Essays on deliberate errors in surveys

Lukas Olbrich

München 2024

Essays on deliberate errors in surveys

Lukas Olbrich

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

eingereicht von
Lukas Olbrich
am 22. Oktober 2024

Erster Berichterstatter: Prof. Joseph Sakshaug, Ph.D.
Zweite Berichterstatterin: Prof. Dr. Yuliya Kosyakova
Dritte Berichterstatterin: Prof. Dr. Frauke Kreuter

Tag der Disputation: 27.11.2024

Acknowledgements

I would like to express my sincere gratitude to all those who have supported and advised me throughout this incredible journey of pursuing my Ph.D. In particular, I am deeply thankful to . . .

- . . . *Prof. Joseph Sakshaug, Ph.D. for giving me the opportunity to pursue a Ph.D., introducing me to the academic world, giving me the freedom to explore research ideas I am interested in, providing feedback and discussing research ideas, and being available at any time.*
- . . . *Prof. Dr. Yuliya Kosyakova for recruiting me to academia, providing feedback and discussing research ideas, mentorship and support, and being available whenever I needed advice.*
- . . . *Prof. Dr. Frauke Kreuter for willing to evaluate this dissertation.*
- . . . *Prof. Dr. Jörg Drechsler and Prof. Dr. Christoph Kern for agreeing to be part of my dissertation committee.*
- . . . *all my co-authors (Elisabeth Beckmann, Jonas Beste, Eric Lewandowski, Silvia Schwanhäuser) for the productive and efficient collaborations.*
- . . . *the GradAB for financial support during my Ph.D. and Sandra Huber for handling all the organizational issues; and the GradAB students for providing such a warm research environment, especially the "Covid cohort" (Timon Hellwagner, Eva Kleifgen, Corinna König, and Sonja Malich) for sharing the special experience of starting a Ph.D. during a pandemic.*
- . . . *the Joe research group (Sophie Hensgen, Sebastian Hülle, Corinna König, Benjamin Küfner, Jan Mackeben, Silvia Schwanhäuser) for all the discussions and exchanges of ideas during our Ph.D.*
- . . . *the IAB departments KEM und INTER for feedback for internal presentations and many fruitful discussions of research ideas. In particular, I would like to thank the department heads Jörg Drechsler, Yuliya Kosyakova, and Herbert Brücker for giving me the freedom to work on research ideas. I would also like to give a special thanks to members of the KEM-SE (Benjamin Baisch, Jonas Beste, Sebastian Hülle, Marieke Volkert, Stefan Zins) for their feedback and helpful advice.*
- . . . *attendants of my presentations at the following conferences and seminars who provided valuable feedback to my work: ESRA 2021; AAPOR 2021 and 2022; LMU Summer retreats; GOR 2024; WAPOR 2023; OeNB Seminar 2022 and 2024; IAB Brownbag Seminar.*
- . . . *my family and friends for their support during my studies and in particular during my Ph.D.*
- . . . *my partner Marie for unconditionally supporting me during all the ups and downs of pursuing a Ph.D. This dissertation would not have been possible without her.*
- . . . *our son Valentin for bringing so much joy to our life and putting all the stress involved with writing a dissertation into perspective.*

Summary

Surveys have always been prone to deliberate errors by the involved actors (e.g., interviewers or respondents). Such errors can bias survey estimates, lead to biased inferences in survey data analysis, and undermine trust in survey data. This dissertation contributes to the literature on preventing and identifying such behavior. The first four contributions focus on deliberate errors by face-to-face interviewers (e.g., fabrication of parts of or entire interviews). The fifth contribution investigates inattentive responding which is a type of deliberate error by web survey respondents.

The first contribution analyzes the deterrence effect of interview audio recordings in face-to-face interviews. With respondent consent, interviews are audio-recorded and these recordings are later used to evaluate the interviewers' behavior. Without recordings, the interviewers' behavior is not observable. Using detailed timestamp data and multiple analysis approaches, we show that audio recordings substantially reduce the prevalence of likely deliberate interviewer errors. The second contribution illustrates how multilevel modeling can be an effective analysis approach to identify fraudulent interviewers. In particular, the developed model focuses on the fraudulent interviewers' behavior over the field period. The method is applied to survey data containing verified falsifications and further data without verified falsifications. The model identifies the verified falsifiers in the former dataset and flags multiple suspicious interviewers in the latter. The third contribution proposes another approach to identifying error-prone interviewers. We exploit that adult self-reported height is stable within short timespans and identify interviewers as error-prone if 1) self-reported heights frequently and substantially differ from measured heights or 2) self-reported heights change frequently and substantially over panel waves. Using multilevel models, we apply the identification approach to four survey datasets and identify several error-prone interviewers. The fourth contribution develops a multivariate approach to analyzing interviewer errors. Using data from ten waves of a yearly panel survey conducted in ten countries (i.e., 100 country-years), we apply multiple indicators of interviewer error both on the interviewer and country-year level. To identify exceptional country-years and interviewers, we use isolation forests and show that interviewer errors are particularly prevalent in several country-years. The results led to the exclusion of multiple country-years from the publicly released data and emphasize the importance of taking the fieldwork institute into account when analyzing interviewer errors. The fifth contribution focuses on preventing and identifying inattentive responding (i.e., providing responses without regard to the question content) in web surveys. As a preventive measure, we experimentally tested the efficacy of so-called commitment pledges that ask respondents to commit to providing accurate responses but found no effect on multiple indicators of inattentive responding. Concerning identification measures, we conducted a further experiment on widely used attention checks and show that large proportions of respondents likely pass such checks by chance. As an alternative, we propose a timestamp-based clustering approach to identify clusters of likely inattentive respondents which is applied to multiple datasets.

The contributions on measures to prevent deliberate errors may guide practitioners in designing surveys. The developed and tested identification methods may guide practitioners and applied researchers who seek to assess the quality of their data. In sum, this dissertation contributes to avoiding the prevalence and (potentially detrimental) consequences of deliberate errors in surveys.

Zusammenfassung

Surveys sind anfällig für absichtliche Fehler der beteiligten Akteure (zum Beispiel der Interviewenden oder der Befragten), die zu verzerrten Schätzungen und fehlerhaften Inferenzen führen und das generelle Vertrauen in Survey-Daten verringern können. Diese Dissertation behandelt Strategien und Methoden zur Prävention und Identifikation solcher Fehler. Die ersten vier Artikel befassen sich mit absichtlichen Fehlern von Face-to-Face-Interviewenden. Der fünfte Artikel beschäftigt sich mit Befragten in Web-Surveys, wobei insbesondere "Inattentive Responding" behandelt wird.

Der erste Artikel untersucht den Effekt von Interviewmitschnitten auf das Interviewendenverhalten in Face-to-Face-Befragungen. Hierbei werden mit dem Einverständnis der Befragten – unter anderem zur späteren Kontrolle der Interviewenden – Tonspuren der Interviews aufgezeichnet. Ohne diese Mitschnitte ist das Verhalten der Interviewenden während des Interviews nicht beobachtbar. Anhand detaillierter Zeitstempeldaten und unterschiedlicher Analyseansätze zeigen wir, dass Interviewmitschnitte (vermutlich absichtliche) Interviewendenfehler erheblich reduzieren. Der zweite Artikel veranschaulicht, wie Multilevel-Modelle zur Identifikation von Interviewendenfälschungen genutzt werden können. Das Modell konzentriert sich auf das Verhalten der fälschenden Interviewenden über den Feldverlauf hinweg. Wir testen die Methode mit Survey-Daten mit verifizierte Fälschungen, die wir identifizieren können, und finden in einem weiteren Datensatz mehrere verdächtige Interviewende. Im dritten Artikel wird ein weiterer Ansatz zur Identifikation fehleranfälliger Interviewender entwickelt. Unter der Annahme, dass sich die Körpergröße von Erwachsenen in kurzen Zeitabständen nicht verändern sollte, klassifizieren wir Interviewende als fehleranfällig, wenn 1) die angegebene Größe ihrer Befragten häufig und erheblich von der gemessenen Größe abweicht oder 2) sich die angegebene Größe ihrer Befragten häufig und erheblich zwischen Panelwellen verändert. Wir verwenden dafür Multilevel-Modelle und identifizieren mehrere fehleranfällige Interviewende in vier Datensätzen. Im vierten Artikel wird ein multivariater Ansatz zur Analyse von Interviewendenfehlern entwickelt. Wir verwenden Daten aus zehn Jahren einer jährlichen Querschnittsbefragung, die in zehn Ländern durchgeführt wurde (insgesamt 100 Länder-Jahre), und wenden mehrere Indikatoren für Interviewendenfehler auf Interviewenden- und Länder-Jahr-Ebene an. Um auffällige Länder-Jahre und Interviewende zu identifizieren, verwenden wir Isolation Forests und zeigen, dass mehrere Länder in bestimmten Jahren besonders auffällige Indikatorwerte aufweisen. Die Ergebnisse führten zum Ausschluss mehrerer Länder-Jahre aus den veröffentlichten Daten und veranschaulichen die Bedeutung des Erhebunsinstituts für die Analyse von Interviewendenfehlern. Der fünfte Artikel befasst sich mit der Prävention und Identifikation von Inattentive Responding (Befragte, die ohne Rücksicht auf den Inhalt der Frage antworten) in Web-Surveys. Als Präventionsansatz untersuchen wir den Effekt von "Commitment Pledges", bei denen sich die Befragten zu Beginn des Interviews verpflichten, bestmögliche Antworten zu geben. Wir finden keinen Effekt der Präventionsmaßnahme auf mehrere Indikatoren für Inattentive Responding. Zur Identifikation führen wir ein Experiment zu "Attention Checks" durch und zeigen, dass ein großer Anteil der Befragten solche Checks wahrscheinlich zufällig besteht. Als Alternative entwickeln und testen wir einen zeitstempelbasierten Clustering-Ansatz, mit dem Cluster, die zu großen Teilen aus Inattentive Respondents bestehen, identifiziert werden können.

Die Artikel über Maßnahmen zur Prävention von absichtlichen Fehlern können bei der Entwicklung von Strategien zur Qualitätssicherung von Surveys unterstützen. Die entwickelten und getesteten Identifikationsmethoden können in der Praxis und angewandten Forschung zur Bewertung der Qualität von Survey-Daten angewandt werden. Insgesamt trägt diese Dissertation dazu bei, die Häufigkeit und die nachteiligen Folgen von absichtlichen Fehlern in Surveys zu verringern.

Contents

List of Figures

List of Tables

Part I

Introduction and Background

1 Introduction

1.1 Motivation

For decades, surveys have been the key source of information on people's behavior, opinions, and characteristics in the social sciences (see De Vaus, 2014; Groves, 2011; Groves et al., 2009; Ornstein, 2013, for discussions of the history of surveys). For most of this period, face-to-face interviews have been the gold standard mode of data collection (Groves, 2011). Frequently postulated arguments for relying on interviewers to collect data include high response rates, the possibility of motivating respondents to provide high-quality data, fewer missing responses, and longer interviews (e.g., Schober, 2018), which have long outweighed their well-known disadvantages (i.e., interviewer effects and high costs). However, interviewer-administered surveys are under severe pressure – not only due to alternative data sources (see Sturgis & Luff, 2021, for evidence on the robust use of surveys in social science research) – but also because they have become more and more expensive and suffer from rapidly decreasing response rates (e.g., de Leeuw et al., 2018; Dutz et al., 2021; Luiten et al., 2020; Williams & Brick, 2018). Hence, the previously postulated benefits of interviewer-administered surveys may no longer justify their excessive costs compared to other modes of data collection.

Concurrently, self-administered web surveys have gained increasing popularity (Couper, 2017). This trend is spurred by the availability of online access panels and crowd-sourcing platforms that allow researchers to conduct their surveys with little effort at low cost and receive data in a short time (Baker et al., 2010, 2013; Brick, 2011; Couper, 2017). Although such non-probability-based surveys deviate from established standards developed for probability-based samples over decades of survey methods research, their use in academic research is increasing (Stefkovics et al., 2024). This trend is also driven by empirical evidence on the external validity of experimental research conducted in non-probability-based surveys under certain conditions (e.g., Berinsky et al., 2012; Coppock, 2019; Mullinix et al., 2015).

Although face-to-face interviews and online-administered interviews are on vastly different paths, ensuring high data quality of the collected data is a challenging problem for both modes. In interviewer-administered surveys, decreasing response rates increase the pressure on all involved actors to collect the required number of interviews, which may come at the cost of data quality. For non-probability-based online surveys, researchers unfamiliar with data quality concepts can conduct their surveys and thus fail to thoroughly monitor and test the quality of the collected data.

The standard approach to assess data quality in (probability-based) surveys is the total survey error (TSE) framework (Biemer, 2010; Groves, 2004b; Groves & Lyberg, 2010), consisting of measurement (validity error, measurement error, processing error) and representation errors (coverage error, sampling error, nonresponse error). In this framework, error is the difference between survey responses and true values which is usually measured as the mean squared error (Groves, 2004b).

The applicability of the TSE framework to non-probability surveys is limited as the representation error is unknown (Baker et al., 2013; Cornesse et al., 2020). Nonetheless, measurement errors can be assessed. Given the available resources, researchers can use the TSE framework to select the survey design with the expected minimal error (Groves, 2004b).

In this dissertation, I focus on types of error that are difficult to plan for: *deliberate errors*. Deliberate errors differ from unintentional errors in that the respective actor "[...] is aware that the action deviates from the guidelines and instructions" (Groves, 2004a, p. 2). Throughout the survey lifecycle, all involved actors may be the source of deliberate errors: researchers, contracted fieldwork institutes, interviewers, and respondents. However, the respective guidelines and instructions differ widely across actors. For example, interviewers usually receive in-depth training on adequate interviewer behavior, while survey respondents are only implicitly expected to read questions and respond honestly. This dissertation contributes to the literature on deliberate interviewer errors in face-to-face surveys and deliberate respondent errors in web surveys.

1.2 Outline

The dissertation consists of five articles (listed in Table 1.1) on preventing and identifying deliberate errors in interviewer-administered face-to-face surveys and web surveys. The next chapter introduces both types of deliberate errors and summarizes previous literature. I will provide a brief summary of the articles when the respective corresponding part of the literature is discussed to illustrate which gaps in the literature are addressed.

Table 1.1: Overview of contributions.

Title	Co-Authors	Publication status
Off the record? Effects of interview audio recordings on interviewer behavior	Jonas Beste, Joseph W. Sakshaug, Silvia Schwanhäuser	under review
Detecting interviewer fraud using multilevel models	Yuliya Kosyakova, Joseph W. Sakshaug, Silvia Schwanhäuser	Published in *Journal of Survey Statistics and Methodology*
The reliability of adult self-reported height: The role of interviewers	Yuliya Kosyakova, Joseph W. Sakshaug	Published in *Economics & Human Biology*
Multivariate assessment of interviewer errors in a cross-national economic survey	Elisabeth Beckmann, Joseph W. Sakshaug	Revise and resubmit at *Journal of the Royal Statistical Society: Series A (Statistics in Society)*
Evaluating methods to prevent and detect inattentive respondents in web surveys	Joseph W. Sakshaug, Eric Lewandowski	Revise and resubmit at *Sociological Methods & Research*

Literature

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., & Lavrakas, P. J. (2010). Research synthesis: AAPOR report on online panels. *Public Opinion Quarterly*, *74*(4), 711–781.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–105.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848.

Brick, J. M. (2011). The future of survey sampling. *Public Opinion Quarterly*, *75*(5), 872–888.

Coppock, A. (2019). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods*, *7*(3), 613–628.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36.

Couper, M. P. (2017). New developments in survey data collection. *Annual Review of Sociology*, *43*, 121–145.

De Vaus, D. A. (2014). *Surveys in social research* (6. ed). Routledge.

de Leeuw, E., Hox, J., & Luiten, A. (2018). International nonresponse trends across countries and years: An analysis of 36 years of Labour Force Survey data. *Survey Insights: Methods from the Field*, 1–11. https://surveyinsights.org/?p=10452

Dutz, D., Huitfeldt, I., Lacouture, S., Mogstad, M., Torgovitsky, A., & Dijk, W. V. (2021). *Selection in surveys: Using randomized incentives to detect and account for nonresponse bias.* NBER Working Paper No. 29549. https://doi.org/10.3386/w29549

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

Groves, R. M. (2004b). *Survey errors and survey costs.* John Wiley & Sons, Inc.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75*(5), 861–871.

Groves, R. M., Fowler, F. J., Couper, M., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology.* Wiley & Sons.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849–879.

Luiten, A., Hox, J., & De Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. *Journal of Official Statistics*, *36*(3), 469–487.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*(2), 109–138.

Ornstein, M. (2013). *A companion to survey research.* SAGE Publications Ltd.

Schober, M. F. (2018). The future of face-to-face interviewing. *Quality Assurance in Education*, *26*(2), 290–302.

Stefkovics, Á., Eichhorst, A., Skinnion, D., & Harrison, C. H. (2024). Are we becoming more transparent? Survey reporting trends in top journals of social sciences. *International Journal of Public Opinion Research*, *36*(2), edae013. https://doi.org/10.1093/ijpor/edae013

Sturgis, P., & Luff, R. (2021). The demise of the survey? A research note on trends in the use of survey data in the social sciences, 1939 to 2015. *International Journal of Social Research Methodology*, *24*(6), 691–696.

Williams, D., & Brick, J. M. (2018). Trends in U.S. face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, *6*(2), 186–211.

2 Background

2.1 Deliberate errors by face-to-face interviewers

2.1.1 Interviewer effects and interviewer variance

Due to the variety of their tasks and the related burden, interviewers have long been identified as a source of error in surveys (e.g., Hyman, 1954; Kish, 1962). Tasks assigned to interviewers (sometimes) include sampling respondents, establishing contact with target persons, convincing them to participate, conducting the interview, and entering responses (Groves et al., 2009; West & Blom, 2017). Hence, interviewers are subject to substantive burden, sometimes enhanced by inaccurate instructions and low-quality instruments. In the TSE framework (Groves et al., 2009), face-to-face interviewers can contribute to coverage, unit nonresponse, measurement, and processing errors (see West & Blom, 2017). In contrast, telephone interviewers work in environments with less daunting tasks and incentives to deviate and are therefore not the focal point of this review and dissertation.

As mentioned above, interviewers can induce a variety of errors. In this literature review, I focus on measurement error. Kreuter (2008) lists four potential sources of interviewer effects on measurement: 1) social desirability bias induced by respondents responding in line with societal norms due to the presence of interviewers, 2) observable interviewer characteristics such as the age, gender, or race influencing the respondent, 3) verbal and nonverbal interviewer behavior such as reactions to responses, body language, or accentuation influencing respondents, and 4) errors when delivering and recording questions and their answers, for example by skipping parts of the question or response categories. Sources 1) and 2) represent unintentional errors induced by respondent reactions to the interviewer. Source 3) represents a mixture of deliberate and unintentional errors; for example, some interviewers may naturally speak relatively fast and thus unintentionally increase the respondents' burden, while other interviewers might deliberately read the questions fast to finish the interview as fast as possible. Source 4) primarily represents deliberate interviewer errors.

The consequences of interviewer error can be distinguished into interviewer bias and interviewer variance (Kish, 1962). Interviewer bias corresponds to systematic differences between survey estimates and "true" values due to the interviewers (e.g., Blaydes & Gillum, 2013; D. W. Davis, 1997; Hatchett & Schuman, 1975; Kerwin & Ordaz Reynoso, 2021; Mensch & Kandel, 1988; Schaeffer, 1980; West et al., 2018a). Such biases may arise from social desirability bias or the observable characteristics of interviewers. Investigating interviewer bias is complicated by the requirement of benchmark values or valid comparison groups, which are rarely available (Kreuter, 2008). Interviewer variance corresponds to the correlation of responses by respondents interviewed by the same interviewer, which depicts a source of random error. Such correlations may arise from differences in how questions and response options are delivered to the respondents across interviewers, either due to deliberate or unintentional errors. However, interviewer variance estimates merely provide evidence that responses nested within interviewers are correlated (Kreuter, 2008). The literature on interviewer effects has focused on interviewer variance, and both terms are often used interchangeably (e.g., Kish, 1962).

Standardized interviewing is the most established approach to counter such effects (Fowler & Mangione, 1990). By instructing interviewers to follow detailed guidelines on conducting interviews (i.e., on reading questions, probing, or recording responses, see Kreuter, 2008), differences across interviewers should be mitigated and thus minimize error. To implement standardized interviewing, in-depth interviewer training, and thorough interviewer supervision and monitoring are recommended (Kreuter, 2008). However, monitoring whether interviewers follow standardized interviewing practices is difficult as interviewers usually work alone in the field, and their work is hard to observe.

In this review, I follow previous literature and focus on interviewer variance. Interviewer variance is the proportion of (residual) variance attributable to the interviewers (Kish, 1962). This proportion is termed intra-interviewer correlation (IIC) or intra-class correlation (ICC). It is often processed to calculate design effects that approximate the variance inflation caused by interviewers relative to a simple random sample (Kish, 1962; Schnell & Kreuter, 2005).

Assuming a random distribution of respondents to interviewers (i.e., a fully interpenetrated design, Mahalanobis, 1946) and following Kish (1962), the classical approach to estimating interviewer variance for a continuous outcome is multilevel modeling (Hox, 1994; Hox et al., 1991):

$$y_{ij} = \beta_0 + \theta_j + \varepsilon_{ij} \tag{2.1}$$

y_{ij} is the outcome observed for respondent i interviewed by interviewer j. β_0 is a constant, θ_j denotes random interviewer intercepts assumed to be distributed with mean zero and variance σ^2_θ, and ε_{ij} is a random error term assumed to be distributed with mean zero and variance σ^2_ε. Following this model, the ICC is $\frac{\sigma^2_\theta}{\sigma^2_\theta+\sigma^2_\varepsilon}$.

In face-to-face surveys, interviewers are rarely randomly assigned to respondents as travel costs would be excessively high. Instead, respondents are usually assigned to interviewers based on regional proximity. Notably, several studies implemented partially interpenetrated designs where interviewers are randomly assigned to respondents within larger regions (e.g., O'Muircheartaigh & Campanelli, 1998; Schnell & Kreuter, 2005; West et al., 2018a). Thus, model 2.1 overestimates interviewer variance since differences in y across interviewers may occur due to differences across regions. In most face-to-face surveys, the region is determined by primary sampling units (PSUs) drawn at the first stage of two-stage sampling designs. To correct for inflated interviewer variance estimates, model 2.1 must be extended to a multilevel cross-classified model by adding random PSU effects:

$$y_{ijk} = \beta_0 + \theta_j + \mu_k + \varepsilon_{ijk} \tag{2.2}$$

The outcome observed for respondent i is now nested in interviewer j and PSU k, and the model incorporates PSU effects μ_k distributed with mean zero and variance σ^2_μ. Note, however, that this approach requires sufficient interpenetration between interviewers and PSUs (i.e., interviewers working in multiple PSUs and multiple interviewers working in the same region). In most studies focusing on measurement error, the model is further extended by respondent and area characteristics

(Hox, 1994; West & Blom, 2017) to account for differential assignment and nonresponse across interviewers (West & Olson, 2010; West et al., 2013, 2018b). Alternative methods exploit that interviewer variance varies across questionnaire items (i.e., lower interviewer measurement variance is expected for respondent age than for attitudinal items) to adjust interviewer variance estimates (Elliott et al., 2022). Brunton-Smith et al. (2017) further extended the standard multilevel model by estimating multilevel location-scale models that allow for estimating interviewer effects on the mean and the variance of the respective survey outcome.

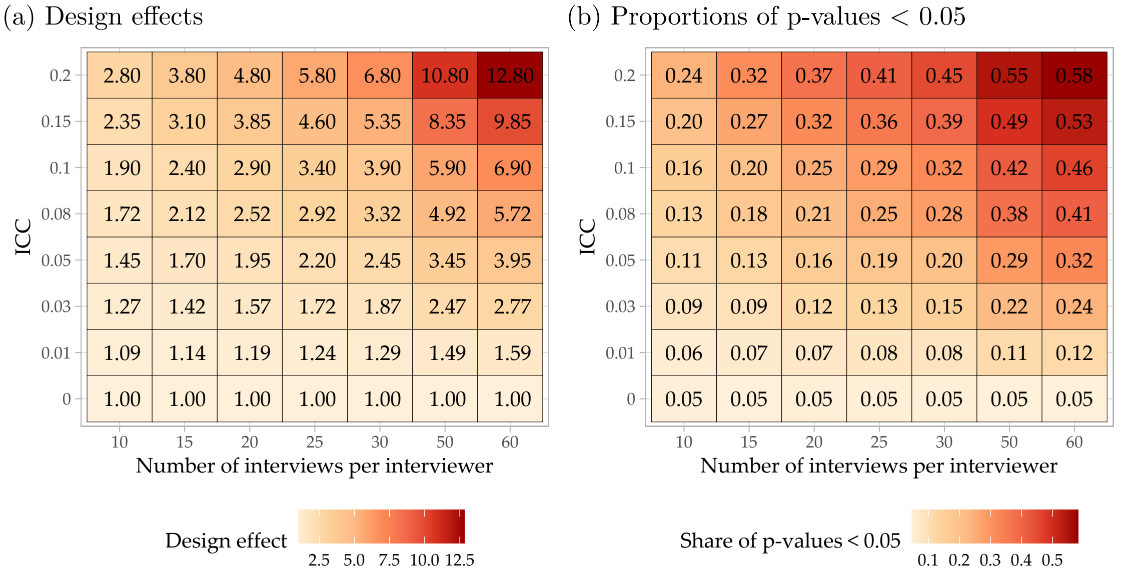
Consequences of interviewer variance

Regardless of its source (i.e., measurement error or nonresponse), interviewer variance inflates the variance of survey estimates. This variance inflation is usually approximated by the so-called design effect D_{eff} (Kish, 1962):

$$D_{eff} = 1 + (m - 1)\rho \tag{2.3}$$

m is the average interviewer workload, and ρ is the estimated ICC. Thus, the design effect increases both with the ICC and the average interviewer workload. Figure 2.1a shows design effects associated with various combinations of ICCs and interviewer workloads. The reported values in each cell denote the respective variance inflation, i.e., ICCs of 0.05 with an average interviewer workload of 20 inflate the variance by 1.95 (Section 2.1.1 below provides an overview of estimated ICCs). Generally, this suggests that practitioners should keep workloads as low as possible. However, this is rarely possible as more interviewers must be hired and trained for each survey. Furthermore, interviewers who are particularly successful in recruiting respondents would be limited to few respondents. Low interviewer workloads are, therefore, hard to implement in times of low response rates and scarce skilled personnel.

Not accounting for interviewer variance in substantive analyses leads to underestimating standard errors. To illustrate the consequences of this effect, I simulated a vector of length 1,500 and mean zero with combinations of ICCs and interviewer workloads depicted on the y-axis and x-axis in Figure 2.1b. I ran 10,000 repetitions for each combination, tested whether the mean differed from zero without accounting for the interviewer variance, and calculated the proportion of p-values below 0.05. The expected proportion in the absence of interviewer variance is 5 percent. As illustrated in Figure 2.1b for large workloads, even small ICCs can lead to a substantial proportion of false positives. For example, the proportion of p-values below 0.05 for an ICC of 0.2 and 10 interviews per interviewer is as high as with an ICC of 0.03 and 60 interviews per interviewer. Hence, not accounting for interviewer variance in survey data analysis can lead to severe consequences, particularly when interviewer workloads are high. Going beyond univariate analyses, Fischer et al. (2019) and Crossley et al. (2021) conducted in-depth evaluations of the effects of interviewers on regression coefficients and highlighted potential attenuation biases due to interviewer measurement error in independent variables. Multiplying the respective coefficient with the inverse of the reliability ratio (i.e., $\frac{1}{1-ICC}$) can limit such biases (Crossley et al., 2021; Fischer et al., 2019).



(a) Design effects

ICC	10	15	20	25	30	50	60
0.2	2.80	3.80	4.80	5.80	6.80	10.80	12.80
0.15	2.35	3.10	3.85	4.60	5.35	8.35	9.85
0.1	1.90	2.40	2.90	3.40	3.90	5.90	6.90
0.08	1.72	2.12	2.52	2.92	3.32	4.92	5.72
0.05	1.45	1.70	1.95	2.20	2.45	3.45	3.95
0.03	1.27	1.42	1.57	1.72	1.87	2.47	2.77
0.01	1.09	1.14	1.19	1.24	1.29	1.49	1.59
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00

(b) Proportions of p-values < 0.05

ICC	10	15	20	25	30	50	60
0.2	0.24	0.32	0.37	0.41	0.45	0.55	0.58
0.15	0.20	0.27	0.32	0.36	0.39	0.49	0.53
0.1	0.16	0.20	0.25	0.29	0.32	0.42	0.46
0.08	0.13	0.18	0.21	0.25	0.28	0.38	0.41
0.05	0.11	0.13	0.16	0.19	0.20	0.29	0.32
0.03	0.09	0.09	0.12	0.13	0.15	0.22	0.24
0.01	0.06	0.07	0.07	0.08	0.08	0.11	0.12
0	0.05	0.05	0.05	0.05	0.05	0.05	0.05

Figure 2.1: Consequences of interviewer variance. Source: Own illustration.

Literature on interviewer variance

The proportion of the residual variance explained by interviewers has been estimated in multiple studies. These studies differ in their designs (i.e., the assignment of respondents to interviewers) and often compare the interviewer variance across question types. Kish (1962) investigated interviewer variance in two small-scale surveys conducted in the U.S. with an interpenetrated design where respondents were randomly assigned to interviewers. He differentiated between critical, ambiguous, and factual items but found no evidence that items from a specific group are more or less susceptible to interviewer variance. The estimated ICCs never exceeded 10 percent in both studies. P. Davis and Scott (1995) investigated interviewer variance in a survey of oral health in New Zealand and found that interviewer variance is substantially smaller than sampling point variance. In addition, they differentiate between attitudinal items, socio-demographics, items on recent behavior, and items on distant behavior. Their results show that interviewer variance is larger for the attitudinal items and the items on recent behavior, though the average ICCs across are below 1 percent for all item categories. O'Muircheartaigh and Campanelli (1998) studied interviewer variance in the British Household Panel Study with a partially interpenetrated design, where addresses were randomly assigned to interviewers within a restricted geographic area. They found that the effects of the interviewers and sampling points were of equal importance and found no difference between factual and attitudinal items. Overall, ICCs were rarely above 10 percent. Schnell and Kreuter (2005) investigated interviewer variance in a German survey with a partially interpenetrated design, where three interviewers worked at each sampling point. Their results show that for most of the 118 items they investigate, interviewer variance plays a more important role than sampling point variance. They categorize the items into several categories and find, on average, higher interviewer variance for sensitive than nonsensitive items, nonfactual than factual items, open-ended questions than closed questions, and no differences between difficult and easy items. Brunton-Smith et al. (2012) used data from the British Crime Survey and found that questions requiring interviewer effort (i.e., probing and showcards) are subject to substantially higher interviewer variance (around 5 percent) than questions requiring effort (i.e., factual questions) (below 1 percent). Brunton-Smith et al. (2017) used cross-classified multilevel location-scale models to study interviewer effects on several questionnaire items in the Understanding Society: the UK Household Longitudinal Study (UKHLS). The model allows for assessing interviewer influences on the item means and its variability, and they report substantial heterogeneity in variability across interviewers. Crossley et al. (2021) estimated interviewer variance for several questions in a German panel on household finances. They find that interviewer variance is substantially larger for financial literacy questions (ICCs up to 38 percent) than for other items. Financial literacy questions quiz respondents on specific financial concepts; thus, interviewers may know the correct response and aid respondents. Beullens and Loosveldt (2016) estimated ICCs for six rounds, 36 countries, and 48 items of the European Social Survey (ESS) and found substantial differences across countries and rounds (averages ICCs over all 48 items vary between 1 and 28 percent). Cernat and Sakshaug (2021) assessed interviewer variance in biomeasures in the U.S. National Social Life, Health, and Aging Project. They find negligible interviewer effects for most measures, though ICCs exceed 10 percent for several variables requiring interviewers to strictly follow detailed instructions (e.g., touch and smell tests). Waldmann et al. (2023) also investigated biomeasures in a cross-national study (Survey of Health, Ageing and Retirement in Europe; SHARE) and found that biomeasures (timed chair stand, walking speed, grip strength, peak flow) are notably more prone to interviewer variance (close to 30 percent of the residual variance explained by interviewers for some measures and countries) than self-reported anthropometric variables. In addition, their results show substantial variation in interviewer variances across countries. In summary, ICCs rarely exceed 10 percent and, if so, for special questionnaire items (i.e., financial literacy) or surveys (i.e., countries with less established survey standards participating in cross-country surveys). In light of the consequences depicted in Figure 2.1, such large values can be detrimental. Note, however, that smaller values are also consequential if interviewer workloads are high.

While earlier literature focused solely on interviewer variance in questionnaire items, the number of studies investigating interviewer variance in data quality measures is increasing. Sources of interviewer variance in data quality measures can be deliberate and unintentional in nature. Several studies found sizeable interviewer variance (larger than 30 percent) in the total triggering rates (the proportion of triggered follow-up questions) or questions that trigger follow-up questions such as the social network size (Böhme & Stöhr, 2014; Brüderl et al., 2013; Herz & Petermann, 2017; Josten & Trappmann, 2016; Kosyakova et al., 2015; Marsden, 2003; Matschinger et al., 2005; Paik & Sanchagrin, 2013; Ruckdeschel et al., 2016; van Tilburg, 1998). Similarly, multiple studies estimated sizeable interviewer variance for the total interview duration (Kirchner & Olson, 2017; Kosyakova et al., 2022; Loosveldt & Beullens, 2013a; Olson & Peytchev, 2007; Vandenplas et al., 2018; West et al., 2018a). With the increasing availability of more fine-grained timestamp data, multiple studies also investigated the durations for questionnaire items or sections (Bergmann & Bristle, 2020; Couper et al., 2013; Holbrook et al., 2020; Kelley, 2020; Loosveldt & Beullens, 2013b; Olson & Smyth, 2015; Sturgis et al., 2021). Further studies report sizeable interviewer variance estimates for item nonresponse (Hox et al., 1991; Pickery & Loosveldt, 2004; Silber et al., 2021) and measures of data quality such as straightlining (selecting the identical response option in sets of adjacent items with the same response scale) or acquiescence (the tendency to "agree") (Hox et al., 1991; Loosveldt & Beullens, 2017; Olson & Bilgen, 2011; Vandenplas et al., 2018).

Previous research has investigated several interviewer characteristics to understand the mechanisms driving interviewer variance. Usually, the explanatory power of such characteristics is analyzed by extending model 2.2 by including the respective characteristic and examining the change in the estimated ICC. West and Blom (2017) provide an in-depth review of considered interviewer characteristics observable to the respondent (i.e., interviewer gender, interviewer age, interviewer race) and interviewer characteristics unobservable to the respondent (i.e., general or within-survey experience). They find no consistent effects of age and gender in the literature but report that responses to racially sensitive questions are affected by the interviewer's race. For general and within-survey experience, effects on response quality are mixed. Hence, the interviewer characteristics typically included in publicly available surveys provide limited insights on interviewer effects and mainly allow for investigating unintentional errors. Recent research by Kühne (2023) showed that interviewer variance can be driven by more complex interviewer-respondent relations (i.e., their mutual perception) than mere socio-demographics.

The introduction of audio recordings of face-to-face interviews has provided opportunities to obtain novel insights into interviewer behaviors during face-to-face interviews that might cause interviewer variance. Several studies analyzed interviewer errors based on audio recordings (e.g., Hicks et al., 2010; Kelley, 2020; Mittereder et al., 2018; C. Sun et al., 2022; Thissen, 2014; Wuyts & Loosveldt, 2022). In such studies, subsets of the available audio recordings are selected, and coders listen to the interviews and classify interviewer behavior, respondent behavior, and their interactions following detailed coding schemes (Ongena & Dijkstra, 2006). However, such studies rarely relate the observed errors to interviewer variance estimates (see Wuyts & Loosveldt, 2022, for an exception). A caveat of using audio recordings to evaluate interviewer behavior is that audio recordings usually require respondent consent, and several studies have shown that consent itself is subject to substantial interviewer variance (Fee et al., 2015; West et al., 2018a).

2.1.2 Interviewer falsification

While interviewer effects (or variance) may arise from deliberate or unintentional errors, interviewer falsification directly refers to deliberate errors. Interviewer falsification is defined as "the intentional departure from the designed interviewer guidelines or instructions, unreported by the interviewer, which could result in the contamination of data. 'Intentional' means that the interviewer is aware that the action deviates from the guidelines and instructions" (Groves, 2004, p. 2). This intentional deviation encompasses a range of behaviors, including complete fabrications of interviews, fabrications of parts of interviews, misreporting contact protocols, miscoding responses to avoid triggering follow-up questions, and interviewing non-sampled persons (Groves, 2004, p. 2). However, the literature mainly focuses on complete fabrications.

Although survey researchers have been aware of interviewer falsification for decades (see Bennett, 1948; Crespi, 1945; Durant, 1946; Evans, 1961; Harrisson, 1947; Hartkemeier, 1944, for early discussions), interviewer falsification has received less attention than interviewer variance. Note, however, that studies on interviewer variance often describe behaviors clearly falling under the definition of interviewer falsification. For instance, Hanson and Marks (1958) identify "interviewer resistance" to questions as a determinant of interviewer variance, where interviewer resistance implies "a tendency on the part of the interviewer to be hesitant about making the inquiry and possibly a tendency to omit or alter the question or assume the answer" (Hanson & Marks, 1958, p. 641). Sturgis et al. (2021) investigated implausibly short response latencies and reported that "[a]n investigation into the causes of these very short latencies was undertaken by listening to audio recordings of a sample of questions with very short latencies. [...] It also showed that the latencies of 1-3 seconds duration were mainly due to interviewers not reading the questions, although it was not clear from the recordings why the questions had been skipped" (Sturgis et al., 2021, p. 706). Furthermore, studies that find large interviewer effects on triggering rates and social network questions frequently put forward interviewer falsifications as a potential explanation (e.g., Josten & Trappmann, 2016; Paik & Sanchagrin, 2013).

Theoretical approaches to explain interviewer falsification are rather general and are difficult to test. Crespi (1945, pp. 437–440) differentiated between survey questionnaire characteristics (e.g., unreasonable length, overly frequent "why's" and "what for's", apparent repetitions of questions, lengthy wording, complex and difficult questions, and antagonizing questions) and administration-related factors (e.g., part-time work, overly difficult sample assignments, and external factors such as the weather or the roads) that can "make or break" (Crespi, 1945, p. 431) the interviewer's morale. Gwartney (2013) distinguished between motivators that are intrinsic (i.e., related to the actual survey) or extrinsic (i.e., remuneration or personal problems) to the interviewer's job. In more general framings, fieldwork institute culture has been mentioned as a determinant of falsification (e.g., Gwartney, 2013; Kennickell, 2015). Furthermore, several authors described interviewer falsification as a principal-agent problem (e.g., Blasius & Friedrichs, 2012; Blasius & Thiessen, 2021; Finn & Ranchhod, 2017; Kennickell, 2015; Kosyakova et al., 2015; Winker, 2016): interviewers may exploit the fact that their principals, i.e., the fieldwork institute, cannot fully observe the interviewers' actions. Since following the fieldwork institute's prescribed guidelines may not maximize the interviewers' utility, they might consider deviations (Winker, 2016). In this framework, aligning the interviewers' incentives with the fieldwork institute's goals or solving the monitoring problem can prevent interviewer falsification. DeMatteis et al. (2020, pp. 18–19) apply the fraud triangle developed for financial crimes by Cressey (1953) as a theoretical framework for falsifications. The three corners of the fraud triangle are pressure/motivation, opportunity, and rationalization. In the falsification context, pressure/motivation consists of the survey design characteristics and the working environment at the fieldwork institute (similar to the Crespi, 1945, framework), opportunity corresponds to the monitoring situation and the probability of detection, and rationalization is the possibility of aligning the falsification with one's norms and values (DeMatteis et al., 2020). Blasius and Thiessen (2021) take a macro-level approach and use Anomie theory to argue that the general level of corruption in the respective country is an important determinant of interviewer behavior. The described theories have mainly been used to derive prevention measures rather than to test hypotheses (though the principal-agent framework has been used in several studies on interviewer behavior, see D'Haultfœuille & Février, 2020; Finn & Ranchhod, 2017; Kosyakova et al., 2015, for examples). For theoretical explanations of falsification *strategies*, previous research has mainly relied on rational choice approaches (Blasius & Thiessen, 2015), where falsifiers must limit their invested effort to maximize their benefits and, at the same time, minimize the probability of detection.

Prevalence of interviewer falsification

Information on the prevalence of interviewer falsifications, both on the proportion of surveys affected and, if so, to which extent, is scarce. The main reasons for this lack of information are that detections of falsifications are rarely reported and that many surveys do not report whether control procedures for falsifications were implemented. Hence, estimates on the prevalence of falsifications can only be drawn from studies on verified falsifications, which likely represent a biased sample (for instance, due to publication bias, i.e., only studies on surveys containing falsifications are written and published). These studies usually focus on complete or partial fabrications, whereas the prevalence of other falsification types is unknown.

A few studies reported proportions of (partially) fabricated data below 1 percent (Li et al., 2011; Schräpler & Wagner, 2005; Schreiner et al., 1988), while the majority of studies reported proportions of (partially) fabricated data between 1 and 10 percent (Bergmann et al., 2019; Beste et al., 2021; Cohen & Warner, 2021; Finn & Ranchhod, 2017; Koch, 1995; Nelson & Kiecker, 1996; Schräpler & Wagner, 2005; Schreiner et al., 1988; Schwanhäuser et al., 2022; Slavec & Vehovar, 2013; Walzenbach, 2021). Proportions beyond 10 percent are rare (Gomila et al., 2017; Murphy et al., 2004), though some exceptional studies documented proportions of (partially) fabricated data above 30 percent (Bossler et al., 2022; Bredl et al., 2012; Castorena et al., 2023).

Though estimates of the prevalence of interviewer falsifications cannot be derived from previous literature, the listed studies emphasize that all surveys are subject to the risk of falsifications. Of course, such risks vary with survey characteristics influencing interviewer burden such as the target population, the quality of the instrument, or the quality of the interviewer training.

Consequences of interviewer falsification

The most crucial information for applied researchers is whether interviewer falsifications influence their results which has been evaluated in multiple studies. Schnell (1991) and Reuband (1990) used data fabricated by students and showed that they generate data highly similar to real-world data, and thus the influence on analyses should be minimal. In contrast, Schräpler and Wagner (2005) showed that even small proportions of fabricated data can change conclusions derived from a regression analysis. Okeke and Godlonton (2014) asked interviewers to hand out vouchers of varying value to respondents based on rolling a dice. They found that the distribution of vouchers differs significantly from the expected distribution if vouchers were indeed assigned based on rolling a die. Hence, interviewer involvement in experimental designs can jeopardize the intended randomization. Finn and Ranchhod (2017) reported relatively minor effects on regression coefficients for cross-sectional analyses, but larger effects on panel analyses. Gomila et al. (2017) showed that the proportion of fabricated interviews in their data leads, on average, to substantial biases in their survey items. Using simulations of higher proportions of fabricated data, they found that increasing proportions lead to larger biases. Sarracino and Mikucka (2017) investigated the impact of duplicated data on results and showed that even small proportions of duplicated data can lead to high probabilities of obtaining biased estimates. Castorena et al. (2023) showed that fabricated data resembles real-world data concerning univariate and regression analyses. DeMatteis et al. (2020) conducted several simulation analyses with varying falsification scenarios to illustrate the potential impacts of falsified interviews on survey results, though such simulations are based on strong assumptions. Lastly, a few studies report on cases of interviewer falsification where the falsifier's motivation was to explicitly bias survey results. For example, Sharma and Elliott (2020) reported on a case where the goal of falsifications was to increase ratings for a specific TV channel. Kuriakose and Robbins (2016) report on a case where respondents interviewed by two suspicious interviewers showed exceptionally high support for one particular party across surveys, which was likely driven by the political motivation of the interviewers.

In summary, a few studies showed that falsifications have negligible effects on research results. The majority of studies showed that falsifications can bias estimates and statistical significance. The extent to which research results are affected depends on the proportion of falsified data, the type of analysis, the "quality" of the falsified data, and the falsifiers' motivation.

Interviewer falsification can have implications beyond impacting survey estimates. It can also damage the reputation of the specific survey and, by extension, undermine trust in survey findings overall (Johnson, 2018). Thus, while thorough reporting on data quality controls typically signifies high data quality, detecting instances of falsification may be perceived as a red flag for poor data quality. In contrast, if there are no reports on data quality controls for a survey, there are also no explicit signals of low quality (Winker, 2016).

Prevention and detection

Strategies to address interviewer falsification can be divided into prevention and detection methods. For prevention, Groves (2004) recommends several measures. These include being aware of issues with burdensome questionnaires, informing interviewers about the prohibition of falsifications and its consequences, conducting background checks on interviewers during recruitment, notifying interviewers about monitoring of their work, and avoiding incentives that may encourage falsification (Groves, 2004, p. 3). Detection strategies can be categorized into observational, recontact, and data analysis methods (Groves, 2004, pp. 3–4). Observational methods, while serving detection and deterrence purposes, are seldom used in face-to-face surveys due to the logistical challenge of having a supervisor accompany each interviewer. However, audio recordings of interviews and further technological monitoring tools offer a means to gain insights into the interviewers' conduct (Thissen & Myers, 2016). Recontact methods involve verifying a random subsample of respondents to ensure that interviews were genuinely conducted. While effective in identifying complete fabrications, they may only reveal insights into partial fabrications if respondents are asked to re-take specific parts of the questionnaire. Data analysis methods play a role in identifying interviewers and interviews that warrant closer scrutiny. These methods involve analyzing survey data or paradata. Paradata represent the information beyond the actual survey data collected during the survey process (Couper, 1998; Kreuter et al., 2010). Targeted efforts guided by data analysis can enhance the efficiency of recontact procedures (Bushery et al., 1999; Hood & Bushery, 1997; Krejsa et al., 1999).

Academic research has focused on detection rather than prevention methods. While several articles summarize best practices concerning preventing interviewer falsification (Johnson et al., 2001; Murphy et al., 2016; Robbins, 2019; Thissen & Myers, 2016), these rarely include an actual evaluation of prevention methods. For example, several studies point out that payment schemes may incentivize falsification (e.g., Groves, 2004). Still, the few studies on payment schemes usually focus on unit response outcomes (D'Haultfœuille & Février, 2020; Rosen et al., 2011; Tourangeau et al., 2012) or overall data quality (Kreuter et al., 2011; Philipson & Lawless, 1997) rather than interviewer deviations (see Menold et al., 2018, for an exception). Instead, most studies have developed and tested statistical methods to identify interviewer falsification. A few studies also provide qualitative analyses of interviewer falsification, such as those by Harrison and Krauss (2002), Kingori and Gerrets (2016), Kriel and Risenga (2014), and Waller (2013). These studies provide detailed accounts of interviewer falsification and, among others, include interviews with falsifiers, shedding light on their strategies and motivations.

Chapter 3: Off the record? Effects of interview audio recordings on interviewer behavior

In Chapter 3, we contribute to the literature on the evaluation of prevention techniques. In particular, we investigate the deterrence effect of audio recordings on interviewers. In the context of deliberate interviewer deviations, audio recordings have mainly been discussed as a detection approach. However, audio recordings usually require the respondent's consent at the beginning of the questionnaire; thus, interviewers are always aware of whether an interview is recorded. Being aware that recordings are analyzed to evaluate their performance, audio recordings might deter interviewers from deviating from standardized guidelines. At the same time, deviations are not observed when the interview is not recorded. To evaluate this deterrence effect, we use data from the German Panel Study Labour Market and Social Security (PASS). In wave 10 of the survey, audio recordings were introduced in the computer-assisted personal interviewing (CAPI) field. To obtain insights into interviewer behavior during the interview, we use questionnaire-module-level timestamp data, which allows for assessing the time spent on each module. Evaluating the effect of audio recordings on interviewer behavior is complicated by the required respondent consent, as differences between recorded and non-recorded interviews can be driven by respondent self-selection or interviewer behavior. To disentangle these effects, we take two approaches: first, we exploit that the PASS is a multi-mode survey (face-to-face and telephone interviews) and the introduction of a CAPI-by-phone mode during the COVID-19 pandemic. Respondents are asked for their consent to recordings in all modes, though telephone interviewers are closely monitored even in the absence of recordings, as supervisors can listen in during interviews. For CAPI interviewers, recordings enable monitoring, which is not possible without respondent consent. We use this difference in the monitoring situation to estimate the effect of audio recordings. The second approach is based on the introduction of audio recordings in wave ten and the longitudinal nature of the survey. As we can observe respondents before the introduction of the recordings, we can rely on respondent-fixed effects to estimate the effect of audio recordings on interviewer behavior. Both analysis approaches show that audio recordings substantially increase module durations and reduce the prevalence of implausibly short module durations, which aligns with a deterrence effect. Our results also indicate that audio recordings reduce interviewer variance in questionnaire items and data quality indicators.

Statistical identification methods

Research on statistical identification methods can be classified into four categories. First, studies aim to develop and test innovative methods using verified falsified survey data initially identified through classical detection approaches (e.g., recontacts). Second, research involves creating and testing new methods using data falsified in controlled experimental settings. Third, researchers explore and evaluate methods using survey data lacking verified falsifications. Lastly, some studies describe the detection process of verified falsifications or existing quality control procedures. Before discussing different types of studies, I briefly provide an overview of the data sources and indicators used to identify falsifying interviewers.

Statistical identification methods are based on indicators derived from survey data or paradata collected during the survey. Survey data-based approaches can be distinguished into formal and content-related indicators (Bredl et al., 2013). Formal indicators are based on differences in response styles and answer patterns between falsified and real interviews (Bredl et al., 2013, p. 16). These include:

- the proportion of middle or extreme responses to Likert-scaled items (Porras & English, 2004; Schäfer et al., 2005)
- the proportion of rounded values reported for questions asking for numerical values (Menold et al., 2013)
- the proportion of item nonresponse (Bredl et al., 2012)
- the proportion of nonresponse to open-ended questions (Menold et al., 2013)
- the proportion of choosing the "other" option as a response to semi-open ended questions (Bredl et al., 2012)
- the proportion of selecting the first or last response for questions with nominal response options (primacy and recency effects) (Menold et al., 2013)
- the proportion of triggered follow-up questions (Hood & Bushery, 1997)
- the proportion of acquiescent responses (Menold et al., 2013)
- non-differentiation of responses to same-scaled item sets (Menold et al., 2013; Schäfer et al., 2005)
- stereotyping in attitudinal items, measured by Cronbach's Alpha (Reuband, 1990)

These indicators are usually calculated at the interviewer-level (i.e., average indicator values are calculated for each interviewer), and falsifiers are expected to have outlying indicator values. The expected direction of deviation is mainly derived from the principal-agent framework (see Schwanhäuser et al., 2022, for an overview). Beyond the response style-based indicators, simple (near-)duplicate analyses have been proposed to identify (partially) copied sets of responses, though such approaches also target fraud by higher-level staff (Kuriakose & Robbins, 2016).

Content-related indicators are also based on the reported survey data but focus on the plausibility of responses (Bredl et al., 2013, p. 16). Naturally, such indicators are survey-specific and not universally applicable. In some cases, the availability of external validation data may serve as a valuable source for assessing the plausibility of responses (Koch, 1995). A special content-related indicator is Benford's Law (Schräpler & Wagner, 2005), which states that the first digits of continuous variables follow the distribution

$$P(first\ digit = d) = log_{10}(1 + \frac{1}{d}),\ d = 1, 2, ..., 9. \tag{2.4}$$

To identify falsifications, deviations of first-digit distributions from Benford's Law are assessed to identify suspicious interviewers (e.g., Swanson et al., 2003), though several conditions need to be fulfilled for its application (Schräpler, 2011, p. 692).

Concerning paradata-based indicators, timestamp data depict the most important source of information (exemplary timestamp data-based indicators: interview duration, item- or questionnaire-module-level durations, daily interviewer workloads). Technological advances also facilitate using more detailed paradata such as mouse movements (Birnbaum, 2012; Robbins, 2019). Contact data represent another source of information for identifying falsifiers (Bredl et al., 2013). Lastly, the availability of contact information (i.e., email addresses or telephone numbers) and record linkage consent rates may aid in identifying suspicious interviewers (Schwanhäuser et al., 2022).

1) Using verified falsifications

As verified falsifications are rarely available, studies that test novel identification approaches usually focus on single datasets where falsifications were detected. Most studies develop and test multivariate approaches to discriminate between falsified and real interviews or identify the most important predictors of falsifications. Schäfer et al. (2005) and Schräpler (2011) used verified falsifications from the German Socio-Economic Panel (GSOEP). Their analyses tested whether Benford's Law and other indicators can identify falsified interviews. They showed that Benford's Law can flag falsifiers, though several remain undetected. Li et al. (2011) trained a logistic regression model with several paradata-based indicators to predict falsification probabilities and explore re-interview samples with increased for the U.S. Current Population Survey. Bredl et al. (2012) used data from a small Eastern European survey and applied cluster analysis based on three indicators. They showed that falsifiers have lower item nonresponse, less extreme responses, and fewer responses to open-ended questions. Bergmann et al. (2019) used data from the SHARE study and first ran a cluster analysis based on 18 indicators to test how well it identifies falsified data (accuracy: 94.3 percent). Applying the same cluster analysis to subsequent wave data and identifying the most likely falsified interviews using logistic regression provided substantially worse results (only 52 out of 1,226 identified interviews were falsified). Cohen and Warner (2021) trained and tested multiple machine learning models (36 models) using 141 predictors based on survey data, paradata, and monitoring results from the AmericasBarometer study and showed that models based on a subset of 30 indicators perform similarly well. Walzenbach (2021) used a logistic regression approach (predictors: extreme responding, item nonresponse, responses to (semi-)open-ended questions, a trick question, responses to occupation item) and showed that the tendency to provide extreme responses is the most important predictor for their data (cross-sectional survey in Germany). Note, however, that the falsifications were initially identified based on the duration of the interview. Schwanhäuser et al. (2022) used both cluster analyses and a novel meta-indicator approach based on more than 30 falsification indicators to successfully identify verified falsifiers in a survey of refugees in Germany. Using discriminant analysis, they find that indicators based on interview duration, Benford's Law, and indicators based on item batteries are particularly valuable for identifying falsifiers.

In the described studies, the main advantage is the availability of verified falsifications that validate identification methods using real-world fraudulent data. However, such studies usually assume that all cases of identified interviewers were falsified, though this seems unlikely in practice (e.g., Castorena et al., 2023). In addition, in many cases, the initial falsification identification was aided by statistical data analysis of quality indicators, sometimes used in the proposed identification methods. Lastly, it is often unclear whether all presumably real interviews have been thoroughly checked or whether further interviews were (at least partially) falsified. Nonetheless, evidence accumulated over multiple studies using verified falsifications should provide viable insights into falsifier behaviors that can be exploited for identification methods in practice. In particular, analysis approaches or single indicators that work across data sets containing verified falsification can provide valuable guidelines for practitioners.

2) Using experimental data

Given the scarcity of publicly available verified falsifications, several studies rely on experimental

Chapter 4: Detecting interviewer fraud using multilevel models

Chapter 4 contributes to the literature on statistical identification methods for identifying falsifiers. Previous literature assumed that falsifiers follow the same behavior throughout the field period. We address this gap in the literature by deviating from this assumption and allowing for different types of falsifiers. Based on a rational choice approach, we propose four different falsifier types that differ in their behavior over the field: Steady low-effort falsifiers follow the same simplistic falsification strategy throughout the field period. Steady high-effort falsifiers follow the same sophisticated falsification strategy throughout the field period. Learning falsifiers reduce the effort invested in falsifications over the field period. Sudden falsifiers do not falsify from the beginning but start falsifying at some point during the field period. Using data from the IAB-BAMF-SOEP Survey of Refugees containing verified interviewer falsifications, we apply a multilevel modeling approach to identify all four types of falsifiers based on their behavior over the field period. The model identifies the previously verified falsifier and flags two undetected falsifiers in the data. We also applied the model to survey data containing no verified cases and identified several interviewers that show patterns in line with behaviors expected for the four types. The results show that falsifiers can follow different strategies and emphasize the necessity of assessing interviewer behavior over the field period to identify interviewer falsification.

Chapter 5: The reliability of adult self-reported height: The role of interviewers

In Chapter 5, we also contribute to the literature on statistical identification methods. While most studies rely on data collected within the current survey, we propose to investigate inconsistencies between data collected in the current survey and data collected in earlier surveys or with different methods. In particular, we investigate the role of interviewers on implausible changes in self-reported height between waves of data collection in panel surveys and differences between measured and self-reported height. Assuming that height is stable for specific age ranges, self-reported height should be constant over time, and we should observe no difference between self-reported and measured height in the absence of measurement error. Using data from two German large-scale panel surveys (GSOEP and PASS) with multiple height reports over time and two cross-sectional surveys (UKHLS and U.S. National Health and Nutrition Examination Survey) containing measured and reported height, we use multilevel models to identify interviewers whose respondents are particularly prone to changes in the reported height over time or differences between the reported and measured height. We identify interviewers with large and frequent reporting errors in all four surveys. One survey contained verified (partially) falsifying interviewers, and we show that investigating changes in presumably time-constant variables can aid in identifying such cases.

data. Birnbaum et al. (2013) asked students to conduct real interviews and later asked them to falsify interviews in three rounds while providing them with feedback on how well they falsified data after each round. Using a random forest classifier and features based on response data and detailed paradata, the model's accuracy in classifying falsified and real interviews decreases with increasing student knowledge (from 96 percent to 86 percent). Using only response data substantially worsens model performance, which signifies the importance of paradata. A set of studies relied on data from an experiment conducted at the University of Gießen in 2012, where 78 students first collected data from other enrolled students and were then asked to falsify interviews in paper-and-pencil interviewing (PAPI) mode (De Haas & Winker, 2014, 2016; Kemper & Menold, 2014; Landrock, 2017; Menold & Kemper, 2014; Menold et al., 2013; Storfinger & Winker, 2013). For example, Kemper and Menold (2014) used the experimental data to test which indicators allow for differentiating between falsified and real data. They found that indicators based on rating scales are particularly relevant, though their model (linear discriminant analysis) only correctly classified 68 percent of the interviews. De Haas and Winker (2014) and De Haas and Winker (2016) used the data to generate synthetic datasets with either parts of an interviewer's workload falsified or parts of single interviews falsified. Similarly, Storfinger and Winker (2013) generated datasets with different characteristics to assess the sensitivity of their detection approach based on the experimental data. The lack of paradata (i.e., interview duration) in this experiment complicates comparisons to studies with verified falsifications. Lastly, Hernandez et al. (2022) asked students to generate falsified data to assess several distribution-based identification methods (i.e., address digits, phone number digits). They found substantial differences in the distributions, though identifying single falsifying interviewers is impossible with such approaches.

Similar to studies using verified falsifications, the main advantage is that researchers know which interviews are falsified and which are not in experimental settings. However, the data generated by students instructed to falsify might substantially differ from the "quality" of falsifications by experienced interviewers. As Birnbaum et al. (2013) showed, even providing feedback to students on their falsification performance reduces the prediction accuracy of their identification methods. Interviewers with in-depth knowledge about real interviews are likely better at generating falsified data (Castorena et al., 2023), which threatens the external validity of insights generated from experimental data.

3) Using data without verified falsifications
Given the lack of available verified falsified data, a further part of the literature tests novel methods to identify suspicious or exceptional interviewers in survey data containing no verified cases. Pickery and Loosveldt (2004) estimated a multilevel multivariate regression model based on five indicators of item nonresponse and extreme responding to identify so-called exceptional interviewers in a small Belgian survey. Brunton-Smith et al. (2017) showed how multilevel location-scale models may aid in identifying interviewers with susceptible contributions to interviewer variance in survey items in the UKHLS. Applying the same model to timestamp data from the UKHLS, Sturgis et al. (2021) identified interviewers with exceptionally high or low variation in response times. Hoellerbauer (2023) developed and evaluated a mixture model that uses re-interview data to estimate interviewer quality. He used both simulated and real-world data collected in Malawi to illustrate the approach. Multiple studies also provide country-level analyses that highlight presumably interviewer-related errors. Judge and Schechter (2009) showed that several surveys conducted in developing countries substantially deviate from Benford's Law, whereas such deviations are not observed in U.S.-based surveys. A further set of studies focused on duplicated or nearly duplicated records (i.e., entirely or partially identical responses in pairs of interviews) in (mostly cross-country) surveys (Blasius & Sausen, 2023; Blasius & Thiessen, 2021; Koczela et al., 2015; Kuriakose & Robbins, 2016; Slomczynski et al., 2017). However, such manipulations can be driven by higher-level staff, as well (e.g., Blasius & Thiessen, 2015). Furthermore, Simmons et al. (2016) emphasized that the prevalence of (near-)duplicates may also be driven by factors unrelated to data manipulations. Lastly, several studies used the proportion of female respondents in gender-heterogeneous households (Sodeur,

1997) to identify unit nonresponse biases that likely arise due to interviewers (Eckman & Koch, 2019; Kohler, 2007; Menold, 2014). These studies used cross-country surveys to identify survey design features that induce biases.

Naturally, the described studies only provide insights into the prevalence of suspicions of falsifications. However, comparisons of results for different datasets allow for the illustration of the plausibility of detected data patterns. Nonetheless, they can only suggest identification methods that can be applied in quality control procedures. At the same time, such analyses can put the responsible data providers under pressure, and thus, the proposed methods and results should be convincing and published with care.

Chapter 6: Multivariate assessment of interviewer errors in a cross-national economic survey

In Chapter 6, we provide an in-depth analysis of interviewer errors in the OeNB Euro Survey, an annual cross-country survey conducted in ten Central, Eastern, and Southeastern European countries. Using data from 100 country-years, we apply a variety of indicators of interviewer error that reflect potential errors for various interviewer tasks (sampling, recruitment, measurement). The first indicator is the internal unit nonresponse bias measure suggested by Sodeur (1997). The second indicator is the daily interviewer workload. The third indicator is interviewer variance. Notably, the OeNB Euro Survey contains items on financial literacy that quiz respondents about economic concepts. As interviewers likely know the correct responses, they have more leeway to influence respondents than for common survey questions, which may lead to enhanced interviewer variance compared to other survey questions. The fourth indicator is satisficing, which is measured by straightlining and item nonresponse. The fifth indicator is a (near-)duplicate analysis. Interviewers might generate highly similar interviews by (partially) fabricating interviews in a repetitive manner or by heavily influencing their respondents' answers, which might lead to highly similar responses. Of course, (near-)duplicates could also be generated by higher-level employees of the fieldwork institutes. Extending previous literature, we analyze these indicators over time and across countries that are rarely investigated for interviewer errors. Notably, several countries changed fieldwork institutes over time which allows for investigating concurrent changes in the error indicators. Given the ubiquity of indicators and country-years, we use isolation forests both on the country-year level and the interviewer level to identify the most suspicious cases. We combine the isolation forest analysis with Shapley values to identify the indicators causing the respective outlier scores. Our results show that all measures of interviewer error vary substantially over time and across countries, in particular when fieldwork institutes change. The results reveal extreme cases for several indicators. For example, ICCs for the literacy items frequently exceed 50 percent. Furthermore, we find that one country has been heavily affected by (near-)duplicates for several years. Follow-up analyses identified the interviewer supervisors as the source of these manipulated data. The contaminated data have been removed from the public use data file. The isolation forest analysis, in combination with Shapley values, serves as an efficient approach to efficiently identify the most extreme country-years and interviewers. Lastly, we illustrate the consequences of interviewer error for substantial analyses with a focus on design effects and bias due to (near-)duplicates.

4) Describing detection processes
The last type of studies are case studies that report on statistical identification techniques and the results of verification processes. An early set of papers discusses interviewer falsification and their detection in U.S. Census Bureau surveys. Hood and Bushery (1997) showed how focused re-interviews based on outlier analysis (ineligibility rates, screening rates, short interview rate, no telephone number rate) exceed the efficiency of random re-interviews for the U.S. National Health Interview Survey. Using the same survey, Bushery et al. (1999) described using data on workloads and timestamps to identify falsifications. Swanson et al. (2003) and Cho et al. (2003) focused on Benford's Law to identify susceptible interviewers in the U.S. Consumer Expenditure Survey. Turner et al. (2002) described how they identified falsifications in their data based on unrealistic workloads and verification processes in a small epidemiologic U.S.-based survey. Porras and English (2004) showed how they use Benford's Law, the lack of variance within falsifier's responses, and response inconsistencies to identify falsifications in a health survey conducted in the U.S. Koch (1995) took a different approach for the German General Social Survey (ALLBUS) and leveraged inconsistencies between survey data and information from the sampling frame to identify falsifications. Yamamoto and Lennon (2018) described the process of detecting falsified data in the Programme for International Student Assessment (PISA) survey and the Programme for the International Assessment of Adult Competencies (PIAAC) survey based on both response data and detailed paradata. Another set of studies reports on detecting falsifications outside the U.S. and Europe. Finn and Ranchhod (2017) provided an overview of instances of falsifications in South Africa and showed how they identified falsifications in the second wave of a panel survey based on Benford's Law and anthropometric measures. Gomila et al. (2017) reported how audio checks introduced due to suspicious interview durations led to detecting falsifications (14 percent) in a survey in Nigeria. Sharma and Elliott (2020) showed how multilevel modeling led to the identification of data manipulation in a survey on television audience measurement in India. C. Sun et al. (2022) described a quality control process for the China Multi-Ethnic Cohort study where outlier detection methods first identify susceptible interviews. Then, each interview is evaluated based on audio recordings, covering 174,012 questions across 5,025 interviews. They identified interviewer falsification as the primary source of interviewer error.

Most studies describing the process of data-driven identification of falsifications come with similar (dis-)advantages as study type 1) (i.e., high validity due to the use of verified falsifications, uncertainty whether falsifiers falsified their entire workload, uncertainty whether all falsifiers were identified). Hence, the generalizability of the results is often limited.

2.2 Deliberate errors by web survey respondents

Given the ubiquity of interviewer errors, some studies conclude that switching to self-administrative modes (i.e., web surveys) might lead to enhanced data quality (e.g., Crossley et al., 2021). Naturally, all potential interviewer-related errors discussed in the previous section are no longer relevant to web surveys. However, all the advantages of interviewer-administered interviewing do not apply, as well (Kreuter, 2008). Particularly relevant for measurement error, interviewers do not guide the respondents through the questionnaire while keeping their motivation up anymore. Hence, if respondents lack motivation – either from the beginning or at some point during the survey – they might break off the survey or finish the questionnaire as fast as possible to collect the incentive.

Such behavior is particularly problematic for non-probability panels with incentives as the main recruitment device (Cornesse & Blom, 2023, p. 880).

As denoted by Tourangeau et al. (2000), the survey response process consists of comprehension, retrieval, judgment, and response selection. Krosnick (1991, 1999) classified respondent types according to their invested effort in the response process. Optimizing respondents take these steps carefully and invest substantial effort. So-called weakly satisficing respondents also take all four steps but invest less effort and provide a seemingly satisfactory response. Strongly satisficing respondents invest even less effort and skip the retrieval and judgment steps by providing presumably reasonable responses. An aggravation of such behavior is inattentive respondents who respond “without regard to item content” (Meade & Craig, 2012, p. 437). Inattentive responding is the focal point of this literature review. In the response process framework, inattentive respondents skip the comprehension, retrieval, and judgment steps as they respond (Anduiza & Galais, 2017; Silber et al., 2022). Such response behavior has received various terms and definitions, including inattentive responding, random responding (though responses are sometimes not random, for example, due to straightlining), or insufficient effort responding (Berinsky et al., 2024). In practice, however, strong satisficing and inattentive responding are challenging to disentangle as both response behaviors skip the retrieval and judgment steps.

Further detrimental types of “respondents” in web surveys are bots (Griffin et al., 2022) and dishonest respondents who try to bias survey outcomes (Arthur et al., 2021). Both types are particularly problematic for surveys where participation is open (i.e., via links posted online) and are not discussed in this review. We also exclude deliberate errors for specific question types such as filter questions (Daikeler et al., 2022) or intentional misreporting of single responses (i.e., income) from this review.

Preventing inattentive responding

Previous research has investigated multiple methods to prevent and detect inattentive responding. A careful questionnaire design that avoids driving respondents to low motivation in the first place is an obvious guideline researchers should follow when conducting web surveys. However, with samples of experienced respondents who mainly participate in surveys to obtain incentives (as on commercial online access survey platforms) inattentive responding may occur regardless of the questionnaire quality. Besides questionnaire design, prevention techniques can be implemented statically or dynamically. Static methods include commitment pledges at the beginning of the questionnaire that ask respondents whether they are willing to provide high-quality responses (Cibelli, 2017; Clifford & Jerit, 2015; Conrad et al., 2017; Hibben et al., 2022), which were successfully applied in interviewer-administered surveys (Cannell et al., 1981). Hibben et al. (2022) and Cibelli (2017) also provided promising results for probability-based panels. Similarly, a few studies found positive effects of warning statements that inform respondents that their responses will be tested for inattentive responding in student samples and, in case of detection, they will not receive credits for participation (e.g., Huang et al., 2012; Meade & Craig, 2012). However, such approaches seem less adequate in general population surveys. In a more dynamic setting, several studies tested warnings that pop up when respondents engage in undesirable behavior, for example, if respondents respond too fast (Conrad et al., 2017; H. Sun et al., 2023; Zhang & Conrad, 2018). While dynamic prevention techniques can be effective measures, they require a-priori thresholds for triggering warnings in dynamic settings. A further concern is that warnings might annoy respondents and induce break-offs, though existing research found no negative effects (Conrad et al., 2017; H. Sun et al., 2023; Zhang & Conrad, 2018).

Detecting inattentive responding

Not all potentially inattentive respondents are deterred by prevention techniques, so detection methods are indispensable. These detection approaches can be classified into a-priori and ex-post methods. A-priori methods are items specifically included to flag respondents. Even commitment pledges can be used as a detection method since respondents who do not commit to providing high-quality responses will likely provide low-quality data. However, these proportions are very small (Cibelli, 2017; Conrad et al., 2017; Hibben et al., 2022). More widely used detection methods are instructed manipulation checks (IMC; also called attention checks or trick questions) as introduced by Oppenheimer et al. (2009). IMCs instruct respondents to fulfill a specific task, i.e., click on a particular word in the question text. A subgroup of IMCs is instructed response items (IRI) (Gummer et al., 2021; Meade & Craig, 2012), which request respondents to provide a specific response. The idea of IMCs is that inattentive respondents likely do not read the instructions, and thus, respondents who fail to follow the instructions are deemed inattentive. While the simplicity of using IMCs is tempting, several problems with IMCs have been identified. First, IMCs add to the respondents' burden, and attentive respondents might interpret IMCs as a signal of distrust and disrespect (Silber et al., 2022). In extreme cases, this might even lead to respondents not following the instructions on purpose (Liu & Wronski, 2018; Silber et al., 2022). Second, depending on the complexity of the IMC, respondents with lower cognitive skills might be more likely to fail (e.g., Silber et al., 2022). Third, in IMCs where respondents are instructed to provide a specific response, inattentive respondents might select the correct response by chance. Fourth, IMCs measure the respondents' attention to a particular point in the questionnaire, which prohibits detecting inattentive respondents at an earlier or later point in the survey (Bowling et al., 2021; Welz & Alfons, 2024). Several studies suggest including multiple IMCs in a single questionnaire (Berinsky et al., 2014; Meade & Craig, 2012), which adds to the respondents' burden. A related detection approach is so-called bogus items where respondents receive an implausible statement such as "I can control the weather with my mind" (Arthur et al., 2021, p. 112). Respondents who agree to such implausible statements are deemed inattentive. For bogus items, caveats similar to those for IMCs apply. Another a-priori approach is items that explicitly ask respondents if they were attentive or whether they think their data should be used for research at the end of the questionnaire (Meade & Craig, 2012). As with using commitment pledges as a detection device, this method depends on the respondents' honesty. In summary, a-priori approaches are widely used in research and well-established in practice, likely due to their simplicity. However, research increasingly shows that these approaches should be applied with care.

Ex-post methods involve analyzing survey data or paradata collected during the web survey. As reviewed by Arthur et al. (2021) as well as Meade and Craig (2012), these approaches include the analysis of screen or item-level durations, the variance of responses in item batteries, response inconsistencies, or outlier detection approaches based on the response data. These methods involve multiple challenges. The analysis of response patterns to item batteries depends on their availability in the respective survey. Similarly, using inconsistency measures requires that clear potential inconsistencies are part of the questionnaire. Furthermore, by definition, outlier approaches assume that inattentive respondents represent a small share of the sample, which may not always be the case (Arthur et al., 2021). For duration data, a significant challenge is identifying an accurate threshold separating the sample into too-fast response times indicating inattentive responding and adequate response times. Here, varying thresholds have been suggested in the literature (see Matjašič et al., 2018, for a review), which will flag varying proportions of presumably inattentive respondents. While most studies have focused on response times for single items, several recent studies have used screen durations along the entire or parts of the questionnaire and more complex modeling approaches to identify inattentive and attentive respondents (Read et al., 2022; Ulitzsch, Pohl, et al., 2024; Ulitzsch, Shin, & Lüdtke, 2024; Ulitzsch et al., 2022). The complexity of these approaches and the necessary data pre-processing steps might hinder their application in practice.

Prevalence of inattentive responding

True values on the prevalence of inattentive responding in web surveys are not available as they rely on the detection methods described above which are subject to both false positives and false negatives. Therefore, any estimate of the prevalence depends on the measurement instrument, its position, the general quality of the questionnaire, and the target population, which prohibits general statements on the prevalence of inattentive responding. Indeed, Arthur et al. (2021) note that widely varying proportions ranging from 10 to 50 percent can be expected for research studies.

Consequences of inattentive responding

Depending on the proportion of inattentive respondents and their response style, inattentive respondents can severely bias substantive analyses. As listed by Ward and Meade (2023) (and assuming random responding by inattentive respondents), inattentive respondents can decrease the reliability of psychometric measures, attenuate correlations (and regression coefficients) and treatment effects, reduce the power for hypothesis testing, bias estimated means, and cause spurious correlations. Nonetheless, recommendations on how to handle inattentive respondents are not entirely clear. While most studies suggest excluding presumably inattentive respondents (see Arthur et al., 2021; Ward & Meade, 2023, in their reviews), removing inattentive respondents can lead to a biased sample composition (e.g., Anduiza & Galais, 2017; Berinsky et al., 2014). Therefore, several studies advise assessing and reporting results with and without inattentive respondents for different definitions of inattention (i.e., show results when only the most inattentive respondents are excluded and when all inattentive respondents are excluded) or to stratify results by levels of attentiveness to enhance transparency (e.g., Berinsky et al., 2014; Kane et al., 2023; Read et al., 2022).

Chapter 7: Evaluating methods to prevent and detect inattentive respondents in web surveys

In Chapter 7, we contribute to the literature on preventing inattentive responding, a-priori and ex-post detection techniques and provide insights on consequences of inattentive responding. These contributions are based on data from a non-probability survey of 16-to-25-year-olds from the U.S., where both the age group and the sample source are prone to inattentive responding. First, the deterrence effect of a commitment pledge at the beginning of the questionnaire is evaluated experimentally. Using several measures of attention and data quality (straightlining, attention check failure, screen durations, break-offs, item nonresponse), we find no effect of the commitment pledge on any outcome. As indicated by the screen durations, the treated respondents likely did not even read the commitment pledge. Hence, such static prevention techniques are not effective in samples prone to inattentive responding. Second, the prevalence of inattentive respondents who randomly pass IRIs is estimated. To do so, respondents were randomly assigned to either a Blank-IRI, where the respondents were instructed to leave the respective item blank, or a Response-IRI, where the respondents were instructed to provide a specific response. If inattentive respondents provide random responses, they could pass the Response-IRI by chance, which is not the case for the Response-IRI. This should lead to higher failure rates for the Blank-IRI. 46.0 percent failed the Response-IRI, while 62.8 percent failed the Blank-IRI, which indicates that substantial proportions of respondents who pass Response-IRIs might do so by chance. These results illustrate the possibility of false negatives of IRI checks and provide further evidence that different types of checks can lead to vastly different results. Third, an easy-to-use paradata-based detection approach is developed and tested. Using screen-level timestamp data, different respondent groups with similar screen duration trajectories are identified by a cluster analysis. In particular, we can identify clusters of likely inattentive respondents with implausibly short screen durations. Furthermore, we show how the cluster analysis results can also identify groups of respondents who speed up at specific sections of the questionnaire. For practitioners, we recommend stratifying substantive analyses based on the identified clusters to assess the potential influence of inattentive respondents. For our data and an additional replication analysis, we show that presumably inattentive respondents can bias results for univariate analysis, regression analysis, and survey experiments.

2.3 Parallels between deliberate interviewer errors and deliberate errors by web survey respondents

While deliberate interviewer errors and deliberate errors by web survey respondents arise from different sources, they share multiple similarities. First, both arise from a lack of monitoring by the researcher or survey manager. Interviewers cannot be observed during interviewing; respondents in self-administered respondents cannot be observed when taking the questionnaire. In both cases, technological advances, such as the collection of detailed timestamps, have been introduced to close this lack of monitoring. Second, incentives can drive both types of error. Fraudulent interviewers may deliberately deviate to maximize their remuneration while inattentive respondents may deviate to receive an incentive after finishing the questionnaire as fast as possible. Third, for both types of error, the prevalence is unknown as data quality control procedures and cases of deliberate errors are rarely reported for face-to-face surveys, and the prevalence of inattentive responding depends on its measurement. Fourth, the challenges posed by lengthy and low-quality questionnaires, particularly on sensitive topics, may push interviewers towards deliberate errors and web survey respondents towards inattentiveness. For face-to-face interviewers, an additional burden comes from the contact and recruitment process. Fifth, to detect both types of deliberate errors, researchers have relied on similar indicators and analysis approaches using both survey data and detailed paradata. Sixth, for both types of error, the consequences on survey estimates depend on the type of analysis, the extent of deliberate errors and their deviation from error-free data, and the motivation of the respective actors.

2.4 Summary

To conclude, neither mode of data collection is unconcerned by deliberate errors by key survey actors. However, as described in the above literature review, researchers can take various steps to reduce the risk of deliberate errors affecting their analyses. These steps can be distinguished into prevention and detection measures. Chapter 3 contributes to the literature on preventing deliberate errors by face-to-face interviewers. Chapters 4, 5, and 6 contribute to the literature on detecting deliberate errors by face-to-face interviewers. Chapter 7 contributes to the literature on preventing and detecting deliberate errors by web survey respondents. In sum, this dissertation may aid in limiting the prevalence and potentially detrimental consequences of deliberate errors in surveys.

Literature

Anduiza, E., & Galais, C. (2017). Answering without reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, *29*(3), 497–519.

Arthur, W., Hagen, E., & George, F. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*, 105–37.

Bennett, A. S. (1948). Toward a solution of the "cheater problem" among part-time research investigators. *Journal of Marketing*, *12*(4), 470–474.

Bergmann, M., & Bristle, J. (2020). Reading fast, reading slow: The effect of interviewers' speed in reading introductory texts on response behavior. *Journal of Survey Statistics and Methodology*, *8*(2), 325–351.

Bergmann, M., Schuller, K., & Malter, F. (2019). Preventing interview falsifications during fieldwork in the Survey of Health, Ageing and Retirement in Europe (SHARE). *Longitudinal and Life Course Studies*, *10*(4), 513–530.

Berinsky, A. J., Frydman, A., Margolis, M. F., Sances, M. W., & Valerio, D. C. (2024). Measuring attentiveness in self-administered surveys. *Public Opinion Quarterly*, *88*(1), 214–241.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*(3), 739–753.

Beste, J., Olbrich, L., & Schwanhäuser, S. (2021). Interviewer:innenkontrolle im Panel Arbeitsmarkt und soziale Sicherung (PASS). *FDZ-Methodenreport 4*.

Beullens, K., & Loosveldt, G. (2016). Interviewer effects in the European Social Survey. *Survey Research Methods*, *10*(2), 103–118.

Birnbaum, B. (2012). *Algorithmic approaches to detecting interviewer fabrication in surveys* [Doctoral dissertation, University of Washington]. https://digital.lib.washington.edu/researchworks/handle/1773/22011

Birnbaum, B., Borriello, G., Flaxman, A. D., DeRenzi, B., & Karlin, A. R. (2013). Using behavioral data to identify interviewer fabrication in surveys. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2911–2920.

Blasius, J., & Friedrichs, J. (2012). Faked interviews. In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, theories, and empirical applications in the social sciences* (pp. 49–56). VS Verlag für Sozialwissenschaften.

Blasius, J., & Sausen, L. (2023). Perceived corruption, trust, and interviewer behavior in 26 european countries. *Survey Research Methods*, *17*(2), 131–145.

Blasius, J., & Thiessen, V. (2015). Should we trust survey data? Assessing response simplification and data fabrication. *Social Science Research*, *52*, 479–493.

Blasius, J., & Thiessen, V. (2021). Perceived corruption, trust, and interviewer behavior in 26 European countries. *Sociological Methods & Research*, *50*(2), 740–777.

Blaydes, L., & Gillum, R. M. (2013). Religiosity-of-interviewer effects: Assessing the impact of veiled enumerators on survey response in Egypt. *Politics and Religion*, *6*(3), 459–482.

Böhme, M., & Stöhr, T. (2014). Household interview duration analysis in CAPI survey management. *Field Methods*, *26*(4), 390–405.

Bossler, M., Gürtzgen, N., Kubis, A., Küfner, B., Olbrich, L., & Schwanhäuser, S. (2022). Revision and new data quality concept due to deviant interviewer behavior in the IAB Job Vacancy Survey. *FDZ-Methodenreport 4*.

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, *24*(4), 718–738.

Bredl, S., Storfinger, N., & Menold, N. (2013). A literature review of methods to detect fabricated survey data. In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewers' deviations in surveys – Impact, reasons, detection and prevention* (pp. 3–24). Peter Lang, Academic Research.

Bredl, S., Winker, P., & Kötschau, K. (2012). A statistical approach to detect interviewer falsification of survey data. *Survey Methodology*, *38*(1), 1–10.

Brüderl, J., Huyer-May, B., & Schmiedeberg, C. (2013). Interviewer behavior and the quality of social network data. In P. Winker, R. Porst, & N. Menold (Eds.), *Interviewers' deviations in surveys. Impact, reasons, detection and prevention* (pp. 147–160). Peter Lang.

Brunton-Smith, I., Sturgis, P., & Leckie, G. (2017). Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(2), 551–568.

Brunton-Smith, I., Sturgis, P., & Williams, J. (2012). Is success in obtaining contact and cooperation correlated with the magnitude of interviewer variance? *Public Opinion Quarterly*, *76*(2), 265–286.

Bushery, J. M., Reichert, J. W., Albright, K. A., & Rossiter, J. C. (1999). Using date and time stamps to detect interviewer falsification. *Proceedings of the Survey Research Method Section, American Statistical Association*, 316–320.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. *Sociological Methodology*, *12*(1981), 389–437.

Castorena, O., Cohen, M. J., Lupu, N., & Zechmeister, E. J. (2023). How worried should we be? The implications of fabricated survey data for political science. *International Journal of Public Opinion Research*, *35*(2), edad007.

Cernat, A., & Sakshaug, J. W. (2021). Interviewer effects in biosocial survey measurements. *Field Methods*, *33*(3), 236–252.

Cho, M. J., Eltinge, J. L., & Swanson, D. (2003). Inferential methods to identify possible interviewer fraud using leading digit preference patterns and design effect matrices. *Proceedings of the Survey Research Method Section, American Statistical Association*, 936–41.

Cibelli, K. L. (2017). *The effects of respondent commitment and feedback on response quality in online surveys* [Doctoral dissertation, University of Michigan]. https://deepblue.lib.umich.edu/bitstream/2027.42/136981/1/kcibelli_1.pdf

Clifford, S., & Jerit, J. (2015). Do attempts to improve respondent attention increase social desirability bias? *Public Opinion Quarterly*, *79*(3), 790–802.

Cohen, M. J., & Warner, Z. (2021). How to get better survey data more efficiently. *Political Analysis*, *29*(2), 121–138.

Conrad, F. G., Couper, M. P., Tourangeau, R., & Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods*, *11*(1), 45–61.

Cornesse, C., & Blom, A. G. (2023). Response quality in nonprobability and probability-based online panels. *Sociological Methods & Research*, *52*(2), 879–908.

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

Couper, M. P., Tourangeau, R., Conrad, F. G., & Zhang, C. (2013). The design of grids in web surveys. *Social Science Computer Review*, *31*(3), 322–345.

Crespi, L. P. (1945). The cheater problem in polling. *Public Opinion Quarterly*, *9*(4), 431–445.

Cressey, D. R. (1953). *Other people's money*. Patterson Smith.

Crossley, T. F., Schmidt, T., Tzamourani, P., & Winter, J. K. (2021). Interviewer effects and the measurement of financial literacy. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(1), 150–178.

Daikeler, J., Bach, R. L., Silber, H., & Eckman, S. (2022). Motivated misreporting in smartphone surveys. *Social Science Computer Review*, *40*(1), 95–107.

Davis, D. W. (1997). Nonrandom measurement error and race of interviewer effects among african americans. *Public Opinion Quarterly*, *61*(1), 183–207.

Davis, P., & Scott, A. (1995). The effect of interviewer variance on domain comparisons. *Survey Methodology*, *21*(2), 99–106.

De Haas, S., & Winker, P. (2014). Identification of partial falsifications in survey data. *Statistical Journal of the IAOS*, *30*(3), 271–281.

De Haas, S., & Winker, P. (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), 643–660.

DeMatteis, J. M., Young, L. J., Dahlhamer, J., Langley, R. E., Murphy, J., Olson, K., & Sharma, S. (2020). *Falsification in surveys*. AAPOR. https://aapor.org/wp-content/uploads/2022/11/AAPOR_Data_Falsification_Task_Force_Report-updated.pdf

D'Haultfœuille, X., & Février, P. (2020). The provision of wage incentives: A structural estimation using contracts variation. *Quantitative Economics*, *11*(1), 349–397.

Durant, H. (1946). The "cheater" problem. *Public Opinion Quarterly*, *10*(2), 288–291.

Eckman, S., & Koch, A. (2019). Interviewer involvement in sample selection shapes the relationship between response rates and data quality. *Public Opinion Quarterly*, *83*(2), 313–337.

Elliott, M. R., West, B. T., Zhang, X., & Coffey, S. (2022). The anchoring method: Estimation of interviewer effects in the absence of interpenetrated sample assignment. *Survey Methodology*, *48*(1), 25–48.

Evans, B. F. (1961). On interviewer cheating. *Public Opinion Quarterly*, *25*(1), 126–127.

Fee, H., Welton, T. A., Marlay, M., & Fields, J. (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*.

Finn, A., & Ranchhod, V. (2017). Genuine fakes: The prevalence and implications of data fabrication in a large South African survey. *World Bank Economic Review*, *31*(1), 129–157.

Fischer, M., West, B. T., Elliott, M. R., & Kreuter, F. (2019). The impact of interviewer effects on regression coefficients. *Journal of Survey Statistics and Methodology*, *7*(2), 250–274.

Fowler, F., & Mangione, T. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. SAGE Publications, Inc.

Gomila, R., Littman, R., Blair, G., & Paluck, E. L. (2017). The audio check: A method for improving data quality and detecting data fabrication. *Social Psychological and Personality Science*, *8*(4), 424–433.

Griffin, M., Martino, R. J., LoSchiavo, C., Comer-Carruthers, C., Krause, K. D., Stults, C. B., & Halkitis, P. N. (2022). Ensuring survey research data integrity in the era of internet bots. *Quality & Quantity*, *56*(4), 2841–2852.

Groves, R. 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, R. M., Fowler, F. J., Couper, M., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. Wiley & Sons.

Gummer, T., Roßmann, J., & Silber, H. (2021). Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research*, *50*(1), 238–264.

Gwartney, P. A. (2013). Mischief versus mistakes: Motivating interviewers to not deviate. In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewers' deviations in surveys – impact, reasons, detection and prevention* (pp. 195–215). Peter Lang, Academic Research.

Hanson, R. H., & Marks, E. S. (1958). Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, *53*(283), 635–655.

Harrison, D. E., & Krauss, S. I. (2002). Interviewer cheating: Implications for research on entrepreneurship in africa. *Journal of Developmental Entrepreneurship*, *7*(3), 319–330.

Harrisson, T. (1947). A British view on "cheating". *Public Opinion Quarterly*, *11*(1), 172–173.

Hartkemeier, H. P. (1944). The use of data collected by poorly trained enumerators. *Journal of Political Economy*, *52*(2), 164–166.

Hatchett, S., & Schuman, H. (1975). White respondents and race-of-interviewer effects. *Public Opinion Quarterly*, *39*(4), 523–528.

Hernandez, I., Ristow, T., & Hauenstein, M. (2022). Curbing curbstoning: Distributional methods to detect survey data fabrication by third-parties. *Psychological Methods*, *227*(1), 99–120.

Herz, A., & Petermann, S. (2017). Beyond interviewer effects in the standardized measurement of ego-centric networks. *Social Networks*, *50*, 70–82.

Hibben, K. C., Felderer, B., & Conrad, F. G. (2022). Respondent commitment: Applying techniques from face-to-face interviewing to online collection of employment data. *International Journal of Social Research Methodology*, *25*(1), 15–27.

Hicks, W. D., Edwards, B., Tourangeau, K., McBride, B., Harris-Kojetin, L. D., & Moss, A. J. (2010). Using CARI tools to understand measurement error. *Public Opinion Quarterly*, *74*(5), 985–1003.

Hoellerbauer, S. (2023). A mixture model approach to assessing measurement error in surveys using reinterviews. *Journal of Survey Statistics and Methodology*, smad037. https://doi.org/10.1093/jssam/smad037

Holbrook, A. L., Johnson, T. P., Kapousouz, E., & Cho, Y. I. (2020). Exploring the antecedents and consequences of interviewer reading speed (IRS) at the question level. In K. Olson, J. D. Smyth, J. Dykema, A. Holbrook, F. Kreuter, & B. T. West (Eds.), *Interviewer effects from a total survey error perspective* (pp. 237–252). Chapman and Hall/CRC.

Hood, C. C., & Bushery, J. M. (1997). Getting more bang from the reinterview buck: Identifying "at risk" interviewers. *Proceedings of the Survey Research Method Section, American Statistical Association*, 820–824.

Hox, J. J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods & Research*, *22*(3), 300–318.

Hox, J. J., de Leeuw, E. D., & Kreft, I. G. G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: A multilevel model. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 439–461). John Wiley & Sons, Inc.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114.

Hyman, H. H. (1954). *Interviewing in social research.* The University of Chicago Press.

Johnson, T. P. (2018). Presidential address: Legitimacy, wicked problems, and public opinion research. *Public Opinion Quarterly*, *82*(3), 614–621.

Johnson, T. P., Parker, V., & Clements, C. (2001). Detection and prevention of data falsification in survey research. *Survey Research*, *32*(3), 1–16.

Josten, M., & Trappmann, M. (2016). Interviewer effects on a network-size filter question. *Journal of Official Statistics*, *32*(2), 349–373.

Judge, G., & Schechter, L. (2009). Detecting problems in survey data using Benford's Law. *Journal of Human Resources*, *44*(1), 1–24.

Kane, J. V., Velez, Y. R., & Barabas, J. (2023). Analyze the attentive and bypass bias: Mock vignette checks in survey experiments. *Political Science Research and Methods*, *11*, 293–310.

Kelley, J. (2020). Accuracy and utility of using paradata to detect question-reading deviations. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, & B. T. West (Eds.), *Interviewer effects from a total survey error perspective* (pp. 267–278). Chapman and Hall/CRC.

Kemper, C. J., & Menold, N. (2014). Nuisance or remedy? The utility of stylistic responding as an indicator of data fabrication in surveys. *Methodology*, *10*(3), 92–99.

Kennickell, A. B. (2015). Curbstoning and culture. *Statistical Journal of the IAOS*, *31*(2), 237–240.

Kerwin, J. T., & Ordaz Reynoso, N. (2021). You know what I know: Interviewer knowledge effects in subjective expectation elicitation. *Demography*, *58*(1), 1–29.

Kingori, P., & Gerrets, R. (2016). Morals, morale and motivations in data fabrication: Medical research fieldworkers views and practices in two Sub-Saharan African contexts. *Social Science and Medicine*, *166*, 150–159.

Kirchner, A., & Olson, K. (2017). Examining changes of interview length over the course of the field period. *Journal of Survey Statistics and Methodology*, *5*(1), 84–108.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, *57*(297), 92–115.

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

Koczela, S., Furlong, C., McCarthy, J., & Mushtaq, A. (2015). Curbstoning and beyond: Confronting data fabrication in survey research. *Statistical Journal of the IAOS*, *31*(3), 413–422.

Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods*, *1*(2), 55–67.

Kosyakova, Y., Olbrich, L., Sakshaug, J. W., & Schwanhäuser, S. (2022). Positive learning or deviant interviewing? Mechanisms of experience on interviewer behavior. *Journal of Survey Statistics and Methodology*, *10*(2), 249–275.

Kosyakova, Y., Skopek, J., & Eckman, S. (2015). Do interviewers manipulate responses to filter questions? Evidence from a multilevel approach. *International Journal of Public Opinion Research*, *27*(3), 417–431.

Krejsa, E. A., Davis, M. C., Hill, J. M., & Bureau, U. S. C. (1999). Evaluation of the quality assurance falsification interview used in the Census 2000 Dress Rehearsal. *Proceedings of the Survey Research Method Section, American Statistical Association*, 635–640.

Kreuter, F. (2008). Interviewer effects. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research* (pp. 369–371). SAGE Publications, Inc.

Kreuter, F., Couper, M., & Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. *Proceedings of the Survey Research Method Section, American Statistical Association*, 282–296.

Kreuter, F., McCulloch, S., Presser, S., & Tourangeau, R. (2011). The effects of asking filter questions in interleafed versus grouped format. *Sociological Methods & Research*, *40*(1), 88–104.

Kriel, A., & Risenga, A. (2014). Breaking the silence: Listening to interviewers when considering sources of non-sampling errors in household survey research in South Africa. *South African Review of Sociology*, *45*(2), 117–136.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

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

Kühne, S. (2023). Interpersonal perceptions and interviewer effects in face-to-face surveys. *Sociological Methods & Research*, *52*(1), 299–334.

Kuriakose, N., & Robbins, M. (2016). Don't get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS*, *32*(3), 283–291.

Landrock, U. (2017). How interviewer effects differ in real and falsified survey data: Using multilevel analysis to identify interviewer falsifications. *methods, data, analyses*, *11*(2), 163–188.

Li, J., Michael Brick, J., Tran, B., & Singer, P. (2011). Using statistical models for sample design of a reinterview program. *Journal of Official Statistics*, *27*(3), 433–450.

Liu, M., & Wronski, L. (2018). Trap questions in online surveys: Results from three web survey experiments. *International Journal of Market Research*, *60*(1), 32–49.

Loosveldt, G., & Beullens, K. (2013a). 'How long will it take?' An analysis of interview length in the fifth round of the European Social Survey. *Survey Research Methods*, *7*(2), 69–78.

Loosveldt, G., & Beullens, K. (2013b). The impact of respondents and interviewers on interview speed in face-to-face interviews. *Social Science Research*, *42*(6), 1422–1430.

Loosveldt, G., & Beullens, K. (2017). Interviewer effects on non-differentiation and straightlining in the European Social Survey. *Journal of Official Statistics*, *33*(2), 409–426.

Mahalanobis, P. (1946). Recent experiments in statistical sampling in the indian statistical institute. *Journal of the Royal Statistical Society*, *109*(4), 325–370.

Marsden, P. V. (2003). Interviewer effects in measuring network size using a single name generator. *Social Networks*, *25*(1), 1–16.

Matjašič, M., Vehovar, V., & Manfreda, K. L. (2018). Web survey paradata on response time outliers: A systematic literature review. *Advances in Methodology and Statistics*, *15*(1), 23–41.

Matschinger, H., Bernert, S., & Angermeyer, M. C. (2005). An analysis of interviewer effects on screening questions in a computer assisted personal mental health interview. *Journal of Official Statistics*, *21*(4), 657–674.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455.

Menold, N. (2014). The influence of sampling method and interviewers on sample realization in the European Social Survey. *Survey Methodology*, *40*(1), 105–123.

Menold, N., & Kemper, C. J. (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, N., Landrock, U., Winker, P., Pellner, N., & Kemper, C. J. (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, N., Winker, P., Storfinger, N., & Kemper, C. J. (2013). A method for ex-post identification of falsifications in survey data. In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewers' deviations in surveys – Impact, reasons, detection and prevention* (pp. 25–47). Peter Lang, Academic Research.

Mensch, B. S., & Kandel, D. B. (1988). Underreporting of substance use in a national longitudinal youth cohort: Individual and interviewer effects. *Public Opinion Quarterly*, *52*(1), 100–124.

Mittereder, F., Durow, J., West, B. T., Kreuter, F., & Conrad, F. G. (2018). Interviewer– respondent interactions in conversational and standardized interviewing. *Field Methods*, *30*(1), 3–21.

Murphy, J., Baxter, R., Eyerman, J., Cunningham, D., & Kennet, J. (2004). A system for detecting interviewer falsification. *American Association for Public Opinion Research Section on Survey Research Methods*, 4968–4975.

Murphy, J., Biemer, P., Stringer, C., Thissen, R., Day, O., & Hsieh, Y. P. (2016). Interviewer falsification: Current and best practices for prevention, detection, and mitigation. *Statistical Journal of the IAOS*, *32*(3), 313–326.

Nelson, J. E., & Kiecker, P. L. (1996). Marketing research interviewers and their perceived of moral compromise interviewers necessity. *Journal of Business Ethics*, *15*(10), 1107–1117.

Okeke, E. N., & Godlonton, S. (2014). Doing wrong to do right? Social preferences and dishonest behavior. *Journal of Economic Behavior and Organization*, *106*, 124–139.

Olson, K., & Bilgen, I. (2011). The role of interviewer experience on acquiescence. *Public Opinion Quarterly*, *75*(1), 99–114.

Olson, K., & Peytchev, A. (2007). Effect of interviewer experience on interview pace and interviewer attitudes. *Public Opinion Quarterly*, *71*(2), 273–286.

Olson, K., & Smyth, J. D. (2015). The effect of CATI questions, respondents, and interviewers on response time. *Journal of Survey Statistics and Methodology*, *3*(3), 361–396.

O'Muircheartaigh, C., & Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *161*(1), 63–77.

Ongena, Y. P., & Dijkstra, W. (2006). Methods of behavior coding of survey interviews. *Journal of Official Statistics*, *22*(3), 419.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872.

Paik, A., & Sanchagrin, K. (2013). Social isolation in America: An artifact. *American Sociological Review*, *78*(3), 339–360.

Philipson, T., & Lawless, T. (1997). Multiple-output agency incentives in data production: Experimental evidence. *European Economic Review*, *41*(3-5), 961–970.

Pickery, J., & Loosveldt, G. (2004). A simultaneous analysis of interviewer effects on various data quality indicators with identification of exceptional interviewers. *Journal of Official Statistics*, *20*(1), 77–89.

Porras, J., & English, N. (2004). Data-driven approaches to identifying interviewer data falsification: The case of health surveys. *Proceedings of the Survey Research Method Section, American Statistical Association*, 4223–4228.

Read, B., Wolters, L., & Berinsky, A. J. (2022). Racing the clock: Using response time as a proxy for attentiveness on self-administered surveys. *Political Analysis*, *30*(4), 550–569.

Reuband, K.-H. (1990). Interviews, die keine sind. "Erfolge" und "Mißerfolge" beim Fälschen von Interviews. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *42*(4), 706–733.

Robbins, M. (2019). New frontiers in detecting data fabrication. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (pp. 771–805). John Wiley & Sons.

Rosen, J., Murphy, J., Peytchev, A., Riley, S., & Lindblad, M. (2011). The effects of differential interviewer incentives on a field data collection effort. *Field Methods*, *23*(1), 24–36.

Ruckdeschel, K., Sauer, L., & Naderi, R. (2016). Reliability of retrospective event histories within the German Generations and Gender Survey: The role of interviewer and survey design factors. *Demographic Research*, *34*(1), 321–358.

Sarracino, F., & Mikucka, M. (2017). Bias and effiency loss in regression estimates due to duplicated observations: A Monte Carlo simulation. *Survey Research Methods*, *11*(1), 17–44.

Schaeffer, N. C. (1980). Evaluating race-of-interviewer effects in a national survey. *Sociological Methods & Research*, *8*(4), 400–419.

Schäfer, C., Schräpler, J.-P., Müller, K.-R., & Wagner, G. G. (2005). Automatic identification of faked and fraudulent interviews in the German SOEP. *Schmollers Jahrbuch*, *125*(1), 183–193.

Schnell, R. (1991). Der Einfluß gefälschter Interviews auf Survey-Ergebnisse. *Zeitschrift für Soziologie*, *20*(1), 25–35.

Schnell, R., & Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, *21*(3), 389–410.

Schräpler, J.-P. (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, J.-P., & Wagner, G. G. (2005). Identification, characteristics and impact of faked interviews in surveys: An analysis by means of genuine fakes in the raw data of SOEP. *Allgemeines Statistisches Archiv*, *89*(1), 7–20.

Schreiner, I., Pennie, K., & Newbrough, J. (1988). Interviewer falsification in Census Bureau Surveys. *Proceedings of the Survey Research Method Section, American Statistical Association*, 491–496.

Schwanhäuser, S., Sakshaug, J. W., & Kosyakova, Y. (2022). How to catch a falsifier: Comparison of statistical detection methods for interviewer falsification. *Public Opinion Quarterly*, *81*(1), 1–31.

Sharma, S., & Elliott, M. R. (2020). Detecting falsification in a television audience measurement panel survey. *International Journal of Market Research*, *62*(4), 432–448.

Silber, H., Roßmann, J., & Gummer, T. (2022). The issue of noncompliance in attention check questions: False positives in instructed response items. *Field Methods*, *34*(4), 346–360.

Silber, H., Roßmann, J., Gummer, T., Zins, S., & Weyandt, K. W. (2021). The effects of question, respondent and interviewer characteristics on two types of item nonresponse. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(3), 1052–1069.

Simmons, K., Mercer, A., Schwarzer, S., & Kennedy, C. (2016). Evaluating a new proposal for detecting data falsification in surveys: The underlying causes of "high matches" between survey respondents. *Statistical Journal of the IAOS*, *32*(3), 327–338.

Slavec, A., & Vehovar, V. (2013). Detecting interviewer's deviant behavior in the Slovenian National Readership Survey. In *Interviewers' deviations in surveys - impact, reasons, detection and prevention* (pp. 131–144). Peter Lang, Academic Research.

Slomczynski, K. M., Powalko, P., & Krauze, T. (2017). Non-unique records in international survey projects: The need for extending data quality control. *Survey Research Methods*, *11*(1), 1–16.

Sodeur, W. (1997). Interne Kriterien zur Beurteilung von Wahrscheinlichkeitsauswahlen. *ZA-Information*, *41*, 58–82.

Storfinger, N., & Winker, P. (2013). Assessing the performance of clustering methods in falsification using bootstrap. In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewers' deviations in* *surveys – impact, reasons, detection and prevention* (pp. 49–65). Peter Lang, Academic Research.

Sturgis, P., Maslovskaya, O., Durrant, G., & Brunton-Smith, I. (2021). The interviewer contribution to variability in response times in face-to-face interview surveys. *Journal of Survey Statistics and Methodology*, *9*(4), 701–721.

Sun, C., Guo, B., Liu, X., Xiao, X., & Zhao, X. (2022). Interviewer error within the face-to-face food frequency questionnaire in large multisite epidemiologic studies. *American Journal of Epidemiology*, *191*(5), 921–925.

Sun, H., Caporaso, A., Cantor, D., Davis, T., & Blake, K. (2023). The effects of prompt interventions on web survey response rate and data quality measures. *Field Methods*, *35*(2), 100–116.

Swanson, D., Cho, M. J., & Eltinge, J. (2003). Detecting possibly fraudulent or error-prone survey data using Benford's Law. *Proceedings of the Survey Research Method Section, American Statistical Association*, 4172–4177.

Thissen, M. R. (2014). Computer audio-recorded interviewing as a tool for survey research. *Social Science Computer Review*, *32*(1), 90–104.

Thissen, M. R., & Myers, S. K. (2016). Systems and processes for detecting interviewer falsification and assuring data collection quality. *Statistical Journal of the IAOS*, *32*(3), 339–347.

Tourangeau, R., Kreuter, F., & Eckman, S. (2012). Motivated underreporting in screening interviews. *Public Opinion Quarterly*, *76*(3), 453–469.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge University Press.

Turner, C. F., Gribble, J. N., Al-Tayyib, A. A., & Chromy, J. R. (2002). *Falsification in epidemiologic surveys: Detection and remediation.* Research Triangle Institute. Washington DC.

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & Von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593–619.

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & Von Davier, M. (2024). Using response times for joint modeling of careless responding and attentive response styles. *Journal of Educational and Behavioral Statistics*, *49*(2), 173–206.

Ulitzsch, E., Shin, H. J., & Lüdtke, O. (2024). Accounting for careless and insufficient effort responding in large-scale survey data—development, evaluation, and application of a screen-time-based weighting procedure. *Behavior Research Methods*, *56*(2), 804–825.

Vandenplas, C., Loosveldt, G., Beullens, K., & Denies, K. (2018). Are interviewer effects on interview speed related to interviewer effects on straight-lining tendency in the European Social Survey? An interviewer-related analysis. *Journal of Survey Statistics and Methodology*, *6*, 516–538.

van Tilburg, T. (1998). Interviewer effects in the measurement of personal network size: A nonexperimental study. *Sociological Methods & Research*, *26*(3), 300–328.

Waldmann, S., Sakshaug, J. W., & Cernat, A. (2023). Interviewer effects on the measurement of physical performance in a cross-national biosocial survey. *Journal of Survey Statistics and Methodology*, smad031. https://doi.org/10.1093/jssam/smad031

Waller, L. G. (2013). Interviewing the surveyors: Factors which contribute to questionnaire falsification (curbstoning) among Jamaican field surveyors. *International Journal of Social Research Methodology*, *16*(2), 155–164.

Walzenbach, S. (2021). Do falsifiers leave traces? Finding recognizable response patterns in interviewer falsifications. *methods, data, analyses*, *15*(2), 125–160.

Ward, M. K., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, *74*, 577–596.

Welz, M., & Alfons, A. (2024). *When respondents don't care anymore: Identifying the onset of careless responding.* arXiv: 2303.07167 `[stat]`.

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, *5*(2), 175–211.

West, B. T., Conrad, F. G., Kreuter, F., & Mittereder, F. (2018a). Can conversational interviewing improve survey response quality without increasing interviewer effects? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(1), 181–203.

West, B. T., Conrad, F. G., Kreuter, F., & Mittereder, F. (2018b). Nonresponse and measurement error variance among interviewers in standardized and conversational interviewing. *Journal of Survey Statistics and Methodology*, *6*(3), 335–359.

West, B. T., Kreuter, F., & Jaenichen, U. (2013). "Interviewer" effects in face-to-face surveys: A function of sampling, measurement error, or nonresponse? *Journal of Official Statistics*, *29*(2), 277–297.

West, B. T., & Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, *74*(5), 1004–1026.

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

Wuyts, C., & Loosveldt, G. (2022). Observing interviewer performance in slices or by traces: A comparison of methods to predict interviewers' individual contributions to interviewer variance. *Survey Research Methods*, *16*(2), 147–163.

Yamamoto, K., & Lennon, M. L. (2018). Understanding and detecting data fabrication in large-scale assessments. *Quality Assurance in Education*, *26*(2), 196–212.

Zhang, C., & Conrad, F. G. (2018). Intervening to reduce satisficing behaviors in web surveys: Evidence from two experiments on how it works. *Social Science Computer Review*, *36*(1), 57–81.

Part II

Contributions

3 Off the record? Effects of interview audio recordings on interviewer behavior

Declaration of Contributions

Lukas Olbrich had the idea of assessing the effects of audio recordings on interviewer behavior. He developed and implemented the analysis approach and wrote the paper.

Contributions of Co-authors

All co-authors provided feedback on the analysis approach and revised and proofread the paper.

Abstract

Survey interviewers are usually instructed to follow standardized interviewing procedures when conducting interviews. The introduction of computer-assisted interviewing facilitated the collection of interview audio recordings to assess if interviewers indeed follow these standardized procedures. However, in most cases, audio recordings require respondent consent, and thus interviewers know whether an interview is recorded or not. In this paper, we analyze the extent to which interviewers change their behavior when interviews are recorded. Using automatically collected timestamp data and both cross-sectional and longitudinal analysis approaches to account for selection effects, we find that interviewers substantially change their behavior when recorded. Our results show that audio recordings lead to substantially fewer very short and very long questionnaire module durations and increase the average time spent on questionnaire modules, which is in line with a deterrence effect. Concerning substantive outcomes, we find larger interviewer variance for estimates derived from non-recorded than for recorded interviews.

Keywords

audio recordings, interviewer behavior, paradata, CARI, interviewer effects, respondent consent

Acknowledgements

Lukas Olbrich gratefully acknowledges financial support from the Graduate Programme of the Institute for Employment Research (IAB) and the University of Erlangen-Nuremberg (GradAB). The authors would like to thank participants of the AAPOR 2022 conference, the LMU summer retreat 2022, and internal IAB seminars for their helpful feedback and comments.

3.1 Introduction

Standardized interviewing is the established method for conducting interviewer-administered survey interviews. Interviewers are instructed to strictly follow standardized interviewing procedures by the survey organization, including reading the questionnaire script to respondents verbatim and using only neutral probes during the interview (Fowler & Mangione, 1990). By following these instructions, differences in interviewer behaviors across interviews should be minimized, thus reducing interviewer effects on survey measurements. However, unlike telephone interviewers who can be monitored live in the telephone studio, face-to-face interviewers usually work alone in the field and survey organizations often do not know whether they follow the standardized procedures.

To counter the lack of monitoring, computer-assisted audio recordings (CARI) have become a standard method to evaluate face-to-face interviewers in recent years (DeMatteis et al., 2020; Edwards et al., 2017; Thissen & Myers, 2016). While such recordings were cumbersome before the introduction of computer-assisted personal interviewing (CAPI) and required additional recording devices (e.g., tape recorders, see Cannell et al., 1981), CAPI heavily simplified recording interviews and transferring the respective audio files to field supervisors (Hicks et al., 2010). The main purposes of these audio recordings are the detection of unstandardized interviewer behavior, monitoring the interviewer-respondent interaction, detecting problematic questionnaire items or modules, and evaluating interviewer performance (Edwards et al., 2017; Hicks et al., 2010; Thissen, 2014).

In this study, we evaluate audio recordings as a *deterrence* device. As interviewers know that they are evaluated based on recordings, they may avoid unstandardized behavior in the first place. However, audio recordings usually require respondent consent (Thissen, 2014). Thus, interviewers must ask respondents whether they consent to the interview being recorded before they start the recording. As a result, interviewers always know which interviews are recorded. In cases where interviewers may want to avoid being recorded, they may also influence the outcome of the consent request. Therefore, the deterrent effect might only work for interviews where the respondent has consented to be recorded. Interviewers have an incentive to perform well in recorded interviews, whereas in non-recorded interviews they can deviate from standardized interviewing with less risk of being detected.

We investigate the extent to which face-to-face interviewers change their behavior when they are recorded using data from the German Panel Study “Labour Market and Social Security” (PASS) (Trappmann et al., 2013, 2019), a large-scale mixed-mode CAPI and computer-assisted telephone interviewing (CATI) survey that collects audio recordings. To disentangle selection effects of respondents consenting to audio recording from interviewer behavior, we take two approaches. In both, we rely on timestamp data automatically collected during the interview to evaluate the duration of reading a question, responding to it, and entering a response.

In the first approach, we make use of the introduction of audio recordings for the CAPI fieldwork in wave 10, which allows for evaluating the effect of recordings on data quality while accounting for respondent fixed effects before and after the audio recordings were introduced. Second, we exploit the mixed-mode setting of the PASS to compare differences in durations across CAPI and CATI interviews and use a forced mode switch for CAPI interviewers who conducted telephone interviews from home during the COVID-19 pandemic. Furthermore, we evaluate whether audio recordings influence substantive outcomes and the prevalence of interviewer variance for several items and data quality measures. We also provide descriptive evidence on differences in durations for the introduction to an embedded trust experiment and its participation rates between recorded and non-recorded interviews.

3.2 Previous literature

3.2.1 Audio recordings and interviewers

Two experiments studied the effects of tape recordings on data quality in paper-and-pencil interviews (Billiet & Loosveldt, 1988; Fowler & Mangione, 1990). Both studies had an additional experimental dimension on interviewer training (basic training vs. extensive training). Their findings indicate positive effects of recordings on data quality, but this depended on the interviewers' training. Further, the recordings did not require respondent consent. In a more recent study, McGonagle et al. (2015) investigated the effect of respondent consent to recording on a variety of data quality outcomes in the Panel Study of Income Dynamics (PSID). Their sample consisted of CATI interviews with an overall consent rate of 94 percent. They found that the total interview duration was substantially shorter for respondents who refused to be recorded. Furthermore, they show that item nonresponse was higher for refusing respondents and that the length of responses to open-ended questions was shorter, suggesting that non-consenters may be less cooperative than consenters. Several further studies investigated differences between recorded and non-recorded interviews with regard to data quality indicators such as item nonresponse or the quality of responses to open-ended questions (e.g., Fee et al., 2016; Sirkis, 2013; Sturgis & Luff, 2015), however, without accounting for selection effects. In sum, the few studies investigating the extent to which recorded interviews differ from non-recorded interviews find that data quality is lower for non-recorded interviews. As pointed out by DeMatteis et al. (2020, p. 27), the effect of recordings on face-to-face interviewers and its role as a potential deterrent, however, is still unknown.

Concerning interviewer effects on recording consent, Fee et al. (2015) analyzed the introduction of audio recordings in the Survey of Income and Program Participation (SIPP) and used multilevel logistic models. They found that interviewers explain almost 45 percent of the variance. Similary, West et al. (2018) found that interviewers explain 20.1 and 12.3 percent of the variance in their experimental groups (conversational interviewing vs. standardized interviewing) in a survey conducted in Germany.

In the growing literature on using audio recordings to investigate interviewer behavior and interviewer-respondent interactions (e.g., Hicks et al., 2010; Kelley, 2020a; Mittereder et al., 2018; Sun et al., 2022; Thissen, 2014; Wuyts & Loosveldt, 2022), non-recorded interviews are often neglected. Consent rates vary widely, and non-recorded interviews can comprise a considerable proportion of the conducted interviews (see Table 3.A1 in the Appendix for a list of consent rates reported in studies on audio recordings in CAPI interviews). Taken together, this calls for a closer analysis of the differences between recorded and non-recorded interviews.

3.2.2 Timestamp data and interviewers

Previous literature on differences between recorded and non-recorded interviews has relied on a variety of rather general data quality measures, such as item nonresponse or total interview duration. We use timestamp data that allow for calculating the duration of questionnaire modules (single items or sets of items on a related topic). These precise durations are particularly valuable for investigating unstandardized interviewing behavior as interviewers require a minimum duration to read the questions and enter respondents' answers. Hence, implausibly short durations enable the identification of unstandardized behavior (Kelley, 2020a; Sturgis et al., 2021). Similarly, exceptionally long durations may indicate that interviewers inappropriately explain questions to respondents or conduct small talk during the interview, which is also not in line with standardized interviewing guidelines (Fowler & Mangione, 1990).

Several previous studies have used timestamp data to investigate the role of interviewers. A set of articles found that only around 2 to 3 percent of the variance in item durations can be attributed to interviewers (Couper & Kreuter, 2013; Garbarski et al., 2020; Olson & Smyth, 2015; Sturgis et al., 2021). Despite these rather small effects, Sturgis et al. (2021) report that 17 percent of the response durations range from only 1 to 3 seconds and state that the main reason for these implausibly short durations was interviewers not reading the questions out. Bergmann and Bristle (2020) used data from the SHARE study to investigate durations of introductory texts to questionnaire sections (e.g., introduction to the interview, introduction to record linkage) where respondents are not required to respond, and thus the durations are entirely driven by interviewers. They find substantial reductions in durations with increasing survey experience and that shorter durations are associated with larger effects on the answers to subsequent items when the introductory texts contain essential information about the respective items. Lastly, Kelley (2020a) combined audio recordings and timestamp data from the Understanding Society Innovation Panel to identify whether and how effective item duration thresholds can identify question-reading deviations. The author coded over 10,000 recorded questions and distinguished between minor (34.5 percent) and major (13 percent) reading deviations. Her findings show that using a 4 words-per-second (WPS) threshold to identify major deviations results in 87.1 percent of correctly classified questions, but only 46.9 percent of all major deviations were detected. Kelley (2020b, Chapter 3) also investigated whether deviations lead to changes in substantive outcomes or data quality, but found no differences.

In summary, previous research has shown that interviewers explain only small proportions of variation in item durations, but that interviewers are particularly important for explaining very short durations. Therefore, we deem item or module durations suitable for investigating interviewer behavior and approximating deviations from standardized interviewing.

3.3 Theoretical framework

The relationship between interviewers and their supervisors can be framed as a principal-agent problem (e.g., D'Haultfœuille & Février, 2020; Kosyakova et al., 2015; Philipson & Lawless, 1997). The interviewers are the agents who work alone in the field and their behavior is difficult to observe. Supervisors are the principals who seek that interviewers follow the standardized interviewing procedures as instructed (e.g., administer all questions, read questions verbatim, read questions at a normal speed). However, interviewers may want to maximize their utility and their utility function likely differs from that of their supervisors. The largest component of interviewers' remuneration in face-to-face surveys is usually a piece-rate wage for every successful interview (especially in Europe). In some cases, a duration-dependent component is added. Thus, following the standardized interviewing protocols is often not incentivized as long as a completed interview is attained. At the same time, following the standardized protocols may require communicating with respondents in a rather rigid and unnatural way which can increase respondent burden (e.g., Schober et al., 2012; Suchman & Jordan, 1990). Hence, under a piece-rate payment scheme, interviewers who conduct a standardized interview receive the same wage as interviewers who deviate from standardization and shorten the question texts either to maximize their remuneration per time spent on the interview or reduce respondent burden by engaging with respondents in a more natural way using conversational principles (e.g., Garbarski et al., 2016; Ongena & Dijkstra, 2007). If supervisors observed the interviewers during the interview, they could detect any deviations and take measures to avoid such behaviors in the future. However, since interviewers typically work alone in the field, supervisors cannot observe whether they follow the instructions, leading to information asymmetry. Interviewers who wish to shorten the questionnaire text to reduce the interview length or prefer a more natural conversational flow may exploit this asymmetry and deviate from the standardized guidelines, which leads to a moral hazard problem.

Audio recordings can reduce this information asymmetry as supervisors can listen to the interviews ex-post and observe whether interviewers deviated from the instructions. If interviewers know when they are being recorded they are less likely to resort to unstandardized behavior, and thus audio recordings can solve the moral hazard problem. Of course, this only holds when interviews are indeed recorded. If respondents do not consent to recording, the moral hazard problem persists and the supervisors cannot observe the interviewers' behavior. Hence, the interviewing situation is systematically different for recorded and non-recorded interviews, which may result in distinct interviewer behaviors in both situations.

Following this framework, we expect fewer deviations from standardization for recorded interviews than for non-recorded interviews due to the difference in information asymmetry. Deviations are not directly observable but should induce the following differences between recorded and non-recorded interviews based on the timestamp data: 1) more very short durations for non-recorded interviews, and 2) more very long durations for non-recorded interviews. Lastly, both 1) and 2) should lead to a narrower distribution of durations for recorded than for non-recorded interviews.

3.4 Data and methods

3.4.1 German Panel Study "Labour Market and Social Security" (PASS)

We use data from the PASS study (Trappmann et al., 2013, 2019). The PASS is a sequential mixed-mode household panel survey launched in 2006 and consists of a combined sample of the residential population and welfare benefit recipients in Germany. Each year, more than 8,000 households participate and all household members aged 15 or older are interviewed. In the sequential mixed-mode design, the starting mode for each new respondent is CAPI but switches to CATI if respondents cannot be contacted or if respondents request so. For subsequent waves, the mode used by the respondent in the previous wave acts as the starting mode.

Both CAPI and CATI interviewers are required to follow standardized protocols. CATI interviews have been monitored via live monitoring (i.e., listening to interviews without recording; Jesske, 2013) since wave 1, while audio recordings – which require the respondent's consent – have been used since wave 4 (in waves 4 to 13, respondents were asked for consent until a specific overall number of successful recordings was reached). In wave 10, audio recordings were also introduced for CAPI interviews, with consent rates ranging between 32 and 37 percent. PASS began collecting timestamps in wave 2. The frequency of collected timestamps substantially increased over time, enabling the calculation of durations in seconds for many questionnaire items and modules.

As timestamps are automatically generated, technical problems might lead to erroneous measurements. To counter such issues, we follow Sturgis et al. (2021) and exclude durations of zero seconds (less than 0.2 percent of durations for the key questionnaire modules). We also exclude respondents older than 65 years who receive a shorter questionnaire. Furthermore, we exclude respondents who used a non-German version of the questionnaire to avoid language differences influencing the results.

3.4.2 Empirical approach

A major challenge with analyzing the effect of audio recordings is the non-random distribution of recordings across respondents. As respondents must consent to be recorded, consenting respondents might systematically differ from non-consenting respondents. In addition, it is difficult to account for this selection process as interviewers can also influence the consent decision. For example, interviewers might anticipate that the interview will be difficult and potentially deviate from the standardized protocol for the consent item. In the forthcoming analysis, we use two distinct methods to counter the effect of selection on the difference between recorded and non-recorded interviews, each described below.

Cross-sectional analysis

In the first approach, we exploit that the PASS is a mixed-mode survey. The CATI interviews are conducted with special software that allows supervisors to listen in during the interview. Hence, CATI interviewers are always subject to the “risk” of monitoring, and whether an interview is also recorded is assumed to not impact the prevalence of deviant interviewer behavior or suspicious durations for CATI interviewers. CAPI interviewers, however, are only subject to the risk of monitoring if the respective interview is recorded. Thus, recordings should make a larger difference for CAPI interviewers. Both CATI and CAPI interviewers must ask their respondents for consent to be recorded. Assuming that the selection process for recording consent is identical between CATI and CAPI interviews, we can use a differences-in-differences approach to estimate the recording effect on CAPI interviewers.

A major threat to that assumption is that the interview situation is substantially different between face-to-face and telephone interviews. In face-to-face interviews, the interviewer is in the respondent’s house, and recording the interview could be perceived as an invasion of privacy. In telephone interviews, the interviewer is more distant and the respondent always has the option to simply hang up. This discrepancy might lead to differences in the selection process. We examine this possibility by exploiting the fact that the 14th wave of the PASS was subject to a forced mode switch due to the COVID-19 pandemic (Jesske & Schulz, 2021). Data collection took place from February to September 2020 but was interrupted by the pandemic. The survey started with CATI and CAPI interviewers working in the telephone studio and in the field, respectively. After the onset of the COVID-19 pandemic in Germany, data collection came to a halt but soon resumed with telephone interviewers working from home using the CATI software and face-to-face interviewers working from home and conducting the interviews via telephone using the CAPI software. Hence, the interview situation was identical to the CATI interviews, except for the live monitoring capability. If the interview situation affected the selection process, we do not expect differences between CATI interviews and the CAPI-by-phone sample.

To analyze differences between CATI and CAPI-by-phone, we focus on differences in the prevalence of very short and very long durations. As the PASS duration measures are only available in seconds and therefore the duration distribution is not sufficiently continuous for RIF (recentered influence functions) or quantile regressions (Biewen et al., 2022; Chernozhukov et al., 2013), we rely on the distribution regression approach where the distribution of the dependent variable is modeled by many logistic regressions with dummy variables for multiple thresholds as dependent variables (Biewen et al., 2022; Chernozhukov et al., 2013).

$$\begin{gathered} P(duration_i \leq y \mid mode_i, recorded_i, X_i) \equiv \\ F(y \mid mode_i, recorded_i, X_i) = \\ \frac{exp(\beta_0 + \beta_1 recorded_i + \beta_2 mode_i + \tau(mode_i \times recorded_i) + X\gamma)}{1 + exp(\beta_0 + \beta_1 recorded_i + \beta_2 mode_i + \tau(mode_i \times recorded_i) + X\gamma)} \end{gathered} \tag{3.1}$$

The duration for interview i is denoted by $duration_i$. We select the thresholds y based on the distribution for the respective questionnaire module and use every second between the 1st and 97.5th percentiles as outcomes. In the regression equation, β_0 is the constant, β_1 denotes the coefficient on whether the interview is recorded (= 1) or not (= 0), and β_2 denotes the coefficient for the modes (CATI as reference category). The key parameter is τ, which denotes the coefficient for the interaction between the modes and recording and thus measures the difference in the differences between recorded and non-recorded interviews across modes. Furthermore, we include several respondent control variables X (age, gender, German citizenship, education, month in field period, number of previous PASS participations, see Table 3.B1) with coefficients γ. We depict the results by showing the cumulative factual and counterfactual distributions by CAPI mode for the recorded interviews. The factual distributions are obtained by predicting the proportion of observations less than or equal to the current threshold for the respective recorded CAPI group, the counterfactual distributions are obtained by the factual minus τ.[1]

Table 3.1 shows the number of interviews by mode and whether the interview was recorded or not. As mentioned above, we can differ between CATI, CAPI face-to-face, and CAPI-by-phone interviews in the wave 14 data. Respondents with missings in the covariates were excluded from the analysis (N=12). The majority of interviews were conducted in the CATI mode. The rates of recorded interviews in the sample differ vastly across modes. In CATI, it is close to 70 percent, while it is around 30 percent for both types of CAPI modes. The difference between CATI and CAPI-by-phone is particularly interesting as the interview situation for the respondent is identical in both modes. Of course, different types of respondents might have self-selected into the CAPI and CATI modes, which might drive the large difference. An alternative explanation is that CATI interviewers could always be monitored both in the telephone studio and in the home office, while CAPI-by-phone monitoring is only possible with audio recordings. Concerning the interviewers' role in audio recordings, we fitted simple multilevel logistic regressions with a binary variable (1 = recorded, 0 = not recorded) for each mode separately and found intra-interviewer correlation coefficients (abbreviated by ICC or IIC) of 0.342 for CATI, 0.303 for CAPI face-to-face, and 0.590 for CAPI-by-phone.

[1] To ensure monotonicity in the estimated distributions (see Chernozhukov et al., 2009), we use the rearrangement technique implemented in the R package `Rearrangement` (Graybill et al., 2016)

Table 3.1: Recorded and non-recorded interviews by mode, PASS W14.

Mode		Non-recorded	Recorded	All
CATI	N	1373	3048	4421
	% row	31.06	68.94	100.00
CAPI, face-to-face	N	1101	560	1661
	% row	66.29	33.71	100.00
CAPI-by-phone	N	731	350	1081
	% row	67.62	32.38	100.00
All	N	3205	3958	7163
	% row	44.74	55.26	100.00

Across modes, recorded respondents are slightly older, have more panel experience, are more likely to have German citizenship, and are slightly less likely to be unemployed (see Tables 3.C1 to 3.C3). As mentioned above, these differences can be driven by respondent self-selection or interviewers avoiding recordings for specific respondents.

For the module durations, we focus on the *Social participation* and *Life satisfaction* modules as we also use them in the longitudinal analysis (see Section 3.4.2). The former contains questions on the feeling of being part of society and one's position within society, while the latter consists of only one question on general life satisfaction (see Table 3.2 for the questionnaire texts). We also analyze 11 further questionnaire modules distributed over the questionnaire (see Table 3.D1 for questionnaire text) but discuss their results only briefly. Figure 3.E1 in the Appendix shows density plots for the duration distributions for all 13 modules by mode and by recorded and non-recorded interviews. The distributions for recorded and non-recorded CATI interviews and recorded CAPI face-to-face and CAPI-by-phone interviews are very similar, while the distributions for non-recorded CAPI face-to-face and CAPI-by-phone interviews are wider and shifted to the left (indicating shorter durations) in most cases.

Table 3.2: Questionnaire text for social participation and life satisfaction.

Variable	Text
Social participation: Part of society	Let us now move on to a couple of other questions. One may have the feeling of being integrated into everyday social life and being a real part of society or one may feel rather excluded. What about your case? To what extent do you feel a part of society or do you feel rather excluded? Please use the numbers from 1 to 10 to rate your opinion. '1' means that you feel excluded from social life. '10' means, you feel part of it. The numbers from '2' to '9' allow you to grade your assessment.
Social participation: Position in society	There are groups in our society, which are considered to be rather at the top while other groups seem to be positioned at the bottom. Where would you see yourself using the numbers 1 to 10? 1 means that you see yourself at the very bottom, 10 means that you are positioned at the very top. [The numbers from 2 to 9 allow you to grade your assessment.] *The last phrase in brackets was excluded from wave 12 onward.*
Life satisfaction	And now one final question: In general, how satisfied are you currently with your life on the whole? '0' means that you are 'very dissatisfied', '10' means that you are 'very satisfied'. The numbers '1' through '9' allow you to grade your assessment.

Longitudinal analysis

In our second approach to countering the selection problem, we use the longitudinal design of the PASS and the introduction of audio recordings in wave 10 of the CAPI field. The PASS includes several modules that appear in every wave, which allows for assessing the development and potential changes of durations for specific modules over time. We focus on the durations for the questionnaire modules *Social participation* and *Life satisfaction.* For these items, timestamps have been collected since wave 7.

Figure 3.1 provides descriptive evidence on the timestamp data and their development over time with a focus on rather short durations. For both social participation and life satisfaction, we calculate a threshold that denotes a 4 WPS limit by counting the words in the questionnaire text and dividing them by 4 (Kelley, 2020a). The resulting thresholds are 30.25 seconds (33.25 before wave 12 due to a slight change in questionnaire text) for social participation and 11.5 seconds for life satisfaction. As depicted in Figure 3.1, approximately 30 percent of durations were shorter than the 4 WPS limit before the introduction of audio recordings for social participation. After the introduction of audio recordings, the percentage stays relatively high for non-recorded interviews but steadily decreases to 20 percent in wave 14. On the contrary, the proportion is close to 2.5 percent in wave 10 and decreases further until wave 14 for recorded interviews. In sum, the recorded interviews drag down the overall proportion of very short durations. For life satisfaction, the developments for recorded and non-recorded interviews are very similar, although the proportion below the threshold for non-recorded interviews is lower and varies between 15 and 20 percent. A potential explanation for this difference is that the social participation module contains more explanatory text that interviewers may skip or shorten.

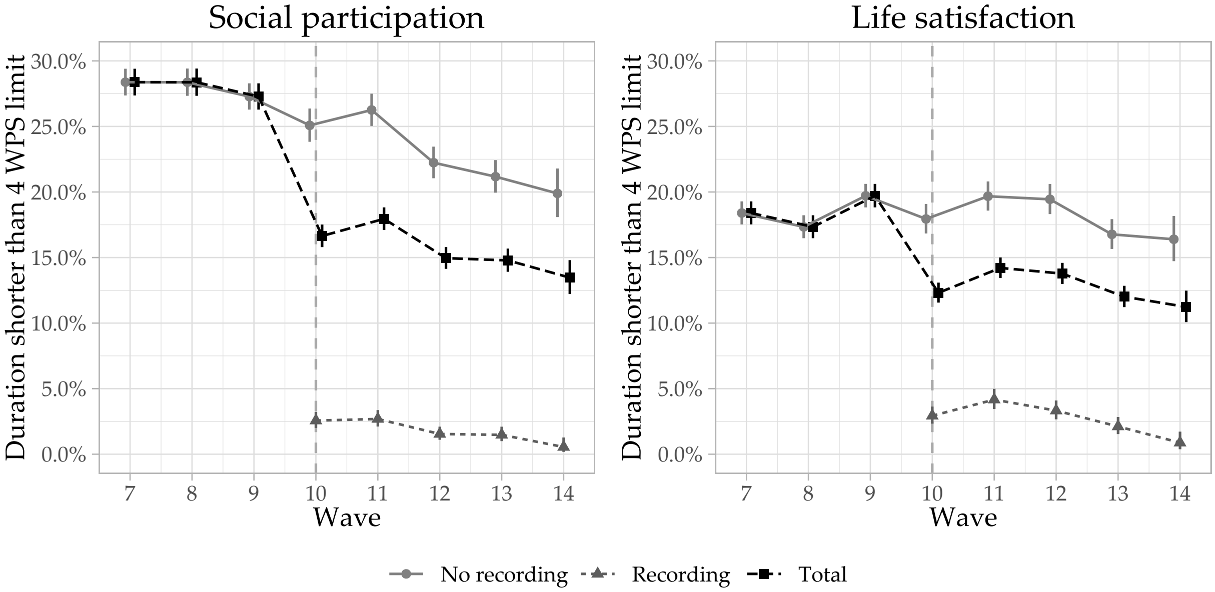


Figure 3.1: Development of the proportion of durations shorter than the 4 WPS limit over time by recorded and non-recorded interviews and for the full sample, only CAPI interviews (with 95 percent confidence intervals).

Although the patterns shown in Figure 3.1 indicate that module durations for recorded and non-recorded interviews differ, alternative factors such as changes in sample and interviewer composition or respondent self-selection into recordings could influence the results. To account for these factors, we rely on fixed effects regressions and exploit the fact that we can observe respondents before the introduction of audio recordings in wave 10. We create a treatment group of respondents whose interviews were recorded in wave 10 and a control group of respondents who were not recorded in wave 10. In this differences-in-differences type setting, we first examine the duration data for both groups in waves 7 to 9 to evaluate whether they differ in their development over time. For wave 10, we estimate the audio recording effect by comparing the observed value for the recorded interviews to the counterfactual that is derived from the value of the control group and the previous development of the treatment group.

We restrict the analysis to CAPI respondents who participated in all waves 7 to 10. Earlier data is not used due to the lack of timestamps and later data are also omitted since respondents might change their recording status in later waves. Lastly, we restrict the data to respondents who were interviewed by the same interviewer over all four waves to ensure that changes in the interviewer composition do not influence the results. To ensure that extreme outliers do not influence the results, respondents with durations exceeding 100 seconds per item are excluded (Sturgis et al., 2021). Hence, we refrain from analyzing the effect of audio recordings on the prevalence of very long durations in this analysis. Section 3.F in the Appendix shows results for alternative thresholds for extreme outliers.

In summary, we end up with a sample of 8,660 observations from 2,165 respondents (795 recorded in Wave 10) for the social participation module and 7,452 observations from 1,863 respondents (706 recorded in Wave 10) for the life satisfaction module. The ICCs for audio recordings estimated from a multilevel logistic model for wave 10 are 0.341 for the social participation sample and 0.317 for the life satisfaction sample. For the longitudinal analysis, the distribution regression cannot be used as there is no variation over time for some respondents in the tails of the duration distribution, i.e., the durations for some respondents are never below (or above) the respective threshold. Therefore, we use the natural logarithm of the durations as the dependent variable (Couper et al., 2013; Sturgis et al., 2021). Figure 3.G1 in the Appendix shows the raw development of the natural logarithm over time for both groups and modules.

The main assumption for our approach is that the durations for the respondents who consented to recordings in wave 10 would have developed the same as the durations for the respondents who did not consent. For both social participation and life satisfaction, the average durations are slightly higher for the recorded group. Tests for differences in trends across groups preceding the introduction of recordings do not indicate statistically significant differences which validates our estimation approach.

The estimation equation for the longitudinal analysis is:

$$\ln(duration_{it}) = \sum_{\substack{t=7\\ t\neq 9}}^{T} \beta_t d_i + \mu_i + \lambda_t + \varepsilon_{it} \tag{3.2}$$

As before, $duration_{it}$ is the duration observed for respondent i in wave t. The model contains respondent fixed effects (μ_i) that account for time-constant respondent-specific heterogeneities (such as the differences in respondent characteristics reported in Tables 3.C4 and 3.C5) and wave fixed effects (λ_t) that account for special circumstances in specific waves. The key parameter is the coefficient of the binary "recorded in wave 10" group variable d_i that is estimated for each wave with wave 9 as the reference group. Hence, this coefficient reports the extent to which the recorded and non-recorded groups differ in the respective wave compared to wave 9.

3.5 Results

3.5.1 Cross-sectional analysis

Figure 3.2 shows the results of the distribution regression for social participation in wave 14.[2] For both CAPI types, we depict the factual (with recording) and counterfactual (without recording) cumulative distribution of the module durations for the recorded interviews. The difference between both distributions represents the effect of audio recordings. For the CAPI face-to-face interviews, both distributions differ substantially. While almost none of the durations are below 30 seconds for the factual distribution, more than 20 percent of the durations are below 30 seconds for the counterfactual. At the same time, the proportion of durations exceeding 80 seconds is lower for the factual distribution. Hence, the prevalence of very short and very long durations is lower due to audio recordings. For the CAPI-by-phone interviews, the prevalence of rather short durations is

[2] See Figure 3.H1 in the Appendix for differences between recorded and non-recorded CATI interviews. Except for the insurance module, we find little evidence of differences between both CATI groups. The insurance module asks for the type of health insurance, although the vast majority are insured in a statutory health insurance fund (the first response option).

also substantially lower for the factual than for the counterfactual distribution, although there is no difference in the prevalence of rather long durations.

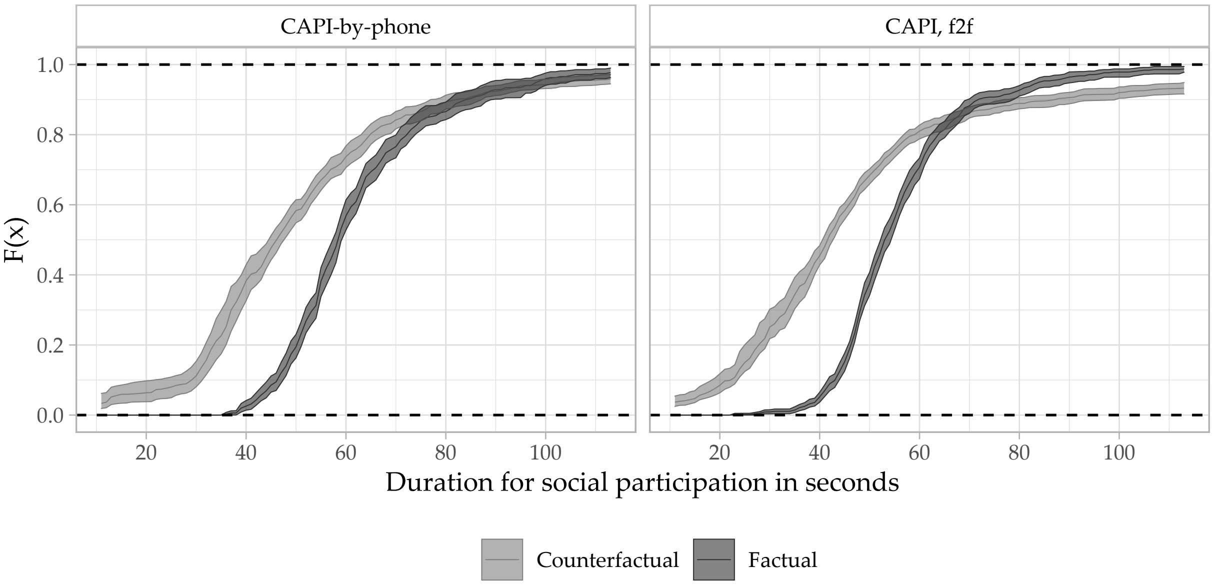


Figure 3.2: Cumulative factual and counterfactual distribution of duration for social participation for recorded interviews by CAPI groups, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

Figure 3.3 shows the results for life satisfaction. Note that this module appears at the end of the questionnaire. Thus the results for very long durations should be interpreted cautiously since technical issues, such as forgetting to exit the interviewing software application, could influence them. However, such errors should not vary across recorded and non-recorded interviews and modes. For CAPI face-to-face, the prevalence of very short durations is lower for the factual than for the counterfactual distribution, i.e., the proportion of durations below 10 seconds is 10 percent for the counterfactual and close to zero for the factual distribution. At the same time, very long durations are also less frequent for the factual distribution. For CAPI-by-phone, the difference in very short durations is similar to CAPI face-to-face, but the difference between the factual and the counterfactual distribution is more significant for very long durations. For example, in the counterfactual distribution, 20 percent have durations exceeding 60 seconds, while such durations are very rare for all other groups. Investigating the non-recorded CAPI-by-phone interviews more closely shows that these extremely long durations accumulate for several interviewers, but it remains unclear why it only happens for non-recorded CAPI-by-phone interviews.

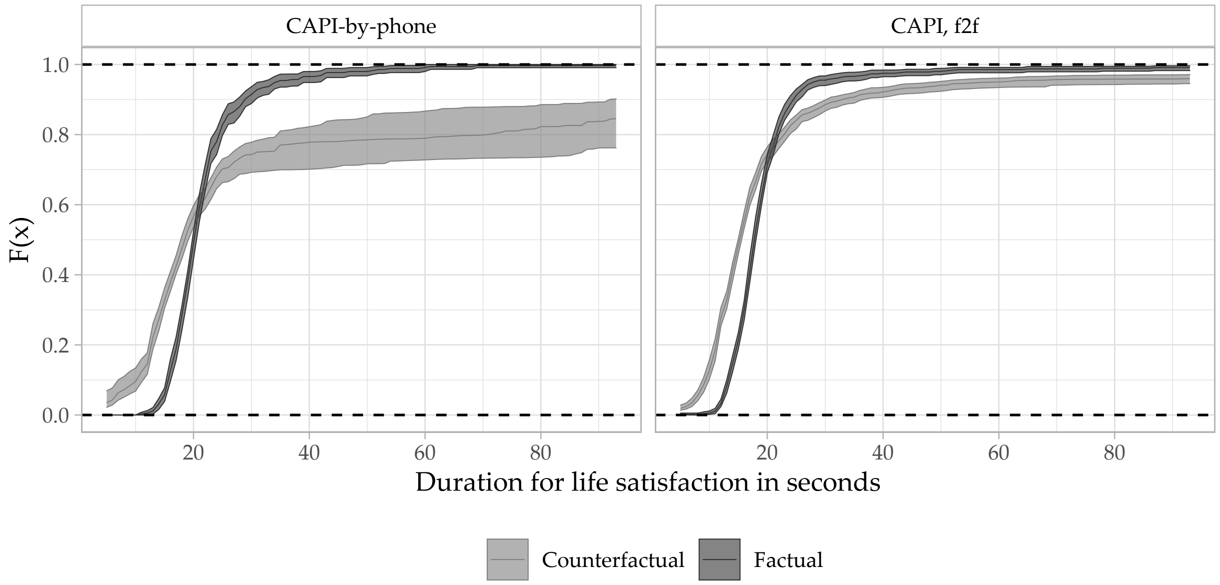


Figure 3.3: Cumulative factual and counterfactual distribution of duration for life satisfaction for recorded interviews by CAPI groups, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

Figures 3.I1 to 3.I11 in the Appendix show the results for the 11 other questionnaire modules. For every module, the prevalence of very short durations is lower for the factual than for the counterfactual distributions. The proportion of long durations for most modules and modes is also lower for the recorded interviews. This pattern is slightly more pronounced for CAPI face-to-face than for CAPI-by-phone.

3.5.2 Longitudinal analysis

The results of the fixed effects longitudinal regression analysis are shown in Figure 3.4. The estimates depict the difference between respondents who consented to recordings versus respondents who did not consent in wave 10. Wave 9 is used as the reference year in the estimation. For social participation, the introduction of audio recordings increased the duration by $exp(26.3) \times 100 = 30.0$ percent (or approx. 15.0 seconds after re-transforming to duration in seconds) while there were no differences before the introduction. For life satisfaction, the introduction of audio recordings increased the duration by $exp(14.8) \times 100 = 16.0$ percent (or approx. 2.9 seconds after re-transforming to duration in seconds). Before the introduction, there were no significant differences.

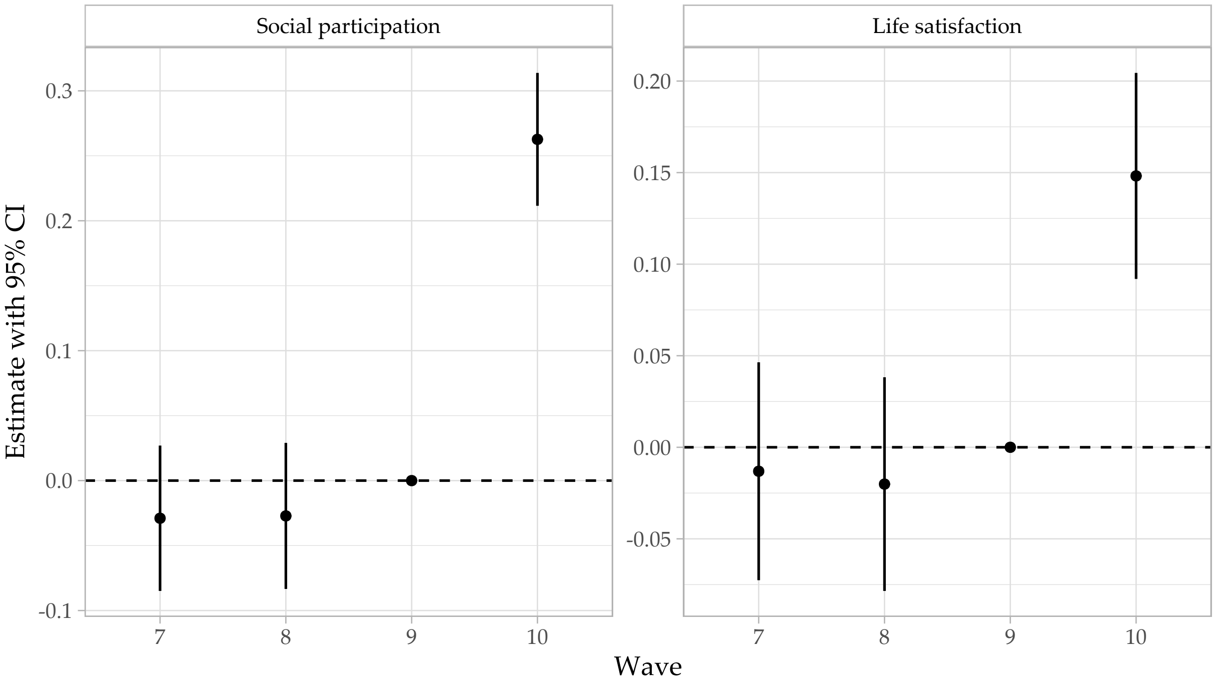


Figure 3.4: Wave-specific effects of audio recordings on the logarithm of the module durations with 95 percent confidence intervals (clustered on respondent-level).

Due to the longitudinal character of the PASS, we can also analyze the stability of durations for respondents who maintained or switched their audio recording consent status from wave 10 to wave 11 (see Table 3.J1 in the Appendix). We refrain from using the distribution regression approach due to the small group sizes. Instead, Figure 3.5 displays the respective proportions for the 4 WPS limit. For both social participation and life satisfaction, only 2.5 percent or less of respondents who are recorded in both waves are faster than 4 WPS. For always non-recorded respondents, these proportions are around 35 percent for social participation and around 25 percent for life satisfaction. Among those who switched from recorded to non-recorded, the proportion increased from below 2 to 31 percent for social participation. Among those who switched from non-recorded to recorded, the proportion decreased from 24 to less than 3 percent. The changes are similar for life satisfaction. These results suggest that duration patterns for recorded and non-recorded interviews do not carry over when recording statuses are switched. Hence, interviewers seem to drive the discrepancies in durations between recorded and non-recorded interviews.

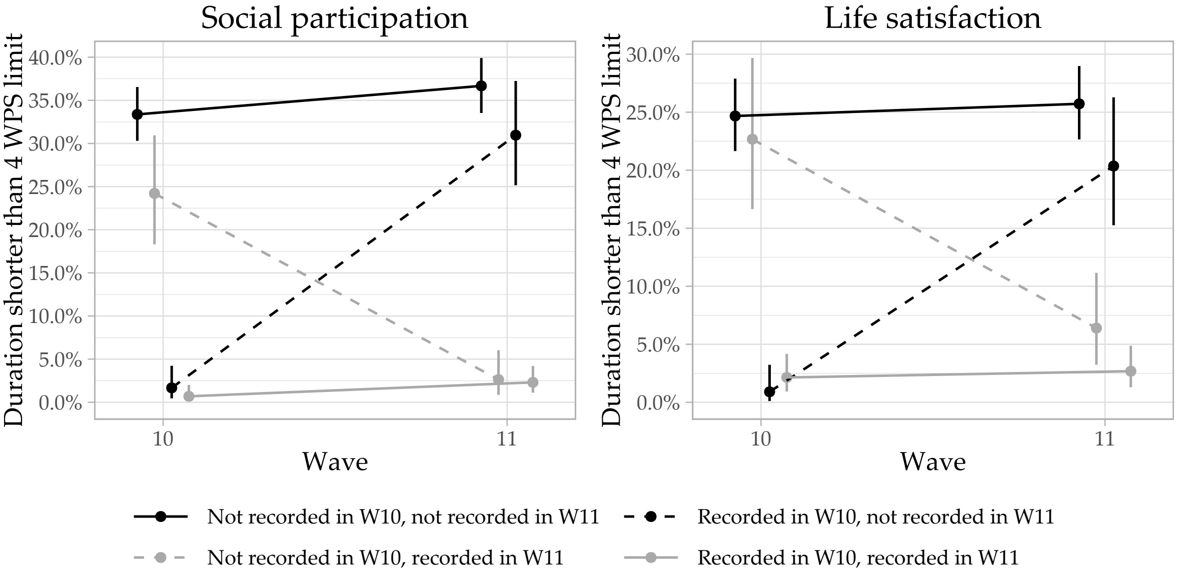


Figure 3.5: Development of the percentage of durations shorter than the 4 WPS limit from wave 10 to wave 11 (with 95 percent confidence intervals).

3.6 Effects on substantive outcomes

To evaluate whether more standardized interviewer behavior leads to changes in measurement, we first return to the longitudinal setting described in section 3.4.2. As dependent variables, we use the social participation and life satisfaction items (see Table 3.2). We fit linear models similar to the model in equation 3.2 with respondent and wave fixed effects, an interaction term of the wave indicator, and an indicator denoting whether the respondent belongs to the group recorded in wave 10 or not. Respondents with item nonresponse in any of the four waves are excluded from the analysis. The results are reported in Table 3.K1 of the Appendix.

We do not find evidence for changes in responses due to the introduction of audio recordings in wave 10 for the social participation and life satisfaction items. Note, however, that respondents in the longitudinal analysis are rather experienced and thus already know the respective questions. In addition, for these particular items interviewers may skip questionnaire text such as “Let us now move on to a couple of other questions” that is unlikely to affect responses, particularly when respondents are experienced.

To investigate a case where skipping words or speeding could have a larger impact on questions unknown to the respondent, we focus on an embedded experiment on trust in the CAPI face-to-face field in wave 14 before the onset of the pandemic. The introductory text contains 203 words and ends with asking the respondent if they are willing to participate (see Appendix 3.L). Applying the same restrictions as before (no respondents older than 65, only interviews conducted in German, no zero durations), almost 88 percent of the sample consented to participate. For recorded interviews, 96.2 percent participated, for the non-recorded interviews, 83.5 percent participated. In sum, 89.6 percent of non-participants also did not consent to being recorded. While the differing participation rates can be purely driven by self-selection (e.g., respondents who don’t consent to be recorded are also less likely to participate in the experiment), the duration data indicate that interviewers might also play a role in administering the introductory text. 24.7 percent of durations for the non-recorded interviews were faster than 4 WPS, while only 2.9 percent were faster among recorded interviews. Of the non-recorded fast interviews, 45.4 percent did not participate.

To evaluate whether audio recordings exacerbate interviewer variance, we estimate ICCs for the items in the social participation and life satisfaction modules, and nondifferentiation measures for three item batteries in the wave 14 data. Nondifferentiation measures the similarity of responses in same-scaled adjacent items (Yan, 2008) and is a frequently used data quality indicator to investigate interviewer and respondent effects on measurement (e.g., Kim et al., 2019; Loosveldt & Beullens, 2017; Olbrich et al., 2024). We measure nondifferentiation for the *Attitude to self*, *Role model*, and *Functions of work* modules (see Table 3.D1) by calculating the standard deviations of responses. The ICCs are based on multilevel models with covariates (variables listed in Table 3.B1) and random effects for the interviewers and denote the residual variance explained by interviewers. The sample is split into four groups (CATI, recorded; CATI, not recorded; CAPI, recorded; CAPI, not recorded) and ICCs are estimated separately for each group. The results are reported in Table 3.3. For the recorded CATI interviews, only the interviewer variance for nondifferentiation in the role model item battery is statistically significant. For the non-recorded CATI interviews, the only statistically significant effect is estimated for the functions of work battery. Concerning the recorded CAPI sample, none of the survey items are subject to interviewer variance, while only the item battery on role models is subject to statistically significant (at the 5 percent level) interviewer variance. For the non-recorded CAPI interviews, we obtain statistically significant interviewer variance for all variables. For the survey items, the estimated ICCs vary between 2 and 3 percent, while the ICCs range between 3.5 and 12.2 percent for the item batteries. These estimates are larger than the estimates for the recorded CAPI interviews. In sum, these results indicate that interviewer variance is larger for non-recorded CAPI interviews.

Table 3.3: ICCs by mode and recording.

Variable	CATI, recorded	CATI, not rec.	CAPI, recorded	CAPI, not rec.
Part of society	N/E	N/E	0.008 (0.320)	0.029 (0.005)
Position in society	0.004 (0.227)	N/E	0.023 (0.082)	0.023 (0.008)
Life satisfaction	0.005 (0.102)	0.002 (0.356)	N/E	0.024 (0.024)
ND attitude to self	0.007 (0.082)	0.012 (0.105)	0.030 (0.050)	0.035 (0.001)
ND role model	0.007 (0.026)	0.009 (0.224)	0.047 (0.020)	0.086 (0.000)
ND functions of work	0.003 (0.156)	0.042 (0.001)	0.016 (0.137)	0.122 (0.000)

Notes: CAPI-by-phone and CAPI face-to-face interviews were pooled to obtain sufficient sample sizes.
p-values in parentheses are based on restricted likelihood ratio test of interviewer variance.
ND = Nondifferentiation. N/E = Model did not converge.

3.7 Discussion

Audio recordings are increasingly used to monitor face-to-face interviewers and their interactions with respondents (DeMatteis et al., 2020; Edwards et al., 2017; Hicks et al., 2010; Thissen & Myers, 2016). While audio recordings have been thoroughly analyzed, their impact on interviewer behavior has been neglected (DeMatteis et al., 2020). In this study, we investigated the extent to which audio recordings deter interviewers from unstandardized behavior. Using timestamp data to approximate unstandardized behavior (Kelley, 2020a), we used two analysis approaches to counter respondent self-selection effects and isolate the effect of audio recording. The first approach used cross-sectional data and exploited a forced mode switch to compare differences between recorded and non-recorded interviews in face-to-face and telephone interviews, and relied on distribution regression techniques to analyze changes across the entire duration distributions for specific questionnaire modules. The second approach used longitudinal data to investigate differences before and after the introduction of audio recordings in a panel setting.

The cross-sectional analysis showed that audio recordings reduce the prevalence of very short and very long durations of questionnaire modules, thus narrowing the duration distributions. The results of the longitudinal analysis showed that the introduction of audio recordings substantially increased module durations. We interpret these findings as evidence that interviewers operate closer to the standardized interviewing protocol when recorded, which is in line with the postulated deterrence effect of audio recordings for unstandardized interviewer behavior (DeMatteis et al., 2020) and previous findings on the effect of audio recording consent on data quality (Fee et al., 2016; McGonagle et al., 2015). We also analyzed the extent to which the introduction of audio recordings affected average reported values for multiple items in the longitudinal setting and found no evidence of changes due to audio recordings. For the cross-sectional analysis, we found evidence that audio recordings might play a more important role in more complex settings such as the introduction of new items and experiments where deviant behavior is more consequential (Bergmann & Bristle, 2020). Furthermore, the results showed that interviewer variance is larger in non-recorded CAPI interviews than in recorded ones.

As stated by Olson et al. (2020, p. 2), “a fundamental goal of research on interviewers is understanding what contributes to (and how to minimize) the IIC.” Our results suggest that audio recordings in CAPI interviews work as an effective deterrence device and thus may enhance overall data quality and reduce ICCs (or IICs). Ensuring high consent rates for audio recordings limits the potential for deviant behavior in non-recorded interviews and may lead to a reduction of ICCs. Furthermore, this study showed that evaluating interviewers solely based on audio recordings can produce a biased picture of their overall work, which is of particular importance for items that require interviewers to read extensive instructions or introductory texts. Studies using audio recordings and behavior coding to disentangle measurement error sources should be aware that recorded interviews may not be representative of the entire sample. In addition, interviewer-induced errors could be systematically smaller in recorded interviews as our analyses have shown that interviewers adapt their behavior when recorded.

This study is not without limitations. First, the audio recordings were not randomly assigned due to the requirement of respondent consent. Hence, differences between recorded and non-recorded interviews could be driven by respondent behavior. We took several steps to counter this potential issue but cannot completely rule out respondent influences. Nonetheless, even if respondents had an impact we still show that recorded interviews proceed differently than non-recorded interviews. Second, the analysis of audio recording effects on substantive outcomes in the longitudinal setting was limited to experienced respondents who were already familiar with the respective items. In this situation, audio recordings are expected to have little effect. Third, our duration measurements could be affected by technical errors, for example, due to interviewers going back and forth during the interview and thus overwriting previous timestamps and generating extremely short durations. However, such errors should not exclusively occur for non-recorded interviews and are thus unlikely to influence the results.

Future research may overcome these limitations and replicate our analysis in other countries and populations where attitudes to being recorded may differ. In addition, results may differ in surveys with higher consent rates for audio recordings. Given the introduction of stricter privacy laws in many countries (e.g., the General Data Protection Regulation in the European Union), assessing differences between consenting and non-consenting respondents for collecting additional data during surveys will become increasingly relevant.

Literature

Ananthpur, K., de Jong, J., Sridharan, G., & Shrivastava, B. (2023). How good is your data? Challenges of ensuring data quality in a large-scale survey: Lessons from the Tamil Nadu Household Panel Survey (TNHPS). *MIDS Working Paper*, *243*.

Arceneaux, T. (2007). Evaluating the computer audio-recorded interviewing (CARI) Household Wellness Study (HWS) field test. *Proceedings of the Survey Research Method Section, American Statistical Association*, 2811–2818.

Bergmann, M., & Bristle, J. (2020). Reading fast, reading slow: The effect of interviewers' speed in reading introductory texts on response behavior. *Journal of Survey Statistics and Methodology*, *8*(2), 325–351.

Biemer, P., Herget, D., Morton, J., Willis, G., & Box, P. O. (2000). The feasibility of monitoring field interview performance using computer audio recorded interviewing (CARI). *Proceedings of the Survey Research Method Section, American Statistical Association*, 1068–1073.

Biewen, M., Fitzenberger, B., & Rümmele, M. (2022). Using distribution regression difference-in-differences to evaluate the effects of a minimum wage introduction on the distribution of hourly wages and hours worked. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4219406

Billiet, J., & Loosveldt, G. (1988). Improvement of the quality of responses to factual survey questions by interviewer training. *Public Opinion Quarterly*, *52*(2), 190–211.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. *Sociological Methodology*, *12*(1981), 389–437.

Chernozhukov, V., Fernandez-Val, I., & Galichon, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, *96*(3), 559–575.

Chernozhukov, V., Fernández-Val, I., & Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, *81*(6), 2205–2268.

Cho, Y. I., Fuller, A., File, T., Holbrook, A. L., & Johnson, T. P. (2006). Culture and survey question answering: A behavior coding approach. *Proceedings of the Survey Research Method Section, American Statistical Association*, 4082–4089.

Couper, M. P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 271–286.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Zhang, C. (2013). The design of grids in web surveys. *Social Science Computer Review*, *31*(3), 322–345.

DeMatteis, J. M., Young, L. J., Dahlhamer, J., Langley, R. E., Murphy, J., Olson, K., & Sharma, S. (2020). *Falsification in surveys*. AAPOR. https://aapor.org/wp-content/uploads/2022/11/AAPOR_Data_Falsification_Task_Force_Report-updated.pdf

D'Haultfœuille, X., & Février, P. (2020). The provision of wage incentives: A structural estimation using contracts variation. *Quantitative Economics*, *11*(1), 349–397.

Edwards, B., Maitland, A., & Connor, S. (2017). Measurement error in survey operations management. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total survey error in practice* (pp. 253–277). John Wiley & Sons, Inc.

Fee, H., Fields, J., & Marlay, M. (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 Washington, D.C.*, 1–20.

Fee, H., Welton, T. A., Marlay, M., & Fields, J. (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.*

Fowler, F., & Mangione, T. (1990). *Standardized survey interviewing: Minimizing interviewer-related error.* SAGE Publications, Inc.

Garbarski, D., Dykema, J., Cate Schaeffer, N., & Edwards, D. F. (2020). Response times as an indicator of data quality: Associations with question, interviewer, and respondent characteristics in a health survey of diverse respondents. In K. Olson, J. D. Smyth, J. Dykema, A. Holbrook, F. Kreuter, & B. T. West (Eds.), *Interviewer effects from a total survey error perspective* (pp. 253–266). Chapman and Hall/CRC.

Garbarski, D., Schaeffer, N. C., & Dykema, J. (2016). Interviewing practices, conversational practices, and rapport: Responsiveness and engagement in the standardized survey interview. *Sociological Methodology*, *46*(1), 1–38.

Graybill, W., Chen, M., Chernozhukov, V., Fernandez-Val, I., & Galichon, A. (2016). *Rearrangement: Monotonize point and interval functional estimates by rearrangement.* https://CRAN.R-project.org/package=Rearrangement

Hicks, W. D., Edwards, B., Tourangeau, K., McBride, B., Harris-Kojetin, L. D., & Moss, A. J. (2010). Using CARI tools to understand measurement error. *Public Opinion Quarterly*, *74*(5), 985–1003.

Jesske, B. (2013). Concepts and practices in interviewer qualification and monitoring. In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewers' deviations in surveys – impact, reasons, detection and prevention* (pp. 91–116). Peter Lang, Academic Research.

Jesske, B., & Schulz, S. (2021). Panel Arbeitsmarkt und Soziale Sicherung PASS-Erhebungswelle 14 – 2020 (Haupterhebung). *FDZ-Methodenreport 6.*

Kelley, J. (2020a). Accuracy and utility of using paradata to detect question-reading deviations. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, & B. T. West (Eds.), *Interviewer effects from a total survey error perspective* (pp. 267–278). Chapman and Hall/CRC.

Kelley, J. (2020b). *Understanding question-reading deviations: Implications for monitoring interviewers, questionnaire design, and data quality* [Doctoral dissertation, University of Essex]. https://repository.essex.ac.uk/30339/

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail – web mixed-mode surveys. *Social Science Computer Review*, *37*(2), 214–233.

Kosyakova, Y., Skopek, J., & Eckman, S. (2015). Do interviewers manipulate responses to filter questions? Evidence from a multilevel approach. *International Journal of Public Opinion Research*, *27*(3), 417–431.

Loosveldt, G., & Beullens, K. (2017). Interviewer effects on non-differentiation and straightlining in the European Social Survey. *Journal of Official Statistics*, *33*(2), 409–426.

McGonagle, K. A., Brown, C., & Schoeni, R. F. (2015). The effects of respondents' consent to be recorded on interview length and data quality in a national panel study. *Field Methods*, *27*(4), 373–390.

Mitchell, S., Fahrney, K., & Strobl, M. (2009). Monitoring field interviewer and respondent interactions using computer-assisted recorded interviewing: A case study. *Proceedings of the Annual Conference of the American Association for Public Opinion Research (AAPOR)*, 5666–5675.

Mittereder, F., Durow, J., West, B. T., Kreuter, F., & Conrad, F. G. (2018). Interviewer– respondent interactions in conversational and standardized interviewing. *Field Methods*, *30*(1), 3–21.

Olbrich, L., Kosyakova, Y., Sakshaug, J. W., & Schwanhäuser, S. (2024). Detecting interviewer fraud using multilevel models. *Journal of Survey Statistics and Methodology*, *12*(1), 14–35.

Olson, K., & Smyth, J. D. (2015). The effect of CATI questions, respondents, and interviewers on response time. *Journal of Survey Statistics and Methodology*, *3*(3), 361–396.

Olson, K., Smyth, J. D., Dykema, J., Holbrook, A. L., Kreuter, F., & West, B. T. (2020). The past, present, and future of research on interviewer effects. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, & B. T. West (Eds.), *Interviewer effects from a total survey error perspective* (1st ed., pp. 3–16). Chapman and Hall/CRC.

Ongena, Y. P., & Dijkstra, W. (2007). A model of cognitive processes and conversational principles in survey interview interaction. *Applied Cognitive Psychology*, *21*(2), 145–163.

Pascale, J. (2011). Using CARI and behavior coding to evaluate questionnaires. *Paper Presented at Federal Computer Assisted Survey Information Collection Workshop (FedCASIC), Washington, DC.*

Philipson, T., & Lawless, T. (1997). Multiple-output agency incentives in data production: Experimental evidence. *European Economic Review*, *41*(3-5), 961–970.

Sala, E., Knies, G., & Burton, J. (2014). Propensity to consent to data linkage: Experimental evidence on the role of three survey design features in a UK longitudinal panel. *International Journal of Social Research Methodology*, *17*(5), 455–473.

Schober, M. F., Conrad, F. G., Dijkstra, W., & Ongena, Y. P. (2012). Disfluencies and gaze aversion in unreliable responses to survey questions. *Journal of Official Statistics*, *28*(4), 555–582.

Sirkis, R. (2013). Impact of the 2012 computer audio recorded interviewing application on Survey of Income and Program Participation Event History Calendar response rates and item-level responses. *Proceedings of the Survey Research Method Section, American Statistical Association*, 3070–3084.

Smith, T. W. (2009). An analysis of computer assisted recorded interviews (CARI) on the 2008 General Social Survey. *GSS Methodology Report*, *117*.

Sturgis, P., & Luff, R. (2015). Audio-recording of open-ended survey questions: A solution to the problem of interviewer transcription. In U. Engel (Ed.), *Survey measurements: Techniques, data quality and sources of error* (pp. 42–57). University of Chicago Press.

Sturgis, P., Maslovskaya, O., Durrant, G., & Brunton-Smith, I. (2021). The interviewer contribution to variability in response times in face-to-face interview surveys. *Journal of Survey Statistics and Methodology*, *9*(4), 701–721.

Suchman, L., & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association*, *85*(409), 232–241.

Sun, C., Guo, B., Liu, X., Xiao, X., & Zhao, X. (2022). Interviewer error within the face-to-face food frequency questionnaire in large multisite epidemiologic studies. *American Journal of Epidemiology*, *191*(5), 921–925.

Thissen, M. R. (2014). Computer audio-recorded interviewing as a tool for survey research. *Social Science Computer Review*, *32*(1), 90–104.

Thissen, M. R., & Myers, S. K. (2016). Systems and processes for detecting interviewer falsification and assuring data collection quality. *Statistical Journal of the IAOS*, *32*(3), 339–347.

Trappmann, M., Bähr, S., Beste, J., Eberl, A., Frodermann, C., Gundert, S., Schwarz, S., Teichler, N., Unger, S., & Wenzig, C. (2019). Data resource profile: Panel study labour market and social security (PASS). *International Journal of Epidemiology*, *48*(5), 1411–1411G.

Trappmann, M., Beste, J., Bethmann, A., & Müller, G. (2013). The PASS panel survey after six waves. *Journal for Labour Market Research*, *46*(4), 275–281.

Uhrig, S. N., & Sala, E. (2011). When change matters: An analysis of survey interaction in dependent interviewing on the British Household Panel Study. *Sociological Methods & Research*, *40*(2), 333–366.

West, B. T., Conrad, F. G., Kreuter, F., & Mittereder, F. (2018). Can conversational interviewing improve survey response quality without increasing interviewer effects? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(1), 181–203.

Wuyts, C., & Loosveldt, G. (2022). Observing interviewer performance in slices or by traces: A comparison of methods to predict interviewers' individual contributions to interviewer variance. *Survey Research Methods*, *16*(2), 147–163.

Yan, T. (2008). Nondifferentiation. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 520–521). SAGE Publications, Inc.

Appendix

3.A Audio recording consent rates in CAPI surveys

Table 3.A1: Audio recording consent rates reported for CAPI surveys.

Source	Survey	Consent rate
Sirkis (2013)	Survey of Income and Program Participation 2012	41.68
Fee et al. (2015)	Survey of Income and Program Participation 2014	66.4
Kelley (2020a)	Understanding Society Innovation Panel, Wave 3	72
Mittereder et al. (2018)	Survey in Germany (2014) - Standardized interviewing	76.7
	Survey in Germany (2014) - Conversational interviewing	62.6
Hicks et al. (2010)	National Home and Hospice Care Survey 2007	96
Pascale (2011)	American Community Survey 2010	64.8
Arceneaux (2007)	Household Wellness Study 2004 - Philadelphia	83.8
	Household Wellness Study 2004 - Detroit	89.1
	Household Wellness Study 2004 - Kansas City	92.3
Biemer et al. (2000)	National Survey of Child and Adolescent Well-Being 1997 - caseworker	85
	National Survey of Child and Adolescent Well-Being 1997 - caregiver	83
	National Survey of Child and Adolescent Well-Being 1997 - child	82
Sala et al. (2014)	Understanding Society Innovation Panel, Wave 4	68.4
Ananthpur et al. (2023)	Tamil Nadu Household Panel Survey 2018	26
Sturgis and Luff (2015)	Wellcome Monitor Survey 2012	64.1
Mitchell et al. (2009)	Study of Community Family Life 2007	93
Smith (2009)	General Social Survey 2008	84
Cho et al. (2006)	Pilot study on respondent experiences, behaviors, and beliefs related to cancer and cancer screening in Chicago	79.8
Uhrig and Sala (2011)	British Household Panel Study Wave 16 pilot	72.1

3.B Covariates

Table 3.B1: Description of covariates.

Variable	Description
Age	Age in years
Education	Years of education
Month in field period	Continuous variable on the month in the field period (ranging from first to last month)
Panel experience	Continuous variable on the number of PASS participations (including the current wave)
Gender	Binary variable (0 = female, 1 = male)
German	Binary variable on whether the respondent has German citizenship or not
Unemployed	Binary variable on whether the respondent is unemployed or not.

3.C Difference between recorded and non-recorded interviews

Table 3.C1: Covariates by audio recording, CATI, wave 14.

		Recorded: No (N=1373)		Recorded: Yes (N=3048)	
		Mean	Std. Dev.	Mean	Std. Dev.
Age		43.08	14.28	44.47	14.14
Education		12.99	3.07	12.94	2.91
Month in field period		3.70	1.66	3.47	1.57
Panel experience		6.90	4.67	7.42	4.71
		N	Pct.	N	Pct.
Gender	Female	719	52.4	1545	50.7
	Male	654	47.6	1503	49.3
German	No	100	7.3	169	5.5
	Yes	1273	92.7	2879	94.5
Unemployed	No	1122	81.7	2570	84.3
	Yes	251	18.3	478	15.7

Table 3.C2: Covariates by audio recording, CAPI, face-to-face, wave 14.

		Recorded: No (N=1101)		Recorded: Yes (N=560)	
		Mean	Std. Dev.	Mean	Std. Dev.
Age		43.06	14.09	44.40	14.23
Education		12.02	2.80	12.15	2.88
Month in field period		1.67	0.60	1.66	0.67
Panel experience		7.08	4.03	7.49	4.03
		N	Pct.	N	Pct.
Gender	Female	581	52.8	307	54.8
	Male	520	47.2	253	45.2
German	No	182	16.5	21	3.8
	Yes	919	83.5	539	96.2
Unemployed	No	857	77.8	441	78.8
	Yes	244	22.2	119	21.2

Table 3.C3: Covariates by audio recording, CAPI-by-phone, wave 14.

		Recorded: No (N=731)		Recorded: Yes (N=350)	
		Mean	Std. Dev.	Mean	Std. Dev.
Age		41.31	13.21	42.37	13.34
Education		12.30	2.90	12.60	2.73
Month in field period		6.63	1.01	6.53	0.91
Panel experience		6.65	4.10	7.04	4.10
		N	Pct.	N	Pct.
Gender	Female	392	53.6	198	56.6
	Male	339	46.4	152	43.4
German	No	102	14.0	18	5.1
	Yes	629	86.0	332	94.9
Unemployed	No	587	80.3	294	84.0
	Yes	144	19.7	56	16.0

Table 3.C4: Social participation sample characteristics by audio recording, wave 10.

		Recorded: No (N=1370)		Recorded: Yes (N=795)	
		Mean	Std. Dev.	Mean	Std. Dev.
Age		43.85	13.13	45.63	13.08
Education		11.71	2.59	11.74	2.50
Month in field period		3.91	1.23	3.56	0.96
Panel experience		6.76	1.99	6.77	2.02
		N	Pct.	N	Pct.
Gender	Female	748	54.6	432	54.3
	Male	622	45.4	363	45.7
German	No	121	8.8	30	3.8
	Yes	1249	91.2	765	96.2
Unemployed	No	1077	78.6	604	76.0
	Yes	293	21.4	191	24.0

Table 3.C5: Life satisfaction sample characteristics by audio recording, wave 10.

		Recorded: No (N=1159)		Recorded: Yes (N=704)	
		Mean	Std. Dev.	Mean	Std. Dev.
Age		43.83	13.18	45.82	12.91
Education		11.80	2.61	11.88	2.60
Month in field period		3.87	1.19	3.59	0.97
Panel experience		6.77	1.98	6.75	2.03
		N	Pct.	N	Pct.
Gender	Female	639	55.1	386	54.8
	Male	520	44.9	318	45.2
German	No	105	9.1	29	4.1
	Yes	1054	90.9	675	95.9
Unemployed	No	928	80.1	543	77.1
	Yes	231	19.9	161	22.9

3.D Questionnaire text for W14 modules.

Table 3.D1: Questionnaire text for W14 modules.

Module		Text
Social Trust	1	We now come to the topic of trust. Speaking very generally, would you say that you can trust most people, or can you never be too careful when dealing with other people? "0" means that you can never be too careful with other people, "10" means that you can trust most people. You can grade your answer with the values in between.
Attitude To Life	1	Let us now deal with your life and your situation in general. How satisfied are you today with the following areas of your life? For your assessment you can use the numbers from "0" to "10". "0" means that you are "very dissatisfied", "10" means you are "very satisfied". The numbers "1" to "9" allow you to grade your assessment. How satisfied are you ... with your health
	2	with your apartment/house
	3	with your standard of living in general
Attitude To Self	1	Whenever unexpected difficulties or problems show up, there are different ways of reacting to that. We grouped some opinions about that topic here. Please tell me, whether to you those opinions "apply completely", "tend to apply", "tend not to apply" or "do not apply at all". I have a solution for every problem.
	2	Even when things happen surprisingly, I believe that I can cope with them.
	3	I have no difficulties in achieving my aims.
	4	I always know how to act in unforeseeable situations.
	5	I can always solve difficult problems if I try to.
Role Model	1	We have now completed all questions concerning your biography. Let's now talk about something completely different. I will read out some opinions about the relation of family and employment. Please tell me if you "completely agree", "agree", "rather disagree" or "strongly disagree". A woman should be ready to reduce her working hours to spend more time with her family.

Continued on next page

Table 3.D1 – *continued from previous page*

Module		Text
	2	It is rather nice to have a job, but what most women want is a home and family.
	3	A working mother can have an equally cordial relationship with her children as a stay at home mother.
	4	It is a husband's duty to earn money, the wife's duty to take care of home and family.
Politics	1	Generally speaking, how much are you interested in politics?
Democracy	1	On the whole, how satisfied are you with the way democracy works in Germany? The value "0" means you are entirely dissatisfied with how democracy works in Germany, the value "10" means you are entirely satisfied with it. You can grade your opinion using other values in between.
Left-Right	1	In politics, people often talk about "left" and "right" when describing different political views. When you think about your own political views, how would you rate them? The value "0" means: "far left", the value "10" means: "far right". You can grade your opinion with the values in between.
Activities	1	Are you actively engaged in one of the following organizations or associations? Union
	2	Political party
	3	Church community
	4	Clubs such as music, sport or culture clubs
	5	Another organization which I have not mentioned yet
Leisure	1	We now have a few questions about your leisure time. We are interested in what you do with people who do not live in the same household as you. I will go through a list of activities. Please tell me how often you do each of these: Go out with friends or acquaintances, for example to the cinema, to cafes, restaurants, pubs or clubs.
	2	Reciprocal visits with neighbours, friends or acquaintances.
	3	Stay in touch with friends and acquaintances by phone, email or via the internet.
	4	Attend sporting events with friends or acquaintances, for example football games or other competitions.
	5	Attend cultural events with friends or acquaintances, for example concerts, theatrical performances, exhibitions or museums.

Continued on next page

Table 3.D1 – ***continued from previous page***

Module		Text
	6	Go on trips or short journeys with friends or acquaintances.
Functions of work	1	The following questions relate to different aspects of your daily life. I will now read out a number of statements on this topic. Please tell me to what extent they apply to you personally. Please respond with one of the values from 1 "completely disagree" to 7 "completely agree". You can use the values in between to rate your opinion. I often feel that I make a meaningful contribution to society.
	2	often feel a valuable part of society.
	3	I hold a valuable position in society.
	4	I often meet new people.
	5	I often go out and meet with others.
	6	I usually have a lot of opportunities to mix with people.
	7	My friends usually value my company.
	8	I am often valued by the people around me.
	9	I am usually important to my friends.
	10	I often have nothing to do.
	11	I often wish I had more things to do to fill up the time in my days.
	12	There is usually too much spare time in my day.
	13	My days are usually well organized.
	14	I find it useful to structure my time.
	15	I have a good balance in my day between responsibilities and free time.
	16	I often have enough money to buy treats for myself.
	17	My income usually allows me to do the things I want.
	18	My level of income usually allows me to make plans for the future.
Insurance	1	What kind of health insurance do you have? Are you ...- in a statutory health insurance fund; - exclusively in a private health insurance fund; - in a private health insurance fund and eligible for additional allowances for public employees ("Beihilfe"); - Do you receive free provision of health services for civil servants ("freie Heilfuersorge")?; - Are you insured differently namely? (please indicate); - Or don't you have any health insurance?

3.E Distribution of module durations in wave 14

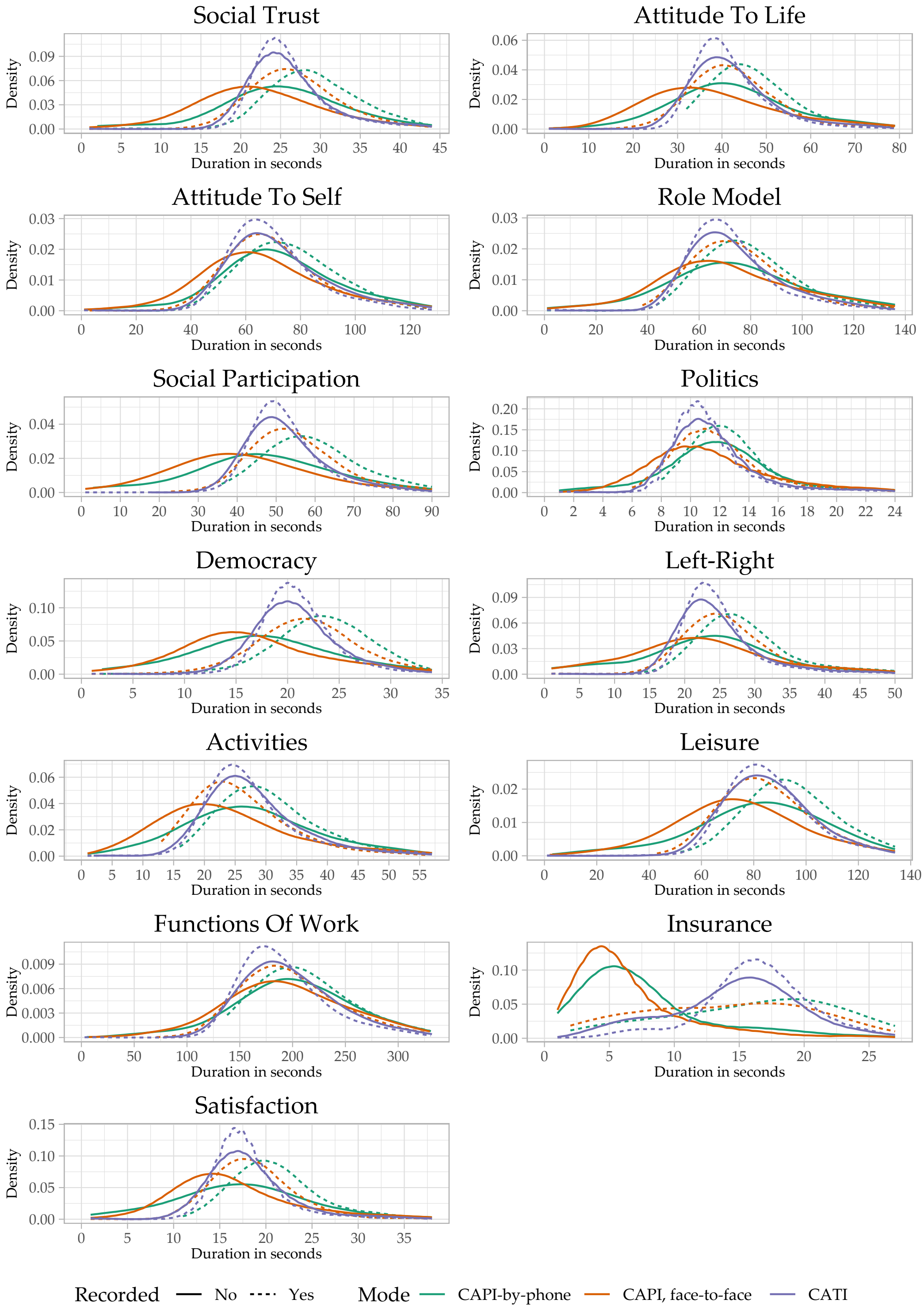


Figure 3.E1: Distributions of module durations in wave 14.

3.F Using different thresholds for excluding outliers

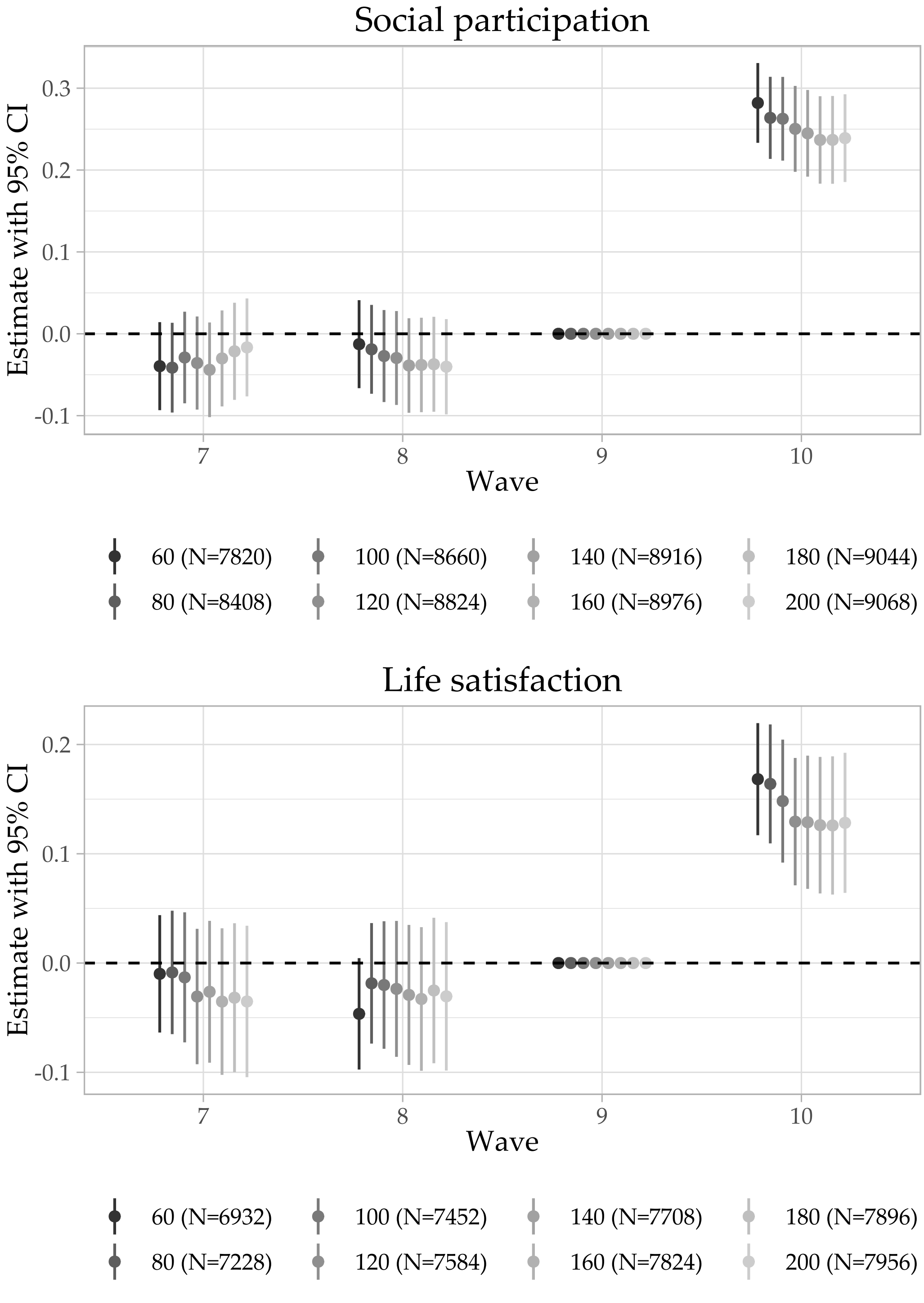


Figure 3.F1: Effects of the introduction of audio recordings when different thresholds (in seconds) are used for excluding outliers (with 95 percent confidence intervals).

3.G Common trends

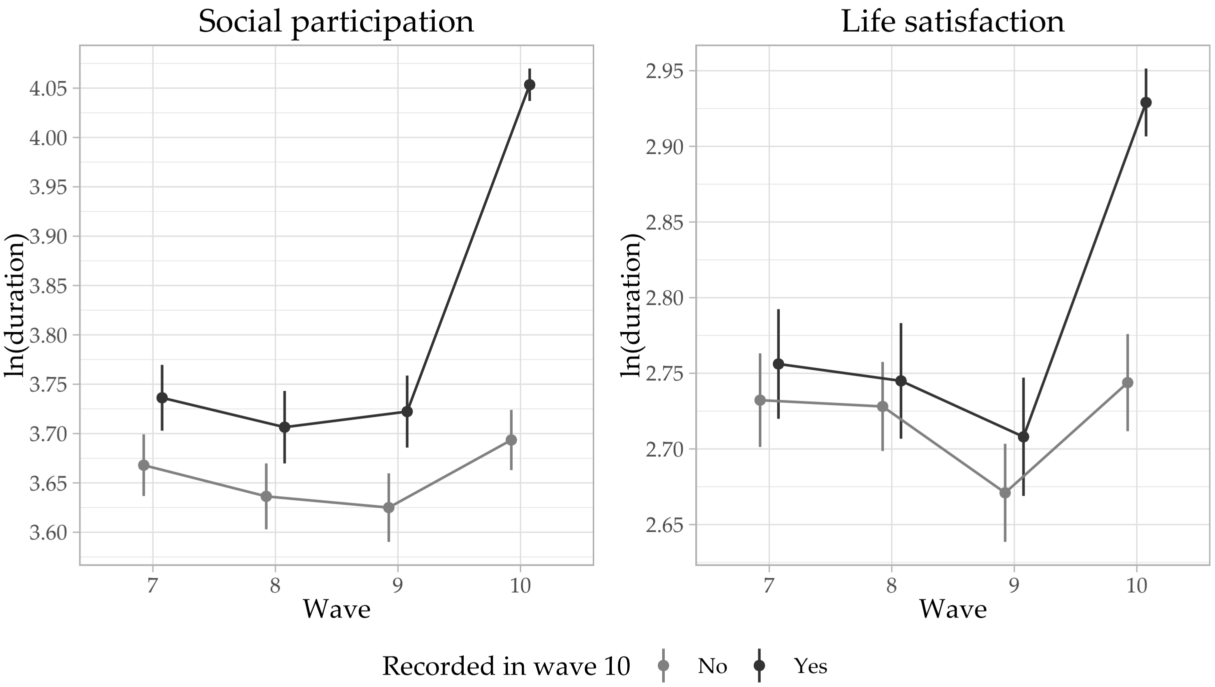


Figure 3.G1: Raw development of average of ln(duration) from waves 7 to 10 (with 95 percent confidence intervals).

3.H Differences between recorded and non-recorded interviews in the CATI sample

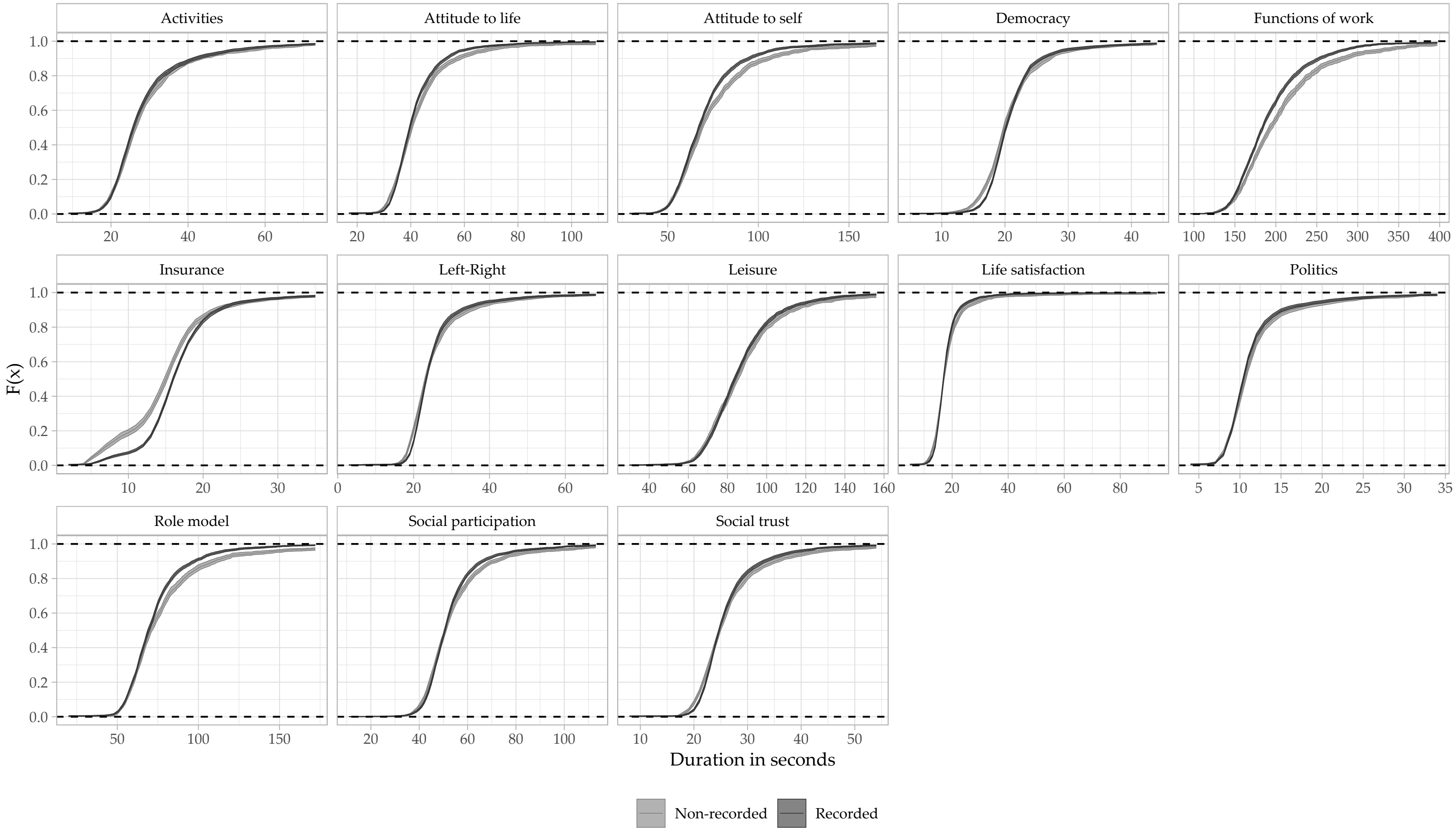


Figure 3.H1: Cumulative distributions of durations for recorded and non-recorded CATI interviews, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

3.I Results for further modules in Wave 14

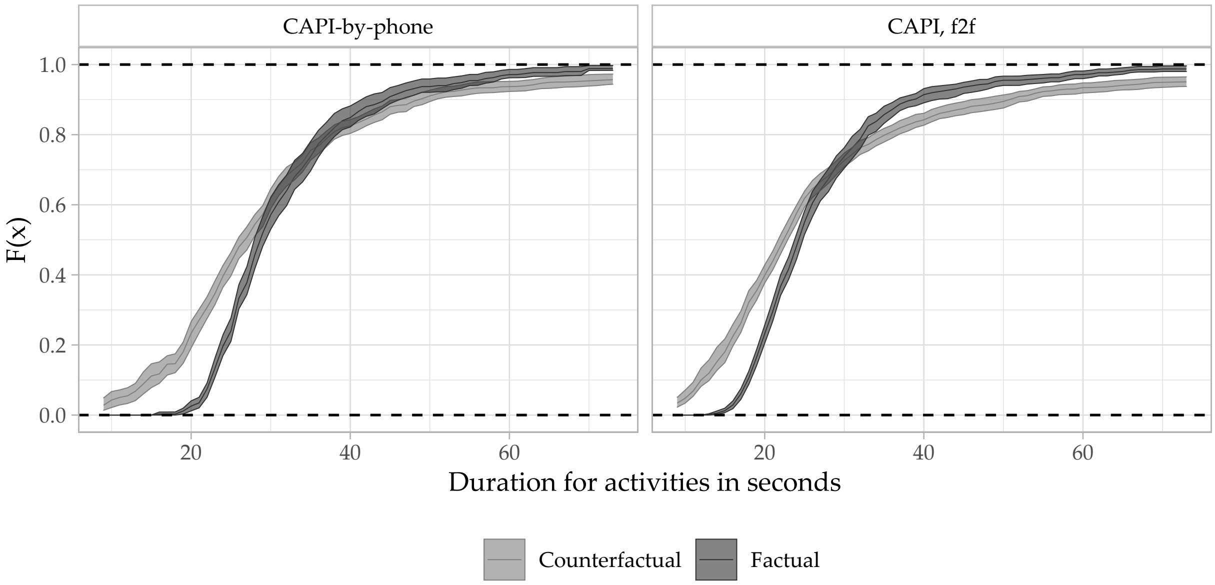


Figure 3.I1: Cumulative factual and counterfactual distribution of duration for activities for recorded interviews by CAPI groups, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

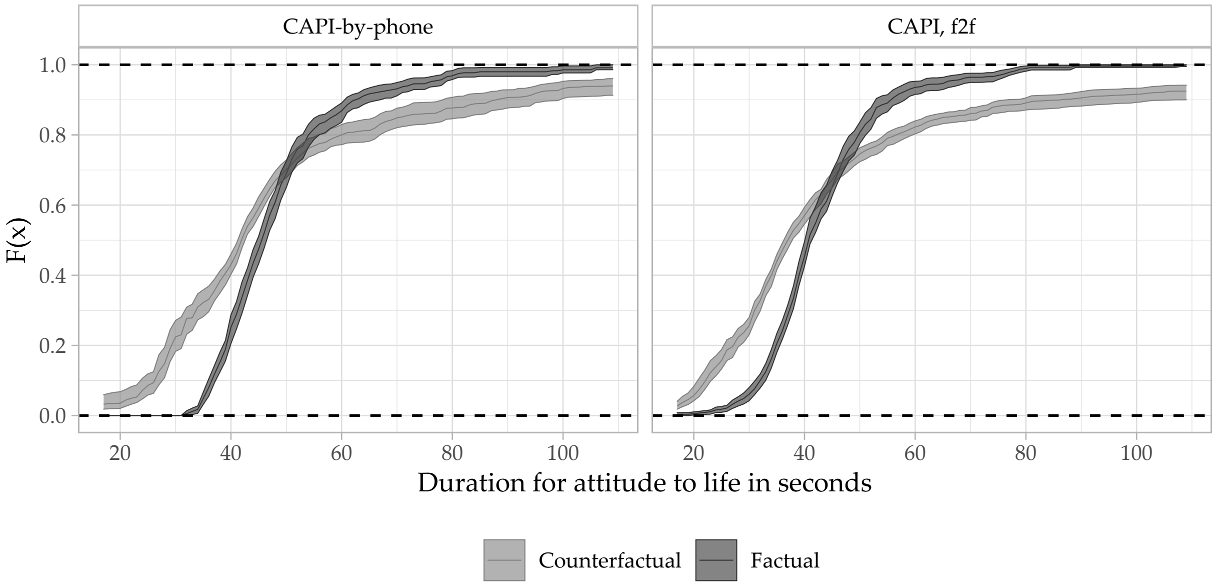


Figure 3.I2: Cumulative factual and counterfactual distribution of duration for attitudes to life for recorded interviews by CAPI groups, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

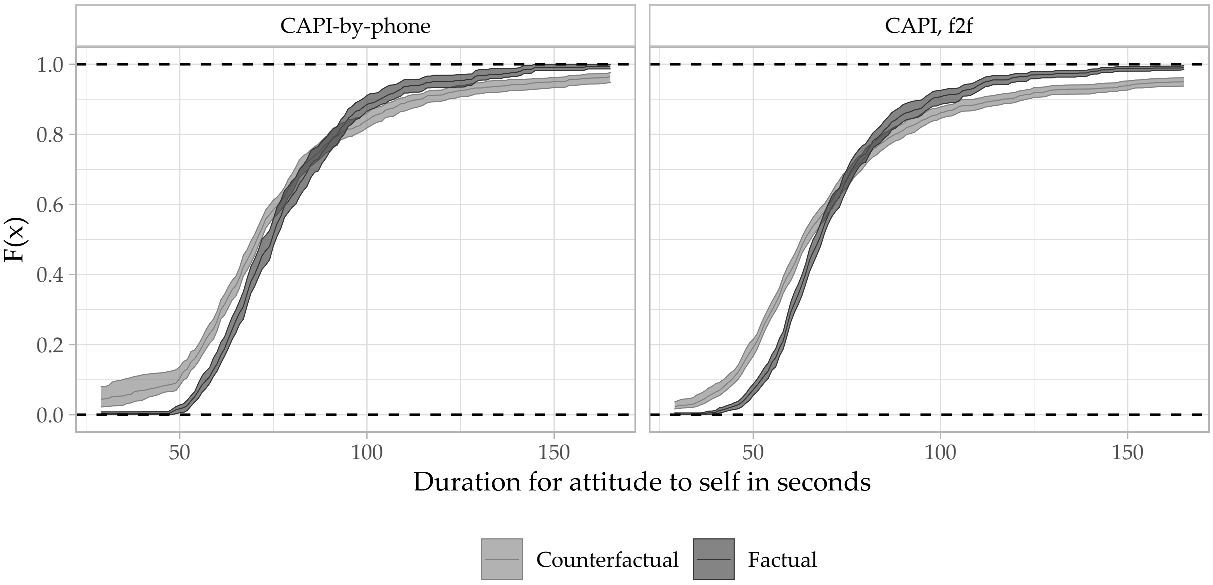


Figure 3.I3: Cumulative factual and counterfactual distribution of duration for attitudes to self for recorded interviews by CAPI groups, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

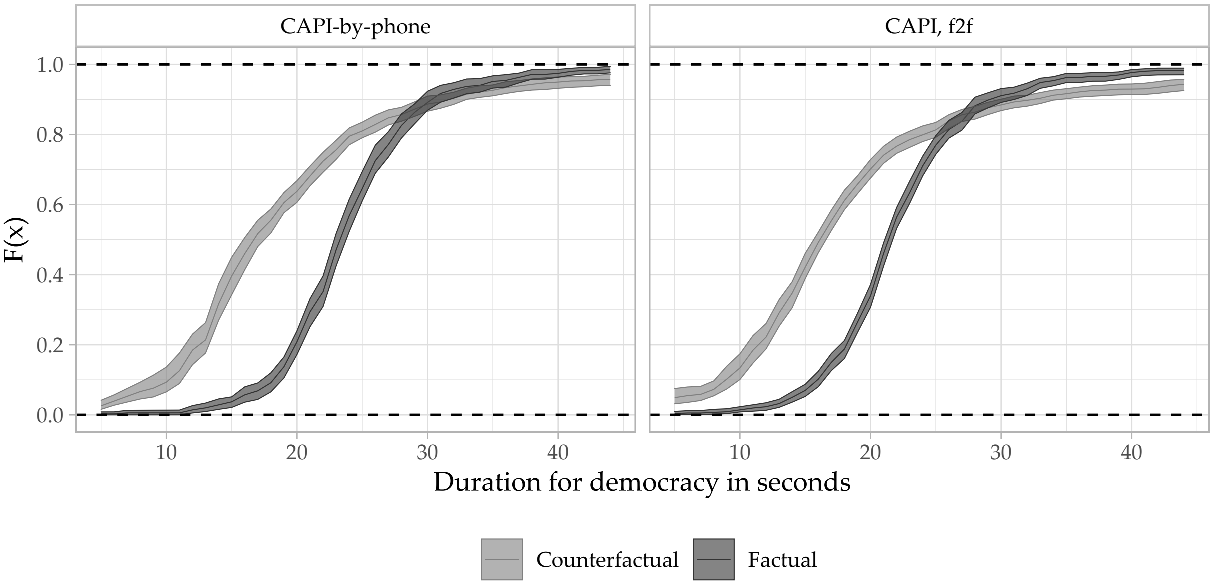


Figure 3.I4: Cumulative factual and counterfactual distribution of duration for democracy for recorded interviews by CAPI groups, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

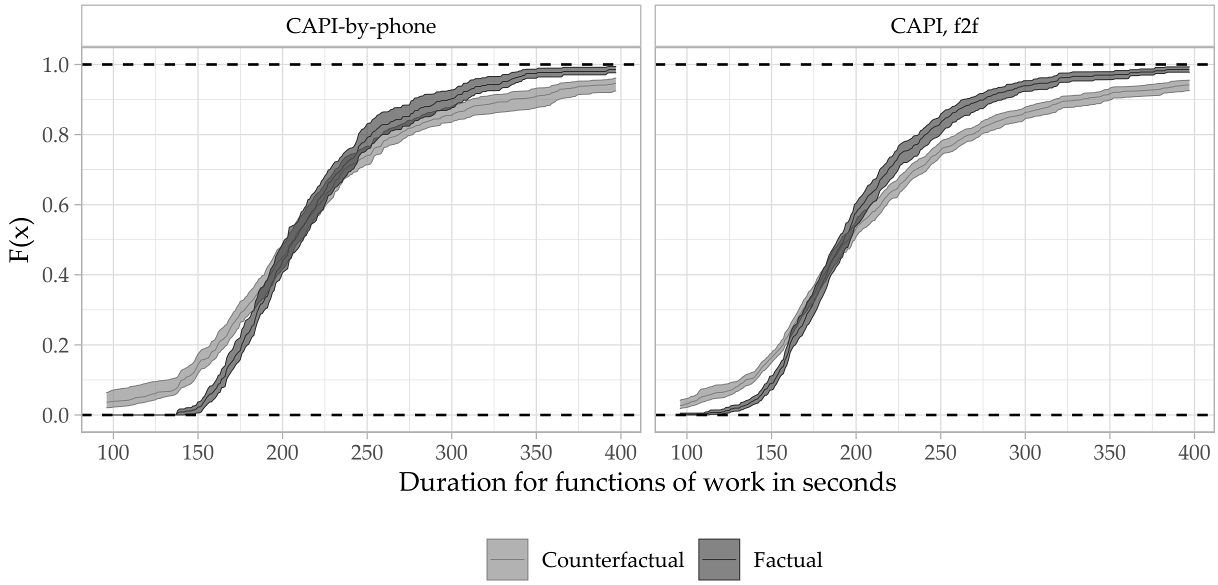


Figure 3.I5: Cumulative factual and counterfactual distribution of duration for functions of work for recorded interviews by CAPI groups, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

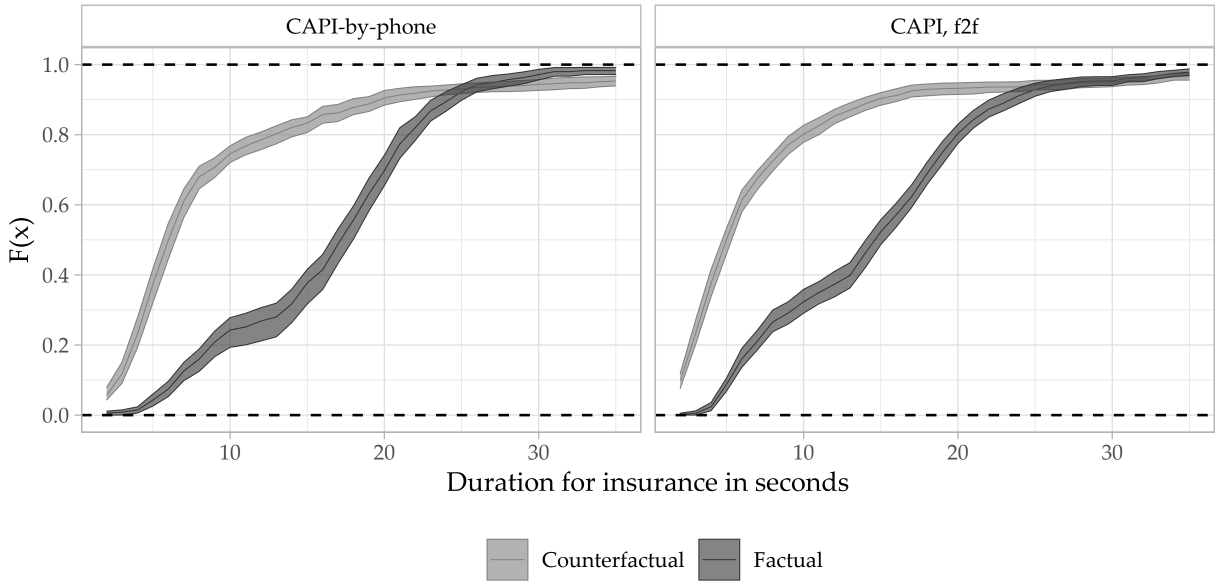


Figure 3.I6: Cumulative factual and counterfactual distribution of duration for insurance for recorded interviews by CAPI groups, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

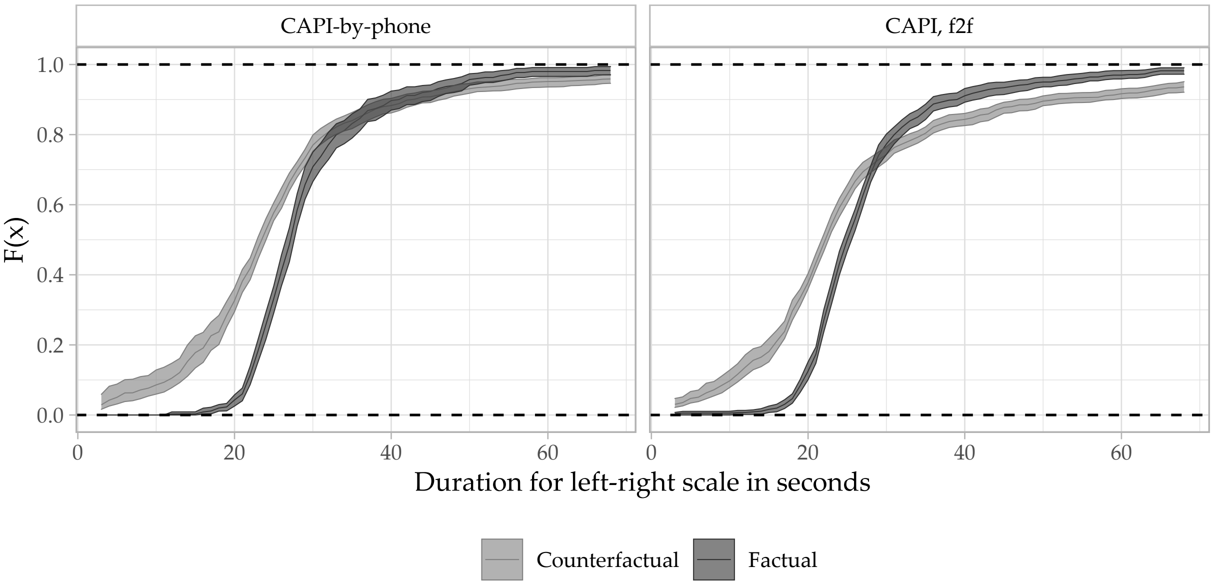


Figure 3.I7: Cumulative factual and counterfactual distribution of duration for left-right scale for recorded interviews by CAPI groups, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

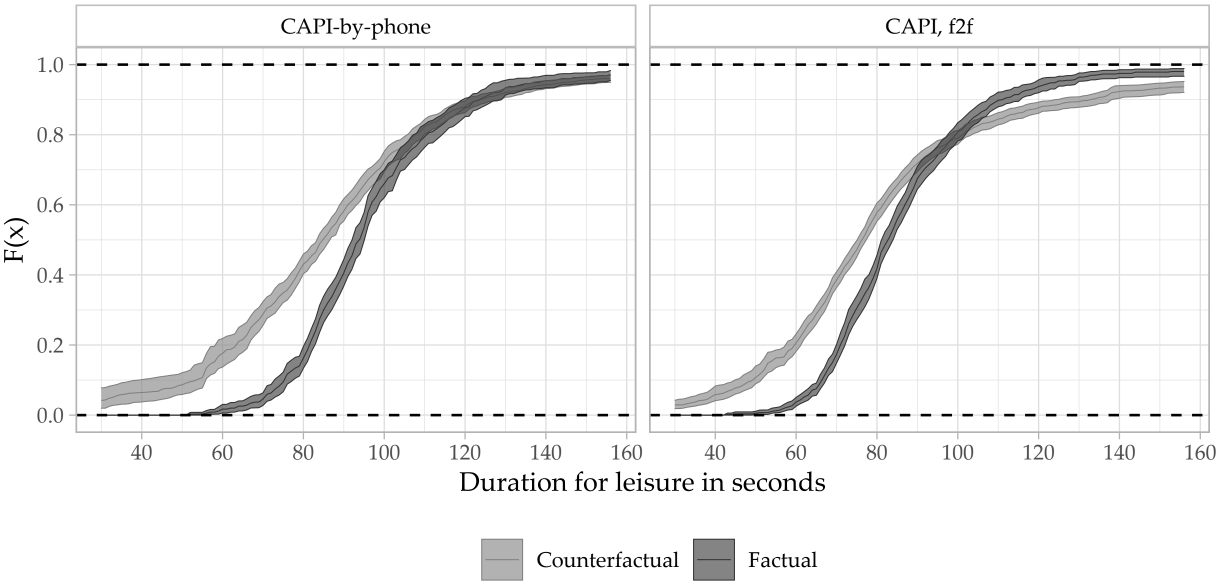


Figure 3.I8: Cumulative factual and counterfactual distribution of duration for leisure for recorded interviews by CAPI groups, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

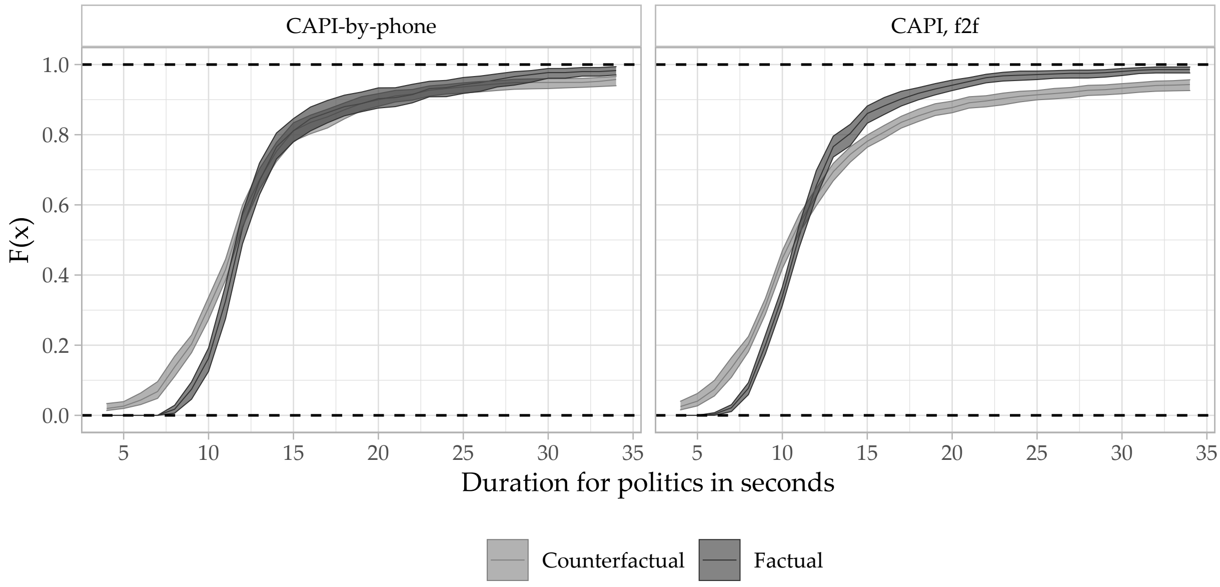


Figure 3.I9: Cumulative factual and counterfactual distribution of duration for politics for recorded interviews by CAPI groups, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

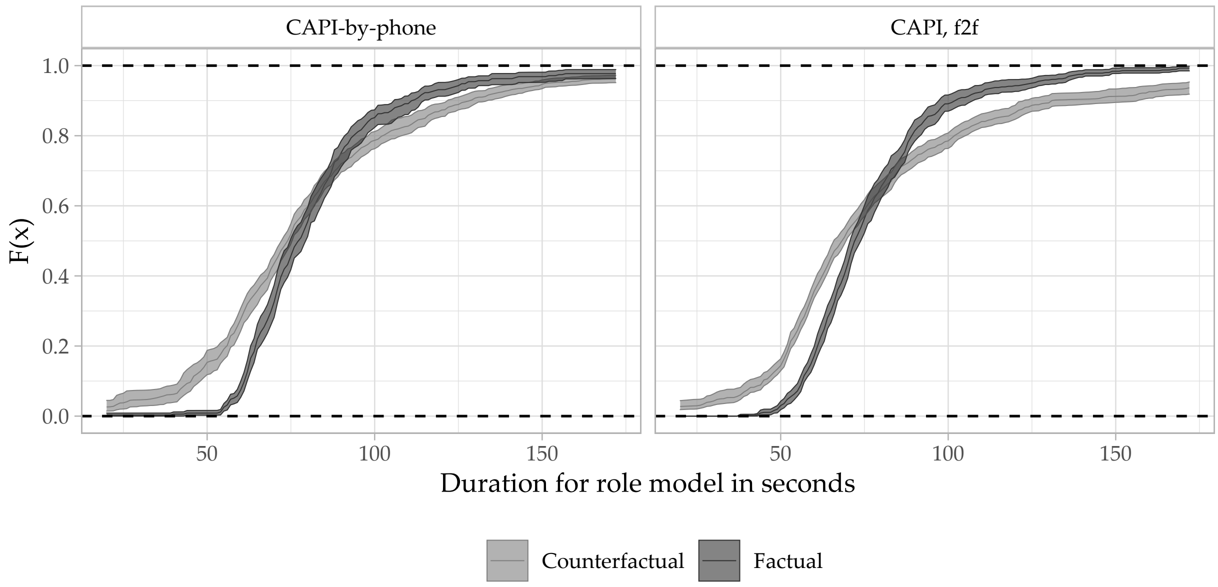


Figure 3.I10: Cumulative factual and counterfactual distribution of duration for role model for recorded interviews by CAPI groups, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

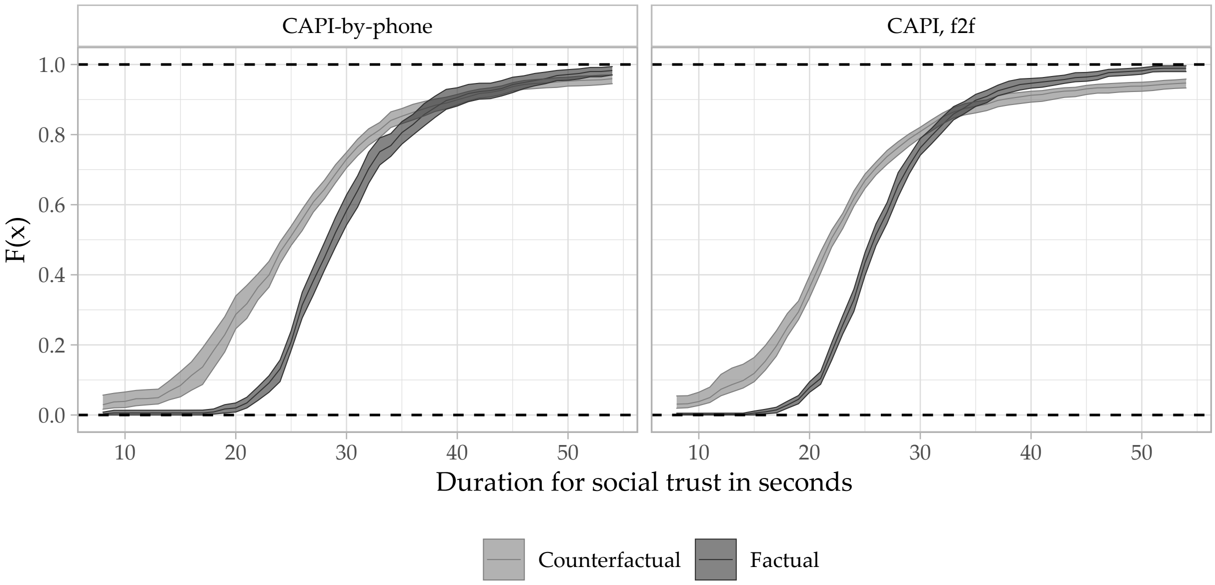


Figure 3.I11: Cumulative factual and counterfactual distribution of duration for social trust for recorded interviews by CAPI groups, wave 14. 95 percent confidence intervals based on 100 bootstrap replications.

3.J Development in wave 11

Table 3.J1: Participation, mode, and audio recording in Wave 11.

	Social participation sample				Life satisfaction sample			
	Not recorded		Recorded		Not recorded		Recorded	
	N	Percent	N	Percent	N	Percent	N	Percent
Participated in W11								
No	250	18.25	107	13.46	203	17.52	93	13.21
Yes	1120	81.75	688	86.54	956	82.48	611	86.79
CATI in W11								
No	1111	99.20	685	99.56	949	99.27	608	99.51
Yes	9	0.80	3	0.44	7	0.73	3	0.49
Recorded in W11								
No	926	82.68	244	35.47	780	81.59	225	36.82
Yes	194	17.32	444	64.53	176	18.41	386	63.18

3.K Effects on outcomes

Table 3.K1: Effects of audio recordings on measurement.

	Position in society (1)	Part of society (2)	Life satisfaction (3)
Treatment×Wave=7	-0.1137	-0.1241	-0.1229
	(0.0711)	(0.0926)	(0.0809)
Treatment×Wave=8	-0.0083	-0.1387	-0.1090
	(0.0663)	(0.0849)	(0.0756)
Treatment×Wave=10	-0.0817	0.0277	-0.0009
	(0.0694)	(0.0837)	(0.0725)
Fixed-effects			
Respondent	Yes	Yes	Yes
Wave	Yes	Yes	Yes
Observations	8,532	8,580	7,440
R^2	0.7053	0.6944	0.7123
Within R^2	0.0006	0.0009	0.0008

Notes: Clustered (Respondent) standard-errors in parentheses.
Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

3.L Participation in experiment in Wave 14

Table 3.L1: Questionnaire text for the experiment in Wave 14.

Variable	Text
Experiment	At the end of the interview, and in collaboration with our partners from the University of Zurich, we would like this year to scientifically examine when and under what conditions people trust each other. This part differs from the usual interview questions. We are interested in how you would behave in two specified decision-making situations. You will now decide whether you want to entrust money to another person. You will also decide whether you want to reciprocate someone else's trust. I will explain exactly how this works later with this card. After completing the survey, we will randomly draw one in ten participants and then actually pay out the amount resulting from the two decisions. Thus the amount of money you end up getting depends on your decisions and the decisions of the other person. You and the other person can receive up to 30 euros. This amount of money will be provided by our partners upon completion of the study and will be paid out by us. This will not affect the 10 euros which you will receive for your participation in the study. The other person is also a participant in this survey. This person will remain completely unknown to you. You will also remain completely unknown to this person. Your decisions will thus be treated in complete confidence. I will not know them either.

Table 3.L2: Experiment participation by duration and recording, PASS CAPI sample Wave 14.

		Not recorded		Recorded	
		N	Percent	N	Percent
Participation	No participation	181	16.47	21	3.78
	Participation	918	83.53	534	96.22
Duration	4 WPS or slower	828	75.34	539	97.12
	Faster than 4 WPS	271	24.66	16	2.88

Table 3.L3: Experiment participation by duration and recording crossings, PASS CAPI sample Wave 14.

Recording	Duration		No participation	Participation	All
Not recorded	4 WPS or slower	N	58	770	828
		% row	7.00	93.00	100.00
	Faster than 4 WPS	N	123	148	271
		% row	45.39	54.61	100.00
Recorded	4 WPS or slower	N	13	526	539
		% row	2.41	97.59	100.00
	Faster than 4 WPS	N	8	8	16
		% row	50.00	50.00	100.00
	All	N	202	1452	1654
		% row	12.21	87.79	100.00

4 Detecting interviewer fraud using multilevel models

Declaration of Contributions

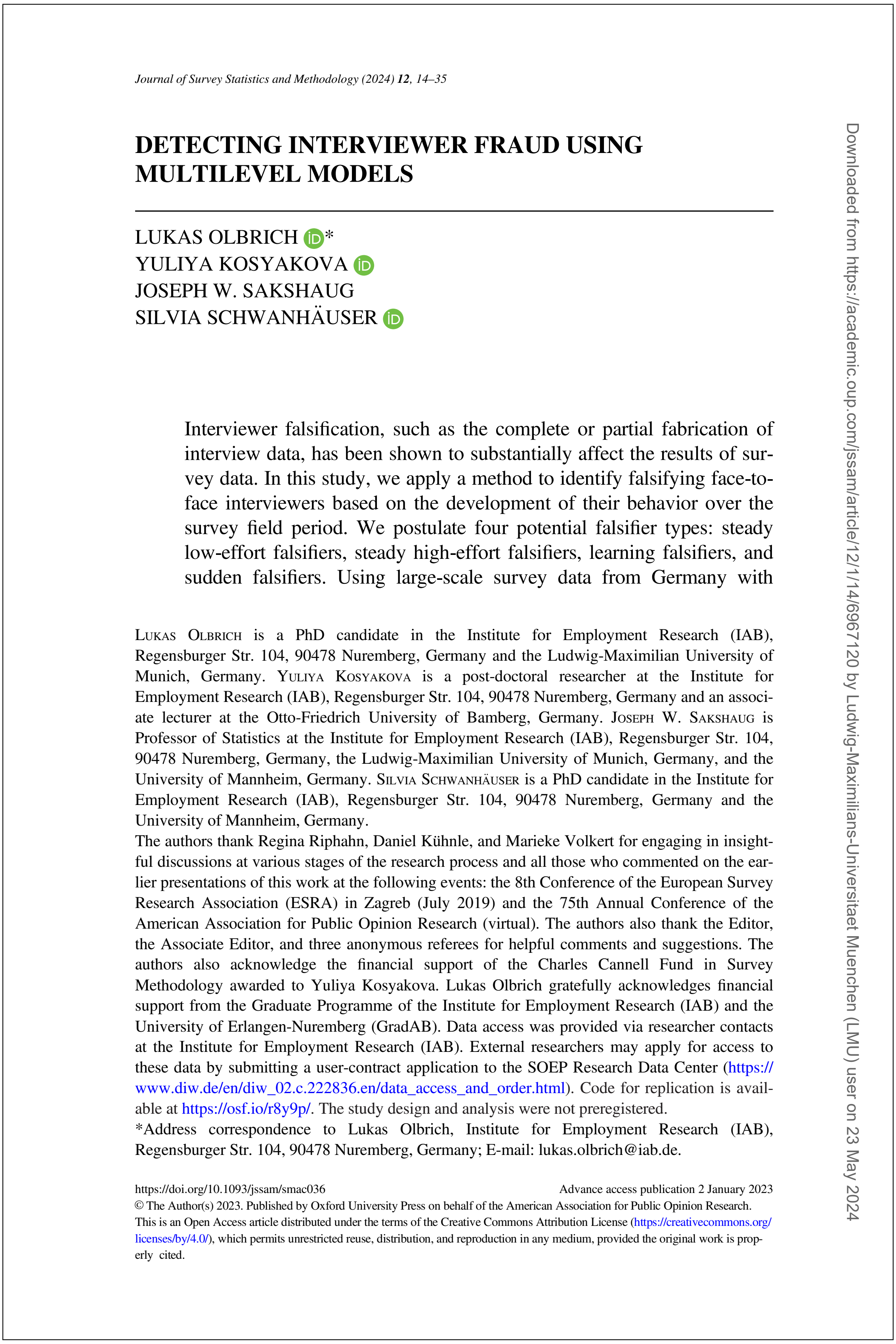
Lukas Olbrich had the idea of analyzing interviewer behaviors over the field period, implemented the analysis, and wrote the paper.

Contributions of Co-authors

All co-authors provided guidance and valuable input, and revised and proofread the paper.

Note

This article builds on Lukas Olbrich's master thesis. Compared to the master thesis, the analysis and modeling approach changed, other datasets were added to the analysis, other indicators were added, and the entire manuscript was rewritten.



Journal of Survey Statistics and Methodology (2024) ***12****, 14–35*

DETECTING INTERVIEWER FRAUD USING MULTILEVEL MODELS

LUKAS OLBRICH*
YULIYA KOSYAKOVA
JOSEPH W. SAKSHAUG
SILVIA SCHWANHÄUSER

Interviewer falsification, such as the complete or partial fabrication of interview data, has been shown to substantially affect the results of survey data. In this study, we apply a method to identify falsifying face-to-face interviewers based on the development of their behavior over the survey field period. We postulate four potential falsifier types: steady low-effort falsifiers, steady high-effort falsifiers, learning falsifiers, and sudden falsifiers. Using large-scale survey data from Germany with

LUKAS OLBRICH is a PhD candidate in the Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany and the Ludwig-Maximilian University of Munich, Germany. YULIYA KOSYAKOVA is a post-doctoral researcher at the Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany and an associate lecturer at the Otto-Friedrich University of Bamberg, Germany. JOSEPH W. SAKSHAUG is Professor of Statistics at the Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany, the Ludwig-Maximilian University of Munich, Germany, and the University of Mannheim, Germany. SILVIA SCHWANHÄUSER is a PhD candidate in the Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany and the University of Mannheim, Germany.

The authors thank Regina Riphahn, Daniel Kühnle, and Marieke Volkert for engaging in insightful discussions at various stages of the research process and all those who commented on the earlier presentations of this work at the following events: the 8th Conference of the European Survey Research Association (ESRA) in Zagreb (July 2019) and the 75th Annual Conference of the American Association for Public Opinion Research (virtual). The authors also thank the Editor, the Associate Editor, and three anonymous referees for helpful comments and suggestions. The authors also acknowledge the financial support of the Charles Cannell Fund in Survey Methodology awarded to Yuliya Kosyakova. Lukas Olbrich gratefully acknowledges financial support from the Graduate Programme of the Institute for Employment Research (IAB) and the University of Erlangen-Nuremberg (GradAB). Data access was provided via researcher contacts at the Institute for Employment Research (IAB). External researchers may apply for access to these data by submitting a user-contract application to the SOEP Research Data Center (https://www.diw.de/en/diw_02.c.222836.en/data_access_and_order.html). Code for replication is available at https://osf.io/r8y9p/. The study design and analysis were not preregistered.

*Address correspondence to Lukas Olbrich, Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany; E-mail: lukas.olbrich@iab.de.

https://doi.org/10.1093/jssam/smac036 Advance access publication 2 January 2023

verified falsifications, we apply multilevel models with interviewer effects on the intercept, scale, and slope of the interview sequence to test whether falsifiers can be detected based on their dynamic behavior. In addition to identifying a rather high-effort falsifier previously detected by the survey organization, the model flagged two additional suspicious interviewers exhibiting learning behavior, who were subsequently classified as deviant by the survey organization. We additionally apply the analysis approach to publicly available cross-national survey data and find multiple interviewers who show behavior consistent with the postulated falsifier types.

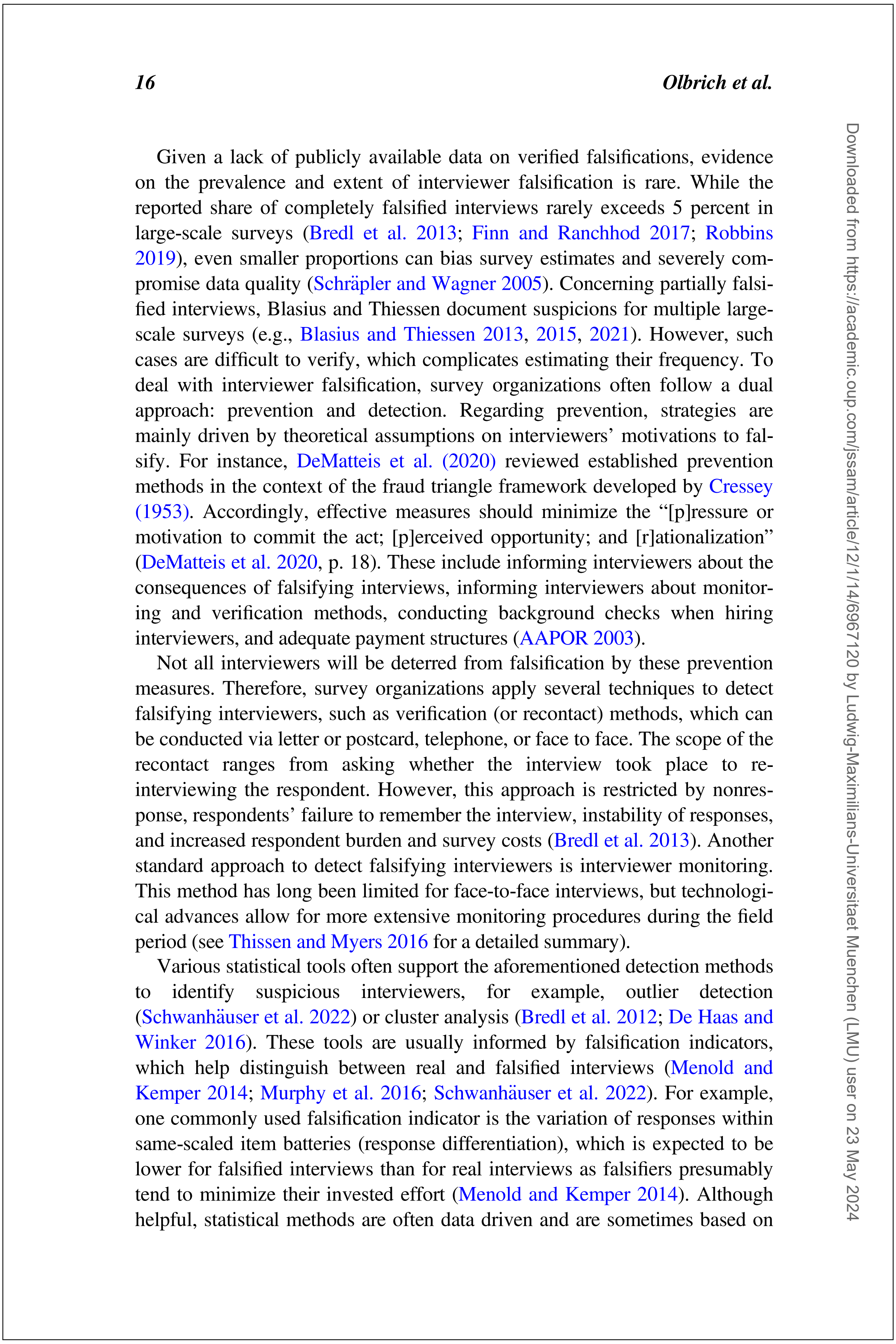
KEYWORDS: Interviewer behavior; Interviewer effects; Interviewer falsification; Multilevel modeling.

Statement of Significance

This study proposes a new method to identify fraudulent interviewers in face-to-face surveys. In particular, we investigate whether falsifying interviewers can be identified by their dynamic behavior over the field period. We postulate four falsifier types: steady low-effort falsifiers, steady high-effort falsifiers, learning falsifiers, and sudden falsifiers. These falsifier types are tested using complex multilevel models applied to German survey data containing verified cases of interviewer falsification. Focusing on the behavior over the field period allows for identifying a verified falsifier and two previously undetected fraudulent interviewers. Applying these methods to further publicly available survey data, we also find behavior expected of the postulated falsifier types. Our findings show that fraudulent interviewers can use sophisticated strategies to avoid detection.

1. INTRODUCTION

Interviewers are a well-known error source in face-to-face surveys. Researchers have intensively investigated unintentional interviewer errors such as accidentally skipping questions or recording responses in error (Weisberg 2005). Less is known about interviewer falsification—defined by the American Association for Public Opinion Research (AAPOR) as “the intentional departure from the designed interviewer guidelines or instructions, unreported by the interviewer, which could result in the contamination of data” (AAPOR 2003, p. 1). Interviewer falsification can take many forms, from strategically miscoding responses to avoid follow-up questions to falsifying complete or partial interviews (AAPOR 2003). In this study, we focus on the falsification of interviews.



Given a lack of publicly available data on verified falsifications, evidence on the prevalence and extent of interviewer falsification is rare. While the reported share of completely falsified interviews rarely exceeds 5 percent in large-scale surveys (Bredl et al. 2013; Finn and Ranchhod 2017; Robbins 2019), even smaller proportions can bias survey estimates and severely compromise data quality (Schräpler and Wagner 2005). Concerning partially falsified interviews, Blasius and Thiessen document suspicions for multiple large-scale surveys (e.g., Blasius and Thiessen 2013, 2015, 2021). However, such cases are difficult to verify, which complicates estimating their frequency. To deal with interviewer falsification, survey organizations often follow a dual approach: prevention and detection. Regarding prevention, strategies are mainly driven by theoretical assumptions on interviewers' motivations to falsify. For instance, DeMatteis et al. (2020) reviewed established prevention methods in the context of the fraud triangle framework developed by Cressey (1953). Accordingly, effective measures should minimize the "[p]ressure or motivation to commit the act; [p]erceived opportunity; and [r]ationalization" (DeMatteis et al. 2020, p. 18). These include informing interviewers about the consequences of falsifying interviews, informing interviewers about monitoring and verification methods, conducting background checks when hiring interviewers, and adequate payment structures (AAPOR 2003).

Not all interviewers will be deterred from falsification by these prevention measures. Therefore, survey organizations apply several techniques to detect falsifying interviewers, such as verification (or recontact) methods, which can be conducted via letter or postcard, telephone, or face to face. The scope of the recontact ranges from asking whether the interview took place to re-interviewing the respondent. However, this approach is restricted by nonresponse, respondents' failure to remember the interview, instability of responses, and increased respondent burden and survey costs (Bredl et al. 2013). Another standard approach to detect falsifying interviewers is interviewer monitoring. This method has long been limited for face-to-face interviews, but technological advances allow for more extensive monitoring procedures during the field period (see Thissen and Myers 2016 for a detailed summary).

Various statistical tools often support the aforementioned detection methods to identify suspicious interviewers, for example, outlier detection (Schwanhäuser et al. 2022) or cluster analysis (Bredl et al. 2012; De Haas and Winker 2016). These tools are usually informed by falsification indicators, which help distinguish between real and falsified interviews (Menold and Kemper 2014; Murphy et al. 2016; Schwanhäuser et al. 2022). For example, one commonly used falsification indicator is the variation of responses within same-scaled item batteries (response differentiation), which is expected to be lower for falsified interviews than for real interviews as falsifiers presumably tend to minimize their invested effort (Menold and Kemper 2014). Although helpful, statistical methods are often data driven and are sometimes based on contradictory theories regarding the expected direction of some falsification indicators.

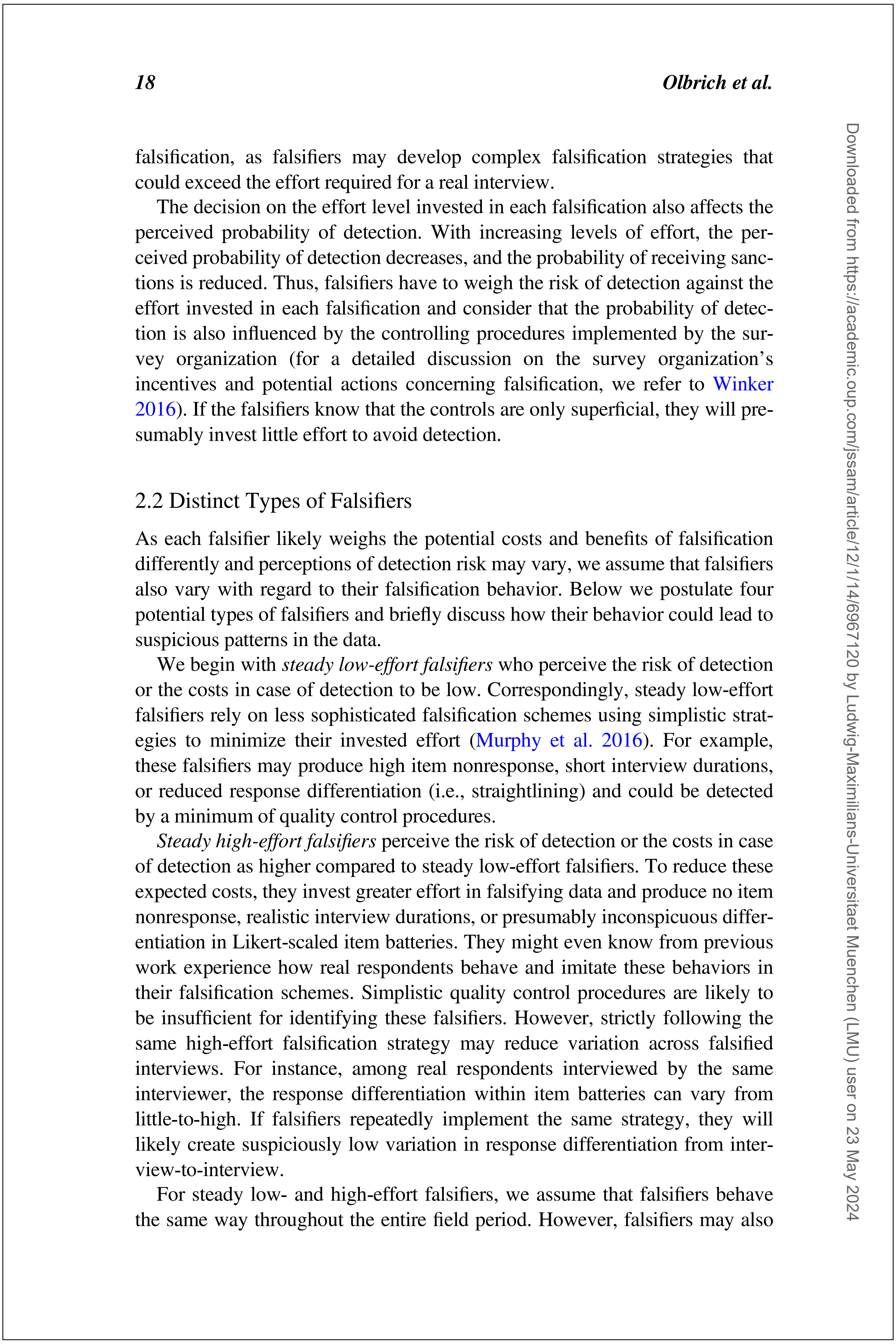
In this study, we investigate whether falsifying interviewers can be identified by their dynamic behavior over the field period. We postulate four distinct falsifier types whom we term as steady low-effort falsifiers, who use simplistic falsification strategies; steady high-effort falsifiers, who rely on complex strategies that require more effort; learning falsifiers, who adapt their behavior over the field period; and sudden falsifiers, who abruptly switch from honest interviewing to falsification during the field period. We argue that they can be distinguished from honest interviewers as their strategies generate suspicious patterns in the data.

We use data from a large-scale survey of refugees in Germany containing verified falsifications to test whether falsifiers indeed follow the postulated strategies. Using response differentiation as an approximation of falsification effort, we employ a multilevel model with interviewer effects on the intercept, the slope of the interview sequence, and the scale. To evaluate the occurrence of the falsifier types in other survey settings, we also apply the model to cross-national survey data of the general population.

2. THEORETICAL FRAMEWORK

2.1 Interviewer Falsification as Rational Behavior

Falsifiers have often been characterized as rational actors who assess their actions' expected costs and benefits to make a decision (Kennickell 2015; Kosyakova et al. 2015; Blasius and Thiessen 2021). Expected costs of falsification include sanctions such as job loss or legal consequences. The latter is rarely relevant as it is complex to provide conclusive proof of falsification, and survey organizations seek to avoid publicity on such delicate cases (Winker 2016; Blasius and Thiessen 2021). These costs only arise if the falsification is detected. Thus, if the perceived probability of detection is low, the expected costs will also be lower. Concerning the expected benefits, falsifiers can save time as it is faster to falsify an interview than to conduct a real interview. Saving time is particularly relevant for widely used piece-rate payment schemes, where interviewers receive fixed amounts for each successfully conducted interview (Kosyakova et al. 2015; Josten and Trappmann 2016). Falsifying instead of conducting a real interview may also reduce interviewer burden and thus the cognitive effort invested in each case. Real interviews require demanding tasks such as convincing the respondent to participate, administering (potentially sensitive or awkward) questions, and recording responses (West and Blom 2017). However, whether falsifying indeed reduces the effort invested in each case depends on the effort invested in the



falsification, as falsifiers may develop complex falsification strategies that could exceed the effort required for a real interview.

The decision on the effort level invested in each falsification also affects the perceived probability of detection. With increasing levels of effort, the perceived probability of detection decreases, and the probability of receiving sanctions is reduced. Thus, falsifiers have to weigh the risk of detection against the effort invested in each falsification and consider that the probability of detection is also influenced by the controlling procedures implemented by the survey organization (for a detailed discussion on the survey organization's incentives and potential actions concerning falsification, we refer to Winker 2016). If the falsifiers know that the controls are only superficial, they will presumably invest little effort to avoid detection.

2.2 Distinct Types of Falsifiers

As each falsifier likely weighs the potential costs and benefits of falsification differently and perceptions of detection risk may vary, we assume that falsifiers also vary with regard to their falsification behavior. Below we postulate four potential types of falsifiers and briefly discuss how their behavior could lead to suspicious patterns in the data.

We begin with *steady low-effort falsifiers* who perceive the risk of detection or the costs in case of detection to be low. Correspondingly, steady low-effort falsifiers rely on less sophisticated falsification schemes using simplistic strategies to minimize their invested effort (Murphy et al. 2016). For example, these falsifiers may produce high item nonresponse, short interview durations, or reduced response differentiation (i.e., straightlining) and could be detected by a minimum of quality control procedures.

Steady high-effort falsifiers perceive the risk of detection or the costs in case of detection as higher compared to steady low-effort falsifiers. To reduce these expected costs, they invest greater effort in falsifying data and produce no item nonresponse, realistic interview durations, or presumably inconspicuous differentiation in Likert-scaled item batteries. They might even know from previous work experience how real respondents behave and imitate these behaviors in their falsification schemes. Simplistic quality control procedures are likely to be insufficient for identifying these falsifiers. However, strictly following the same high-effort falsification strategy may reduce variation across falsified interviews. For instance, among real respondents interviewed by the same interviewer, the response differentiation within item batteries can vary from little-to-high. If falsifiers repeatedly implement the same strategy, they will likely create suspiciously low variation in response differentiation from interview-to-interview.

For steady low- and high-effort falsifiers, we assume that falsifiers behave the same way throughout the entire field period. However, falsifiers may also adapt their behavior over time. Such *learning falsifiers* might behave like steady high-effort falsifiers in the beginning of the field period but adjust their estimate of the risk of detection after learning about the quality control procedures (or lack thereof) used by the survey institute. Learning falsifiers likely reduce their falsification effort to increase the benefits of falsification when they perceive the control procedures to be poor. Such falsifiers are characterized by steadily changing values used for quality control monitoring, such as response differentiation. Methods implemented for detecting steady low- or high-effort falsifiers may not work here, as higher-effort falsifications are mixed with lower-effort falsifications, depending on the learning pace. As another alternation of steady high-effort falsification behavior, learning falsifiers could also switch from fabricating parts of the interview to blatantly fabricating entire interviews.

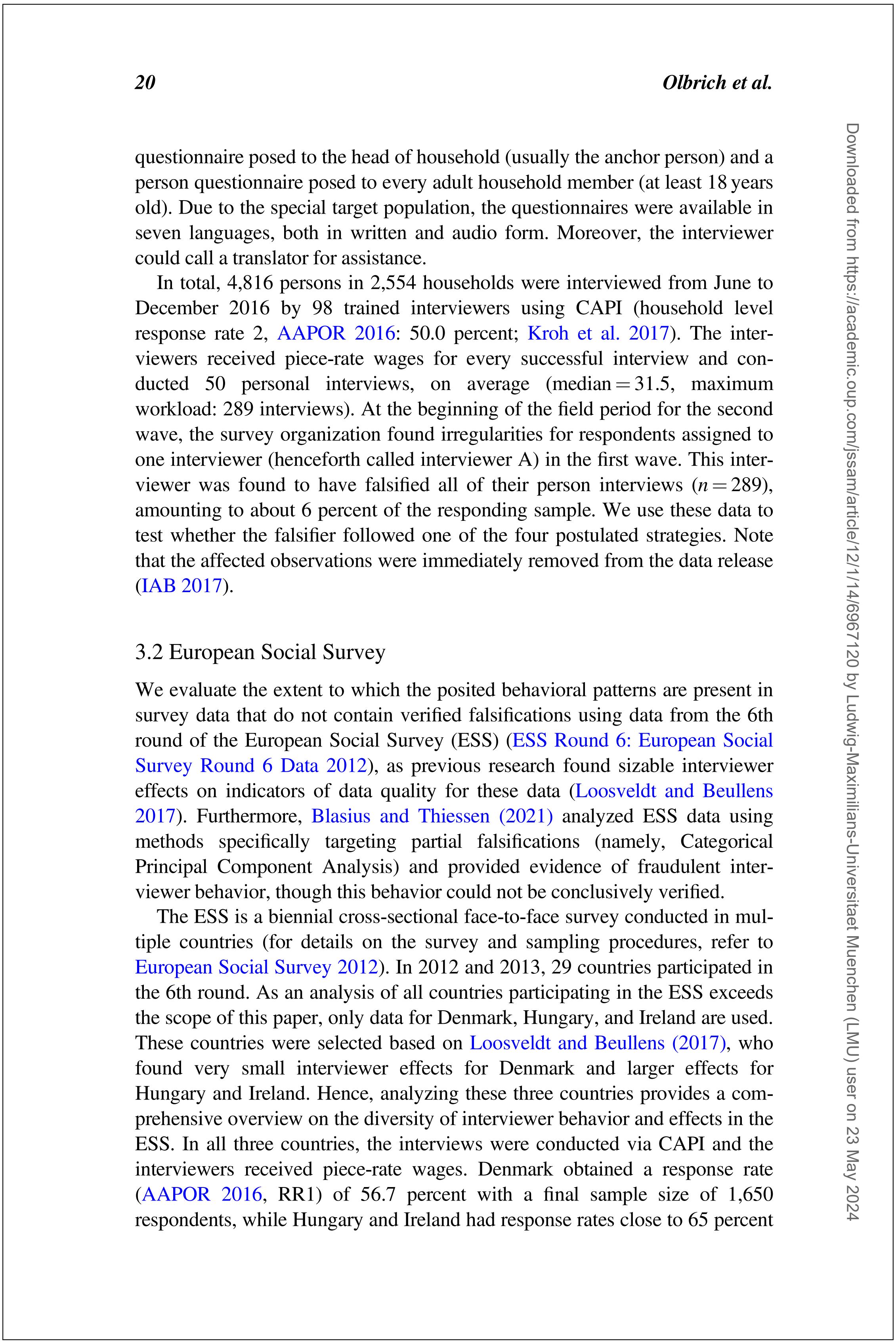
Lastly, some falsifiers may start falsifying at some point during the field period, for instance, because of being overwhelmed by their tasks or frustrated by the lack of respondent cooperation (Crespi 1945; Gwartney 2013). For such *sudden falsifiers*, the change point is ex ante unknown and their quality control values will resemble real interviews up until the switch to falsifying and then follow either the behavior of steady low- and high-effort or even learning falsifiers. Such falsifiers are characterized by changes in data quality measures and high variation in these measures due to the switch from interviewing to falsifying.

3. DATA

First, we test whether falsifiers follow the postulated strategies using large-scale survey data containing verified falsifications. Second, we use large-scale survey data to investigate the occurrence of the posited behaviors in other publicly available survey datasets. We note that the occurrence of real falsifications in the second data set is unknown, and thus, any possible detection of suspicious interviewers does not prove the prevalence of falsifications as this requires formal investigations.

3.1 IAB–BAMF–SOEP Survey of Refugees

The data containing verified falsifications come from the first wave of the IAB–BAMF–SOEP Survey of Refugees in Germany 2016 (version SOEP.v33) (Brücker et al. 2017). After the large influx of refugees in Germany in 2015 and 2016, the panel study was launched to gather information about this population. The multi-stage cluster sample was drawn from the Central Register of Foreign Nationals (*Ausländerzentralregister*; AZR). In addition to the selected anchor person, all household members older than 18 were interviewed, if possible. The interview consisted of a household



questionnaire posed to the head of household (usually the anchor person) and a person questionnaire posed to every adult household member (at least 18 years old). Due to the special target population, the questionnaires were available in seven languages, both in written and audio form. Moreover, the interviewer could call a translator for assistance.

In total, 4,816 persons in 2,554 households were interviewed from June to December 2016 by 98 trained interviewers using CAPI (household level response rate 2, AAPOR 2016: 50.0 percent; Kroh et al. 2017). The interviewers received piece-rate wages for every successful interview and conducted 50 personal interviews, on average (median = 31.5, maximum workload: 289 interviews). At the beginning of the field period for the second wave, the survey organization found irregularities for respondents assigned to one interviewer (henceforth called interviewer A) in the first wave. This interviewer was found to have falsified all of their person interviews ($n = 289$), amounting to about 6 percent of the responding sample. We use these data to test whether the falsifier followed one of the four postulated strategies. Note that the affected observations were immediately removed from the data release (IAB 2017).

3.2 European Social Survey

We evaluate the extent to which the posited behavioral patterns are present in survey data that do not contain verified falsifications using data from the 6th round of the European Social Survey (ESS) (ESS Round 6: European Social Survey Round 6 Data 2012), as previous research found sizable interviewer effects on indicators of data quality for these data (Loosveldt and Beullens 2017). Furthermore, Blasius and Thiessen (2021) analyzed ESS data using methods specifically targeting partial falsifications (namely, Categorical Principal Component Analysis) and provided evidence of fraudulent interviewer behavior, though this behavior could not be conclusively verified.

The ESS is a biennial cross-sectional face-to-face survey conducted in multiple countries (for details on the survey and sampling procedures, refer to European Social Survey 2012). In 2012 and 2013, 29 countries participated in the 6th round. As an analysis of all countries participating in the ESS exceeds the scope of this paper, only data for Denmark, Hungary, and Ireland are used. These countries were selected based on Loosveldt and Beullens (2017), who found very small interviewer effects for Denmark and larger effects for Hungary and Ireland. Hence, analyzing these three countries provides a comprehensive overview on the diversity of interviewer behavior and effects in the ESS. In all three countries, the interviews were conducted via CAPI and the interviewers received piece-rate wages. Denmark obtained a response rate (AAPOR 2016, RR1) of 56.7 percent with a final sample size of 1,650 respondents, while Hungary and Ireland had response rates close to 65 percent

(65.1 and 65.0) (AAPOR 2016, RR1) with final sample sizes of 2,014 and 2,628, respectively (European Social Survey 2012; Beullens et al. 2014). The average number of interviews conducted per interviewer was substantially lower in the ESS samples (Denmark: 15.6; Hungary: 13.0; Ireland: 22.3) than in the IAB–BAMF–SOEP Survey of Refugees.

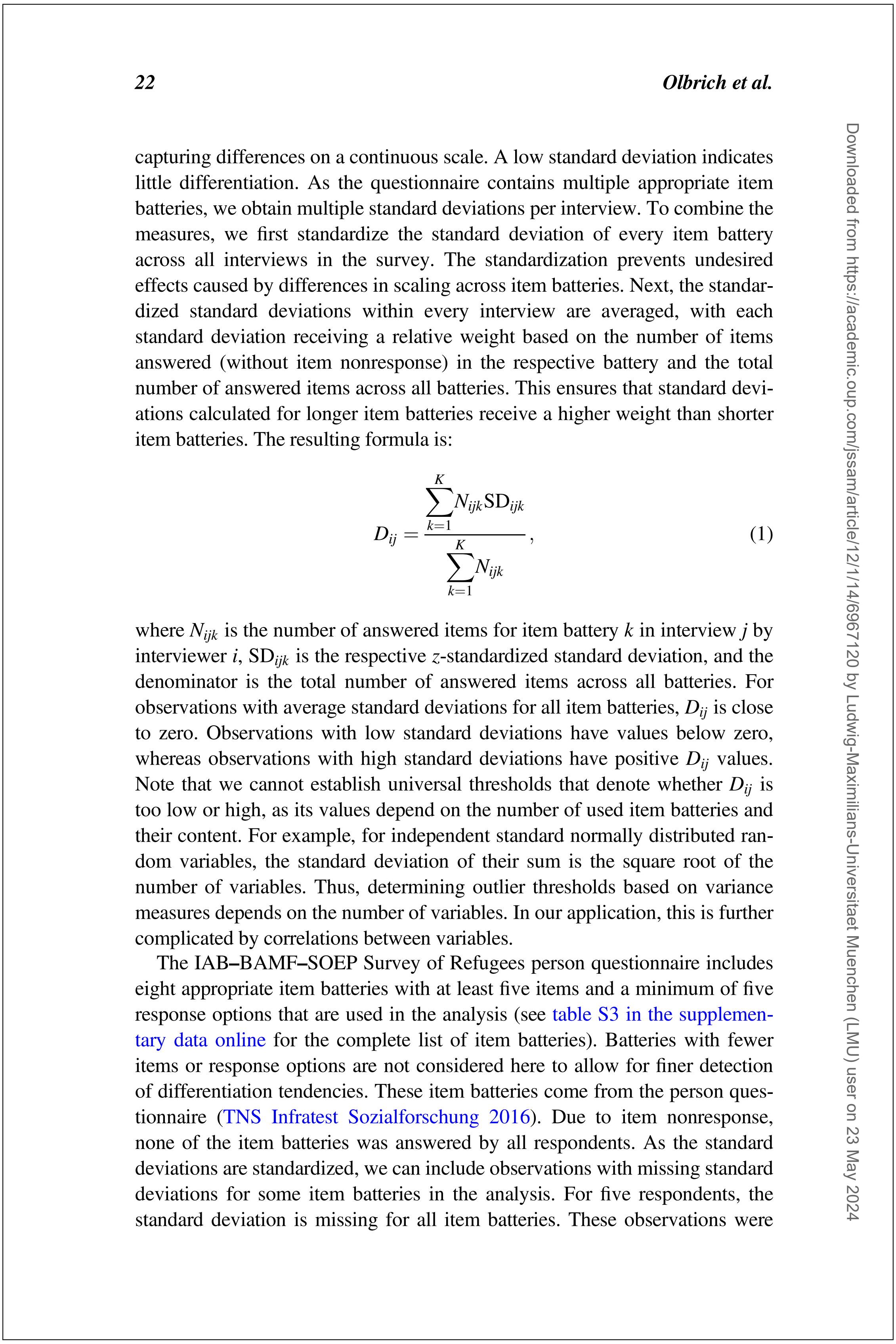
3.3 Dependent Variable

To approximate the falsifier's effort, we rely on a measure of response differentiation for item batteries using the same response scale (Yan 2008). We use response differentiation for two reasons. First, response differentiation is closely related to effort. Less response differentiation implies more similar responses, thereby reducing the cognitive effort of the answering process (Menold et al. 2013; Menold and Kemper 2014). Hence, less differentiation allows for faster completion of the questionnaire (and, in the case of the IAB–BAMF–SOEP Survey of Refugees and the ESS, a higher hourly interviewer wage), if the interviewer chooses the same response options regardless of their content. Second, the IAB–BAMF–SOEP Survey of Refugees questionnaire contains many long item batteries distributed over the entire questionnaire; thus, a measure based on these items will likely provide more detailed insights on the falsifiers' strategy than measures based on few questions in specific sections of the questionnaire.

Low response differentiation implies saving time and effort, thereby increasing the expected benefits of falsification. At the same time, lack of response differentiation may increase the perceived probability of detection as reduced differentiation is a suspicious response pattern, which leads to an increase in the expected costs. Therefore, when generating artificial responses to item batteries, falsifiers must take the outlined tradeoff into account which may result in distinct patterns over the field period for the proposed falsifier types.

As a robustness check and to illustrate the potential application of our approach in surveys lacking item batteries, we also use two further data quality measures: the share of rounded responses for numerical questions (Menold and Kemper 2014) and the share of extreme responses to Likert-scaled questions (Schäfer et al. 2005). Extreme responding has a looser relation to effort, and numeric questions are less frequent than item batteries in the questionnaire. Therefore, we will only briefly discuss their results and implications. Their measurement and the involved variables are described in the supplementary data S1 online.

We measure response differentiation for each interview by calculating the standard deviation of responses for several batteries of same-scaled items (following Kemper and Menold 2014). Although various approaches to measure response differentiation exist (Loosveldt and Beullens 2017; Kim et al. 2019), we use the standard deviation due to its simplicity and the possibility of



capturing differences on a continuous scale. A low standard deviation indicates little differentiation. As the questionnaire contains multiple appropriate item batteries, we obtain multiple standard deviations per interview. To combine the measures, we first standardize the standard deviation of every item battery across all interviews in the survey. The standardization prevents undesired effects caused by differences in scaling across item batteries. Next, the standardized standard deviations within every interview are averaged, with each standard deviation receiving a relative weight based on the number of items answered (without item nonresponse) in the respective battery and the total number of answered items across all batteries. This ensures that standard deviations calculated for longer item batteries receive a higher weight than shorter item batteries. The resulting formula is:

$$D_{ij} = \frac{\sum_{k=1}^{K} N_{ijk}\mathrm{SD}_{ijk}}{\sum_{k=1}^{K} N_{ijk}}, \qquad (1)$$

where N_{ijk} is the number of answered items for item battery k in interview j by interviewer i, SD_{ijk} is the respective z-standardized standard deviation, and the denominator is the total number of answered items across all batteries. For observations with average standard deviations for all item batteries, D_{ij} is close to zero. Observations with low standard deviations have values below zero, whereas observations with high standard deviations have positive D_{ij} values. Note that we cannot establish universal thresholds that denote whether D_{ij} is too low or high, as its values depend on the number of used item batteries and their content. For example, for independent standard normally distributed random variables, the standard deviation of their sum is the square root of the number of variables. Thus, determining outlier thresholds based on variance measures depends on the number of variables. In our application, this is further complicated by correlations between variables.

The IAB–BAMF–SOEP Survey of Refugees person questionnaire includes eight appropriate item batteries with at least five items and a minimum of five response options that are used in the analysis (see table S3 in the supplementary data online for the complete list of item batteries). Batteries with fewer items or response options are not considered here to allow for finer detection of differentiation tendencies. These item batteries come from the person questionnaire (TNS Infratest Sozialforschung 2016). Due to item nonresponse, none of the item batteries was answered by all respondents. As the standard deviations are standardized, we can include observations with missing standard deviations for some item batteries in the analysis. For five respondents, the standard deviation is missing for all item batteries. These observations were excluded from the analysis. The distribution of the resulting indicator is displayed in figure S3 in the supplementary data online.

For the ESS data, we use six item blocks (Loosveldt and Beullens 2017), which are listed in table S5 in the supplementary data online. Each item block contains at least five items and response options ranging from 0 to 10 or 1 to 5. The final measure of response differentiation is calculated in the same way as for the IAB–BAMF–SOEP Survey of Refugees. The distribution of the indicator is shown in figure S4 in the supplementary data online.

4. MODELING APPROACH

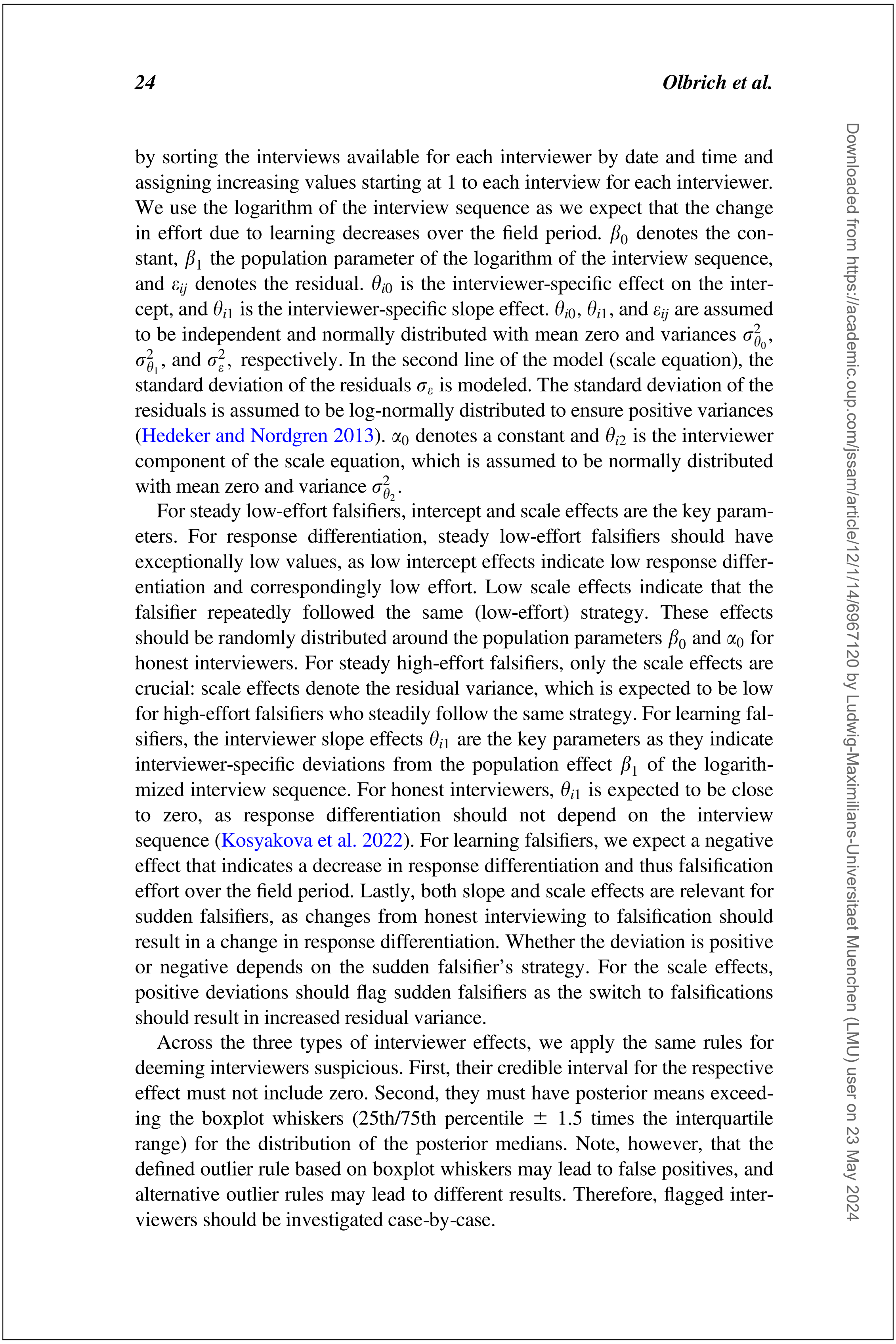
To test whether interviewers show suspicious behaviors over the field period, we employ multilevel modeling to disentangle interviewer from respondent effects and exploit the hierarchical data structure (respondents nested within interviewers) (Hox et al. 1991; Hox 1994). Such models have been applied to investigate interviewer effects on a variety of data quality measures (e.g., Pickery and Loosveldt 2004; Schnell and Kreuter 2005; Olson and Peytchev 2007; Kosyakova et al. 2015; Brunton-Smith et al. 2017; Loosveldt and Beullens 2017; Sharma and Elliott 2020; Sturgis et al. 2021). Among these studies, the effect of interview sequence (or within-survey experience) has also been considered (Olson and Peytchev 2007; Olson and Bilgen 2011; Kosyakova et al. 2015, 2022; Josten and Trappmann 2016; Loosveldt and Beullens 2017). However, these studies focused on overall interviewer effects and did not test whether suspicious individual interviewers can be detected. Moreover, only Brunton-Smith et al. (2017) and Sturgis et al. (2021) analyzed interviewer effects on residual variance. Pickery and Loosveldt (2004) and Sharma and Elliott (2020) are the closest to the present analysis as they use multilevel modeling to detect “exceptional” interviewers (i.e., interviewers with unusual response patterns).

We are interested in differences in the intercept, differences in the slope of the interview sequence, and differences in the residual variance across interviewers. While previous studies examined these differences in separate models, we fit a single model that contains interviewer effects on the intercept, slope, and scale. The base specification of this model is formalized below:

$$D_{ij} = \beta_0 + \theta_{i0} + (\beta_1 + \ \theta_{i1})\log\left(t_{ij}\right) + \varepsilon_{ij}, \qquad (2)$$

$$\log(\sigma_\varepsilon) = \ \alpha_0 + \ \theta_{i2}.$$

The dependent variable in the first line of the model (location equation) is response differentiation D_{ij}, which is calculated using equation (1) for each interviewer i and interview j. The interview sequence variable t_{ij} is generated



by sorting the interviews available for each interviewer by date and time and assigning increasing values starting at 1 to each interview for each interviewer. We use the logarithm of the interview sequence as we expect that the change in effort due to learning decreases over the field period. β_0 denotes the constant, β_1 the population parameter of the logarithm of the interview sequence, and ε_{ij} denotes the residual. θ_{i0} is the interviewer-specific effect on the intercept, and θ_{i1} is the interviewer-specific slope effect. θ_{i0}, θ_{i1}, and ε_{ij} are assumed to be independent and normally distributed with mean zero and variances $\sigma^2_{\theta_0}$, $\sigma^2_{\theta_1}$, and σ^2_{ε}, respectively. In the second line of the model (scale equation), the standard deviation of the residuals σ_ε is modeled. The standard deviation of the residuals is assumed to be log-normally distributed to ensure positive variances (Hedeker and Nordgren 2013). α_0 denotes a constant and θ_{i2} is the interviewer component of the scale equation, which is assumed to be normally distributed with mean zero and variance $\sigma^2_{\theta_2}$.

For steady low-effort falsifiers, intercept and scale effects are the key parameters. For response differentiation, steady low-effort falsifiers should have exceptionally low values, as low intercept effects indicate low response differentiation and correspondingly low effort. Low scale effects indicate that the falsifier repeatedly followed the same (low-effort) strategy. These effects should be randomly distributed around the population parameters β_0 and α_0 for honest interviewers. For steady high-effort falsifiers, only the scale effects are crucial: scale effects denote the residual variance, which is expected to be low for high-effort falsifiers who steadily follow the same strategy. For learning falsifiers, the interviewer slope effects θ_{i1} are the key parameters as they indicate interviewer-specific deviations from the population effect β_1 of the logarithmized interview sequence. For honest interviewers, θ_{i1} is expected to be close to zero, as response differentiation should not depend on the interview sequence (Kosyakova et al. 2022). For learning falsifiers, we expect a negative effect that indicates a decrease in response differentiation and thus falsification effort over the field period. Lastly, both slope and scale effects are relevant for sudden falsifiers, as changes from honest interviewing to falsification should result in a change in response differentiation. Whether the deviation is positive or negative depends on the sudden falsifier's strategy. For the scale effects, positive deviations should flag sudden falsifiers as the switch to falsifications should result in increased residual variance.

Across the three types of interviewer effects, we apply the same rules for deeming interviewers suspicious. First, their credible interval for the respective effect must not include zero. Second, they must have posterior means exceeding the boxplot whiskers (25th/75th percentile ± 1.5 times the interquartile range) for the distribution of the posterior medians. Note, however, that the defined outlier rule based on boxplot whiskers may lead to false positives, and alternative outlier rules may lead to different results. Therefore, flagged interviewers should be investigated case-by-case.

We note that interviewers employed in the IAB–BAMF–SOEP Survey of Refugees and the ESS were not randomly assigned to households across the respective countries but were assigned to regional clusters: primary sampling units. Therefore, the results observed for interviewers may be driven by regional clustering effects (Schnell and Kreuter 2005). To disentangle interviewer and regional cluster effects, we would require sufficient interpenetration, that is, interviewers must work in multiple clusters, and multiple interviewers must work in a given cluster, which is not prevalent in the data used here. For the IAB–BAMF–SOEP Survey of Refugees, however, regional clusters are expected to have only small effects, as the target population are recently arrived refugees subject to state-based residential allocation policies following a political quota (BAMF 2019; Kosyakova et al. 2019). Lastly, controlling for small-scale regional effects could prevent the detection of falsifiers operating or cooperating in the same region (Yamamoto and Lennon 2018; Bergmann et al. 2019).

Nonetheless, we test the robustness of the results by including control variables for respondent and area characteristics in the model for the IAB–BAMF–SOEP Survey of Refugees and ESS data. They include respondents' age, gender, education, living arrangement (only for the IAB–BAMF–SOEP Survey of Refugees), an urban-rural binary variable, as well as federal state/region fixed effects (see supplementary data S4 online). Note that the included variables are more likely to explain overall effects on the dependent variable than extreme slopes or scales observed for single interviewers.

The model is fitted using Markov chain Monte Carlo (MCMC) methods. In particular, we use the No-U-Turn Sampler (Hoffman and Gelman 2014), a version of the Hamiltonian Monte Carlo algorithm implemented in Stan (Carpenter et al. 2017) and accessed via the brms interface (Bürkner 2017, 2018) in R (R Core Team 2020). The model is fitted using eight chains of 8,000 iterations, each with a burn-in period of 3,000 iterations. We specify flat priors for the population-level coefficients and default half student-t priors with three degrees of freedom for the standard deviations of the interviewer effects. We assessed whether the priors for the interviewer effects affect the results by trying different priors such as half Cauchy and inverse Gamma distributions, but the results did not change. Model convergence was evaluated by the $\widehat{R}$ statistic with a critical value of 1.01 for each model parameter (Gelman et al. 2013, p. 285) and by ensuring that there were no divergent transitions (Betancourt 2017). Estimates and credible intervals shown in the results section are based on posterior distributions for the model parameters obtained from the MCMC draws.

5. RESULTS

For each dataset, we first estimate the base specifications with no control variables and then conduct a robustness check with control variables. Note that we are not interested in explaining differences in levels, slopes, or residual variances but in detecting suspicious interviewers.

5.1 Analysis of the IAB–BAMF–SOEP Survey of Refugees

The interviewer effects for the IAB–BAMF–SOEP Survey of Refugees are displayed in figure 1. Figure 1a shows the effects on the intercept, figure 1b shows the effects on the slope, and figure 1c shows the effects on the scale. In each panel, each point corresponds to a single interviewer. The interviewer effects are sorted by size, and 95 percent credible intervals are provided. The dashed horizontal lines depict the boxplot whiskers. The estimation results are reported in column 1 in table S9 in the supplementary data online. The estimated coefficient of the logarithmized interview sequence is positive but negligible in size. From the first to the 10th interview, response differentiation increases by 0.058, which equals roughly 12 percent of one standard deviation. Hence, overall the response differentiation changes only slightly over the field period.

Figure 1a shows that the verified falsifier's intercept effect is not suspicious (ranked 68th). The first-ranked interviewer (interviewer D) is suspicious but conducted only eight interviews. The last-ranked interviewer (interviewer E) has rather high differentiation values and conducted 27 interviews. None of these interviewers was flagged by further statistical identification methods (see Kosyakova et al. 2019) or further checks (such as recontacts) by the survey organization. Thus, relying on the intercept effects alone is insufficient for identifying the verified case. Remember that the intercept effects denote differences at the first interview as interview sequence effects are included in the model.

As displayed in figure 1b, most of the slope effects for the interview sequence are close to zero, or the credible intervals include zero. As with the intercept effects, interviewer A does not deviate from the other interviewers and is ranked 28th. For the first-ranked interviewer (henceforth called interviewer B), however, the slope value deviates substantially from the others, implying a decrease in response differentiation over the field period. Similarly, the second-ranked interviewer (henceforth called interviewer C) is suspicious with a negative slope effect. Interviewers B and C conducted 46 and 16 interviews, respectively. Accordingly, interviewers B and C reveal a suspicious slope effect consistent with learning behavior or switching from honest interviewing to falsification. These results and conclusions drawn from further statistical checks were reported to the survey organization, who verified that interviewers B and C were indeed deviant, although the survey organization could not exactly tell which interviews were falsified (Kosyakova et al. 2019). The published data were immediately revised after the detection (IAB et al. 2019).

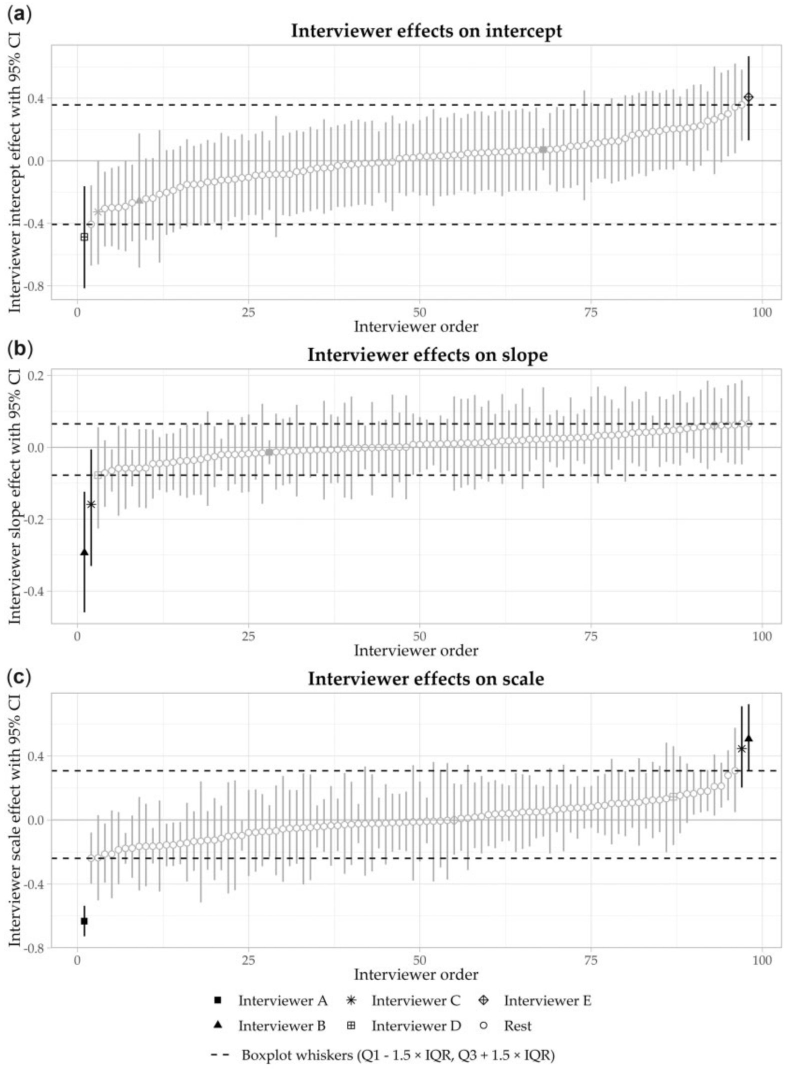


Figure 1. Interviewer Effects on Intercept, Slope, and Scale. Bolded interviewers have median posteriors exceeding the boxplot whiskers and credible intervals not including zero. Predictions are based on model 1 in table S9 in the supplementary data online. IAB-BAMF-SOEP Survey of Refugees 2016.

Lastly, figure 1c shows that the scale effects are distributed homogeneously, except for the first-ranked interviewer and the last-ranked pair of interviewers. The first-ranked interviewer is interviewer A who has a suspiciously low scale effect. This indicates limited variation in response differentiation,

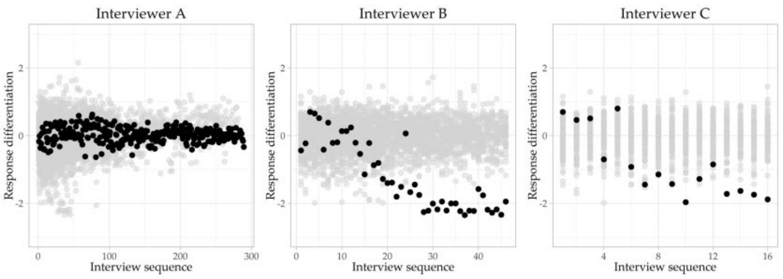


Figure 2. Development of Response Differentiation for Verified Falsifiers. Black dots correspond to the respective falsifier, and gray dots correspond to the rest of the sample. IAB-BAMF-SOEP Survey of Refugees 2016.

which—combined with the inconspicuous intercept effect—is expected for steady high-effort falsifiers. The two last-ranked interviewers are interviewers B and C. The exceptional slope effects observed for these interviewers also lead to a suspicious scale effect.

To investigate the behaviors of interviewer A and the additionally identified interviewers B and C in greater detail, figure 2 shows the development of response differentiation over the field period for each of them. For reference, the differentiation values for the rest of the sample are also shown. Interviewer A has a relatively low variation in response differentiation with values around zero, close to the overall average in the sample. Such a pattern is in line with the behavior expected of a steady high-effort falsifier. In contrast, interviewer B has relatively high response differentiation values at the beginning of the field period that steadily decrease from the 10th interview onward. Toward the end of the field period, response differentiation is clearly below the "normal" values observed for the rest of the sample. This pattern is in line with a learning falsifier, although it is also possible that some of the first interviews were real, and the interviewer switched to falsification. For interviewer C, the pattern is less clear due to the limited number of available observations. The response differentiation values are at the upper end of the distribution in the beginning of the field period, but this strictly changes after the 5th interview. This break may either illustrate learning behavior or a change from conducting real interviews to falsifying interviews. As mentioned above, detailed information on whether every interview of interviewers B and C was falsified is not available.

To test the robustness of the results, we replicate the benchmark models by adding multiple control variables (gender, age, education, accommodation, region, rural/urban) to the location equation. The estimation results of these models are reported in table S10 in the supplementary data online. As illustrated in figure S5 in the supplementary data online, the deviations of interviewers B and C for the slope effects cannot be explained by the included explanatory variables, although the effect for interviewer C is now closer to zero. Similarly, the explanatory variables cannot explain the deviation of interviewer A for the scale effects.

Figures S9 and S10 in the supplementary data online show the results for the model without controls for two further indicators, extreme responding and rounding. For extreme responding, two interviewers are flagged for the intercept effects, but their values are close to the rest of the sample. Interviewers A, B, and C have inconspicuous values. Interviewers B and C have exceptionally negative slope effects, although interviewer C is barely below the boxplot rule. For the scale effects, one interviewer is flagged, and interviewer A is ranked second, but their values seem in line with the distribution for the remaining sample. For rounding, interviewer A has a suspiciously low intercept effect, and interviewer B has a suspiciously large intercept effect. The slope effects depict that several interviewers have negative effects, but these values are not particularly exceptional. Interviewer B is the only interviewer with a relatively large slope effect, denoting that the share of rounded responses increased over the field period. A closer inspection revealed that this interviewer heavily reduced the number of valid responses to open-ended numeric questions over the field period, which led to frequent high shares of rounded answers as, for example, one numeric item in the interview had a valid response, and this response was a rounded number, which results in a share of 100 percent. For the scales, interviewer A has an exceptionally negative value, indicating reduced residual variance. Interestingly, interviewer C is not flagged by any of the estimated interviewer effects. In summary, for neither of the indicators are all three interviewers A, B, and C flagged.

5.2 Analysis of the European Social Survey

For the ESS data, we only discuss the results for Ireland in detail while touching on the results for Denmark and Hungary only briefly for brevity. Figure 3 displays the interviewer effects for Ireland. We also fit the multilevel model for the three countries with covariates, and the results remain robust to these extensions (see supplementary data S5 and S6 online). Figure 3a shows that multiple interviewers have suspiciously low or high values of response differentiation, although most of them do not significantly differ from the unflagged interviewers. As displayed in figure 3b, there is an interviewer who has a suspicious negative slope effect indicating potential learning behavior. One further interviewer has a suspicious positive slope effect. Finally, figure 3c shows two interviewers with suspicious negative scale effects expected of steady high- or low-effort falsifiers. The first-ranked interviewer is the interviewer who is ranked first in figure 3a, which is expected for a steady low-effort falsifier. None of the other interviewers has multiple suspicious effects.

Next, we take a closer look at the development of response differentiation over the field period for the flagged interviewers. Figure 4 shows the

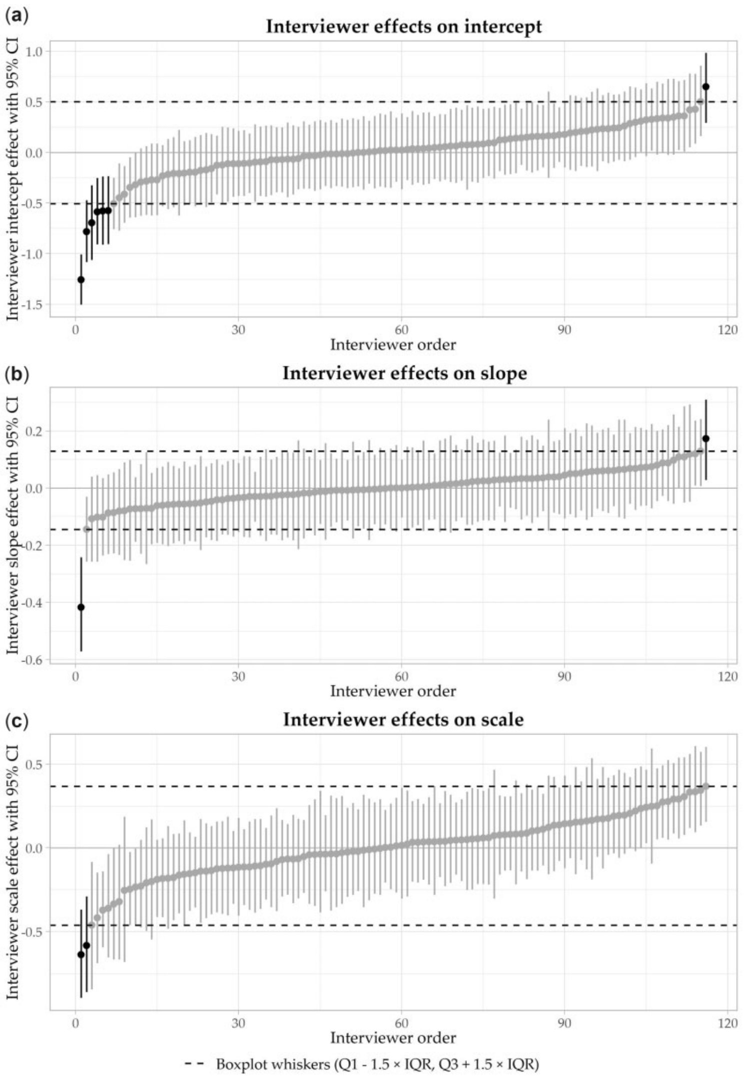


Figure 3. Interviewer Effects on Intercept, Slope, and Scale. Bolded interviewers have median posteriors exceeding the boxplot whiskers and credible intervals not including zero. Predictions are based on model 3 in table S9 in the supplementary data online. ESS Round 6, Ireland.

differentiation results for the interviewer ranked first for the slope effects (interviewer ESS-A), and the interviewers ranked first and second for the scale effects (interviewers ESS-B and ESS-C, respectively). For interviewer ESS-A, response differentiation decreases over the field period and thus follows the

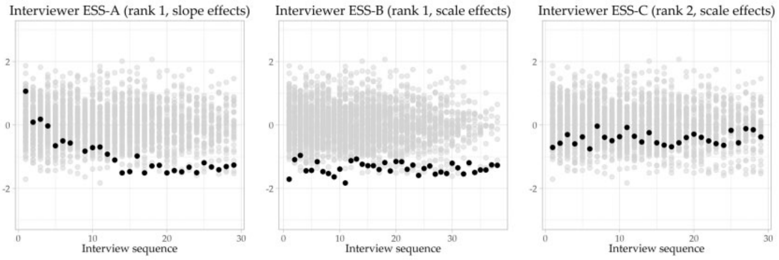


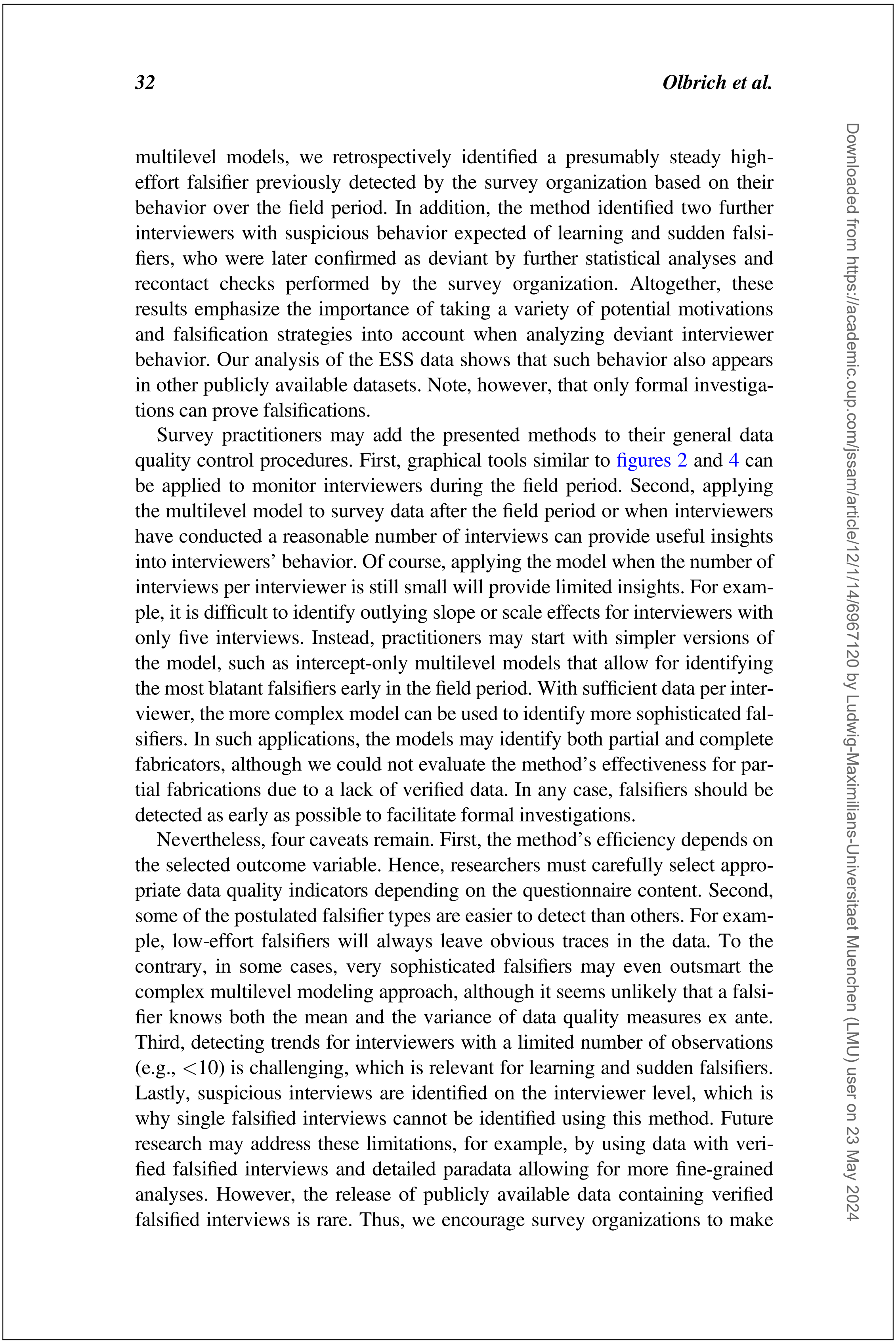
Figure 4. Development of Response Differentiation. Black dots correspond to the respective suspicious interviewer, and gray dots correspond to the rest of the sample. ESS Round 6, Ireland.

behavior expected of a learning falsifier. Interviewer ESS-B—ranked first for the intercept and scale effects—shows limited variation around reduced response differentiation, which is characteristic of a steady low-effort falsifier. Lastly, interviewer ESS-C also has limited variation around close to average response differentiation, suggesting deviant behavior consistent with a more sophisticated falsifier.

The results for Denmark and Hungary are provided in supplementary data S8 online. For Denmark, we find only minor interviewer effects on the intercept, slope, and scale and no suspicious interviewer, which is in line with previous research on interviewer effects in Scandinavian countries (Loosveldt and Beullens 2017). We find more evidence of interviewer effects in the Hungary sample. Multiple interviewers have negative effects on the intercept, although none of these effects is suspicious. There are several interviewers with slope effects below the boxplot whisker line, but only for one interviewer who does not significantly differ from unsuspicious interviewers does the credible interval not include zero. With regard to the scale effects, several interviewers have rather low values indicative of behavior expected of steady high- or low-effort falsifiers, but none of the effects exceeds the boxplot whisker rule. Although the interviewer effects are not as suspicious as for the IAB–BAMF–SOEP Survey of Refugees or Ireland, interviewers with particularly low intercept, slope, or scale effects may have required closer inspection.

6. DISCUSSION

Falsified interviews can substantially bias survey results (e.g., Schräpler and Wagner 2005). To prevent and detect falsifications, survey methodologists should not only use empirical detection methods, but also comprehend falsifiers' motivations and behaviors. In this study, we posited four distinct falsifier types: steady low-effort falsifiers, steady high-effort falsifiers, learning falsifiers, and sudden falsifiers. Using data containing verified falsifications and



multilevel models, we retrospectively identified a presumably steady high-effort falsifier previously detected by the survey organization based on their behavior over the field period. In addition, the method identified two further interviewers with suspicious behavior expected of learning and sudden falsifiers, who were later confirmed as deviant by further statistical analyses and recontact checks performed by the survey organization. Altogether, these results emphasize the importance of taking a variety of potential motivations and falsification strategies into account when analyzing deviant interviewer behavior. Our analysis of the ESS data shows that such behavior also appears in other publicly available datasets. Note, however, that only formal investigations can prove falsifications.

Survey practitioners may add the presented methods to their general data quality control procedures. First, graphical tools similar to figures 2 and 4 can be applied to monitor interviewers during the field period. Second, applying the multilevel model to survey data after the field period or when interviewers have conducted a reasonable number of interviews can provide useful insights into interviewers' behavior. Of course, applying the model when the number of interviews per interviewer is still small will provide limited insights. For example, it is difficult to identify outlying slope or scale effects for interviewers with only five interviews. Instead, practitioners may start with simpler versions of the model, such as intercept-only multilevel models that allow for identifying the most blatant falsifiers early in the field period. With sufficient data per interviewer, the more complex model can be used to identify more sophisticated falsifiers. In such applications, the models may identify both partial and complete fabricators, although we could not evaluate the method's effectiveness for partial fabrications due to a lack of verified data. In any case, falsifiers should be detected as early as possible to facilitate formal investigations.

Nevertheless, four caveats remain. First, the method's efficiency depends on the selected outcome variable. Hence, researchers must carefully select appropriate data quality indicators depending on the questionnaire content. Second, some of the postulated falsifier types are easier to detect than others. For example, low-effort falsifiers will always leave obvious traces in the data. To the contrary, in some cases, very sophisticated falsifiers may even outsmart the complex multilevel modeling approach, although it seems unlikely that a falsifier knows both the mean and the variance of data quality measures ex ante. Third, detecting trends for interviewers with a limited number of observations (e.g., <10) is challenging, which is relevant for learning and sudden falsifiers. Lastly, suspicious interviews are identified on the interviewer level, which is why single falsified interviews cannot be identified using this method. Future research may address these limitations, for example, by using data with verified falsified interviews and detailed paradata allowing for more fine-grained analyses. However, the release of publicly available data containing verified falsified interviews is rare. Thus, we encourage survey organizations to make such data available to researchers to help advance our understanding of interviewer falsification.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

AAPOR (2003), "Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection, and Repair of Its Effects." https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/falsification.pdf.

AAPOR (2016), "Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys," 9th edition, AAPOR.

BAMF (2019), "Initial Distribution of Asylum-Seekers." http://www.bamf.de/EN/Fluechtlingsschutz/AblaufAsylv/Erstverteilung/erstverteilung-node.html.

Bergmann, M., Schuller, K., and Malter, F. (2019), "Preventing Interview Falsifications during Fieldwork in the Survey of Health, Ageing and Retirement in Europe (SHARE)," *Longitudinal and Life Course Studies*, 10(4), 513–530.

Betancourt, M. (2017), "A Conceptual Introduction to Hamiltonian Monte Carlo." http://arxiv.org/abs/1701.02434

Beullens, K., Matsuo, H., Loosveldt, G., and Vandenplas, C. (2014), *Quality Report for the European Social Survey, Round 6*. London: European Social Survey ERIC.

Blasius, J., and Thiessen, V. (2013), "Detecting Poorly Conducted Interviews," in *Interviewers' Deviations in Surveys—Impact, Reasons, Detection and Prevention*, eds. P. Winker, N. Menold, and R. Porst, Frankfurt am Main: Peter Lang, Academic Research, pp. 67–88.

———. (2015), "Should We Trust Survey Data? Assessing Response Simplification and Data Fabrication," *Social Science Research*, 52, 479–493.

———. (2021), "Perceived Corruption, Trust, and Interviewer Behavior in 26 European Countries," *Sociological Methods and Research*, 50(2), 740–777.

Bredl, S., Storfinger, N., and Menold, N. (2013), "A Literature Review of Methods to Detect Fabricated Survey Data," in *Interviewers' Deviations in Surveys—Impact, Reasons, Detection and Prevention*, eds. P. Winker, N. Menold, and R. Porst, Frankfurt am Main: Peter Lang, Academic Research, pp. 3–24.

Bredl, S., Winker, P., and Kötschau, K. (2012), "A Statistical Approach to Detect Interviewer Falsification of Survey Data," *Survey Methodology*, 38(1), 1–10.

Brücker, H., Rother, N., and Schupp, J. (2017), "IAB-BAMF-SOEP Befragung von Geflüchteten 2016: Studiendesign, Feldergebnisse Sowie Analysen zu Schulischer wie Beruflicher Qualifikation, Sprachkenntnissen Sowie Kognitiven Potenzialen," *IAB-Forschungsbericht*, 13, 1–76.

Brunton-Smith, I., Sturgis, P., and Leckie, G. (2017), "Detecting and Understanding Interviewer Effects on Survey Data by Using a Cross-Classified Mixed Effects Location–Scale Model," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(2), 551–568.

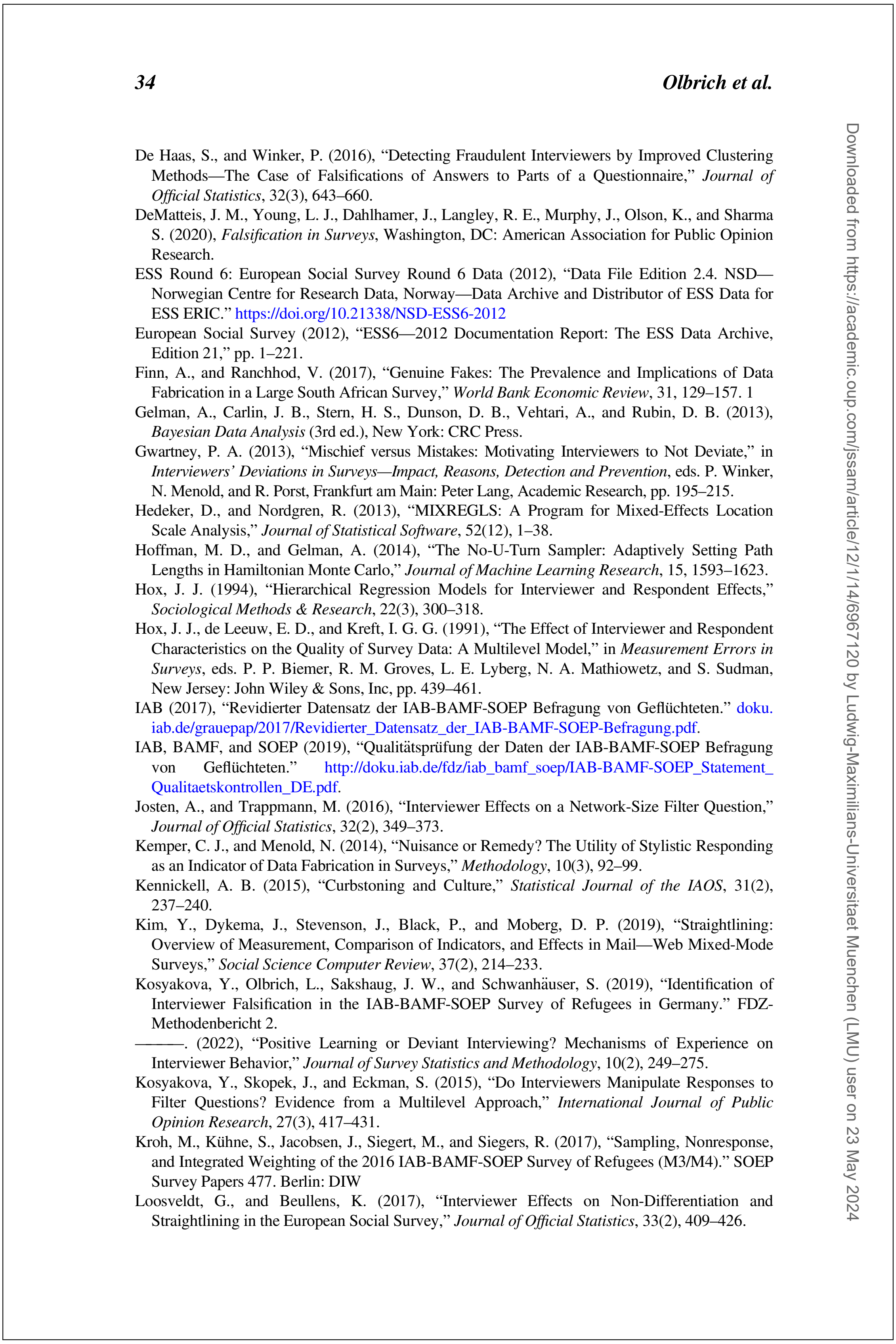
Bürkner, P. C. (2017), "brms: An R Package for Bayesian Multilevel Models Using Stan," *Journal of Statistical Software*, 80(1), 1–28.

———. (2018), "Advanced Bayesian Multilevel Modeling with the R Package brms," *R Journal*, 10(1), 395–411.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017), "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, 76(1), 1–32.

Crespi, L. P. (1945), "The Cheater Problem in Polling," *Public Opinion Quarterly*, 9(4), 431–445.

Cressey, D. R. (1953), *Other People's Money*, Montclair, NJ: Patterson Smith.



De Haas, S., and Winker, P. (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), 643–660.

DeMatteis, J. M., Young, L. J., Dahlhamer, J., Langley, R. E., Murphy, J., Olson, K., and Sharma S. (2020), *Falsification in Surveys*, Washington, DC: American Association for Public Opinion Research.

ESS Round 6: European Social Survey Round 6 Data (2012), "Data File Edition 2.4. NSD—Norwegian Centre for Research Data, Norway—Data Archive and Distributor of ESS Data for ESS ERIC." https://doi.org/10.21338/NSD-ESS6-2012

European Social Survey (2012), "ESS6—2012 Documentation Report: The ESS Data Archive, Edition 21," pp. 1–221.

Finn, A., and Ranchhod, V. (2017), "Genuine Fakes: The Prevalence and Implications of Data Fabrication in a Large South African Survey," *World Bank Economic Review*, 31, 129–157. 1

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis* (3rd ed.), New York: CRC Press.

Gwartney, P. A. (2013), "Mischief versus Mistakes: Motivating Interviewers to Not Deviate," in *Interviewers' Deviations in Surveys—Impact, Reasons, Detection and Prevention*, eds. P. Winker, N. Menold, and R. Porst, Frankfurt am Main: Peter Lang, Academic Research, pp. 195–215.

Hedeker, D., and Nordgren, R. (2013), "MIXREGLS: A Program for Mixed-Effects Location Scale Analysis," *Journal of Statistical Software*, 52(12), 1–38.

Hoffman, M. D., and Gelman, A. (2014), "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, 15, 1593–1623.

Hox, J. J. (1994), "Hierarchical Regression Models for Interviewer and Respondent Effects," *Sociological Methods & Research*, 22(3), 300–318.

Hox, J. J., de Leeuw, E. D., and Kreft, I. G. G. (1991), "The Effect of Interviewer and Respondent Characteristics on the Quality of Survey Data: A Multilevel Model," in *Measurement Errors in Surveys*, eds. P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, New Jersey: John Wiley & Sons, Inc, pp. 439–461.

IAB (2017), "Revidierter Datensatz der IAB-BAMF-SOEP Befragung von Geflüchteten." doku.iab.de/grauepap/2017/Revidierter_Datensatz_der_IAB-BAMF-SOEP-Befragung.pdf.

IAB, BAMF, and SOEP (2019), "Qualitätsprüfung der Daten der IAB-BAMF-SOEP Befragung von Geflüchteten." http://doku.iab.de/fdz/iab_bamf_soep/IAB-BAMF-SOEP_Statement_Qualitaetskontrollen_DE.pdf.

Josten, A., and Trappmann, M. (2016), "Interviewer Effects on a Network-Size Filter Question," *Journal of Official Statistics*, 32(2), 349–373.

Kemper, C. J., and Menold, N. (2014), "Nuisance or Remedy? The Utility of Stylistic Responding as an Indicator of Data Fabrication in Surveys," *Methodology*, 10(3), 92–99.

Kennickell, A. B. (2015), "Curbstoning and Culture," *Statistical Journal of the IAOS*, 31(2), 237–240.

Kim, Y., Dykema, J., Stevenson, J., Black, P., and Moberg, D. P. (2019), "Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail—Web Mixed-Mode Surveys," *Social Science Computer Review*, 37(2), 214–233.

Kosyakova, Y., Olbrich, L., Sakshaug, J. W., and Schwanhäuser, S. (2019), "Identification of Interviewer Falsification in the IAB-BAMF-SOEP Survey of Refugees in Germany." FDZ-Methodenbericht 2.

———. (2022), "Positive Learning or Deviant Interviewing? Mechanisms of Experience on Interviewer Behavior," *Journal of Survey Statistics and Methodology*, 10(2), 249–275.

Kosyakova, Y., Skopek, J., and Eckman, S. (2015), "Do Interviewers Manipulate Responses to Filter Questions? Evidence from a Multilevel Approach," *International Journal of Public Opinion Research*, 27(3), 417–431.

Kroh, M., Kühne, S., Jacobsen, J., Siegert, M., and Siegers, R. (2017), "Sampling, Nonresponse, and Integrated Weighting of the 2016 IAB-BAMF-SOEP Survey of Refugees (M3/M4)." SOEP Survey Papers 477. Berlin: DIW

Loosveldt, G., and Beullens, K. (2017), "Interviewer Effects on Non-Differentiation and Straightlining in the European Social Survey," *Journal of Official Statistics*, 33(2), 409–426.

Menold, N., and Kemper, C. J. (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, N., Winker, P., Storfinger, N., and Kemper, C. J. (2013), “A Method for Ex-Post Identification of Falsifications in Survey Data,” in *Interviewers’ Deviations in Surveys—Impact, Reasons, Detection and Prevention*, eds. P. Winker, N. Menold, and R. Porst, Frankfurt am Main: Peter Lang, Academic Research, pp. 25–47.

Murphy, J., Biemer, P., Stringer, C., Thissen, R., Day, O., and Hsieh, Y. P. (2016), “Interviewer Falsification: Current and Best Practices for Prevention, Detection, and Mitigation,” *Statistical Journal of the IAOS*, 32(3), 313–326.

Olson, K., and Bilgen, I. (2011), “The Role of Interviewer Experience on Acquiescence,” *Public Opinion Quarterly*, 75(1), 99–114.

Olson, K., and Peytchev, A. (2007), “Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes,” *Public Opinion Quarterly*, 71(2), 273–286.

Pickery, J., and Loosveldt, G. (2004), “A Simultaneous Analysis of Interviewer Effects on Various Data Quality Indicators with Identification of Exceptional Interviewers,” *Journal of Official Statistics*, 20(1), 77–89.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at https://www.r-project.org/.

Robbins, M. (2019), “New Frontiers in Detecting Data Fabrication,” in *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, eds. T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, and B. Dorer, John Wiley & Sons, pp. 771–805.

Schäfer, C., Schräpler, J.-P., Müller, K.-R., and Wagner, G. G. (2005), “Automatic Identification of Faked and Fraudulent Interviews in the German SOEP,” *Schmollers Jahrbuch*, 125(1), 183–193.

Schnell, R., and Kreuter, F. (2005), “Separating Interviewer and Sampling-Point Effects,” *Journal of Official Statistics*, 21(3), 389–410.

Schräpler, J.-P., and Wagner, G. G. (2005), “Characteristics and Impact of Faked Interviews in Surveys—An Analysis of Genuine Fakes in the Raw Data of SOEP,” *Allgemeines Statistisches Archiv*, 89(1), 7–20.

Schwanhäuser, S., Sakshaug, J. W., and Kosyakova, Y. (2022), “How to Catch a Falsifier: Comparison of Statistical Detection Methods for Interviewer Falsification,” *Public Opinion Quarterly*, 81(1), 1–31.

Sharma, S., and Elliott, M. R. (2020), “Detecting Falsification in a Television Audience Measurement Panel Survey,” *International Journal of Market Research*, 62(4), 432–448.

Sturgis, P., Maslovskaya, O., Durrant, G., and Brunton-Smith, I. (2021), “The Interviewer Contribution to Variability in Response Times in Face-to-Face Interview Surveys,” *Journal of Survey Statistics and Methodology*, 9(4), 701–721.

Thissen, M. R., and Myers, S. K. (2016), “Systems and Processes for Detecting Interviewer Falsification and Assuring Data Collection Quality,” *Statistical Journal of the IAOS*, 32(3), 339–347.

TNS Infratest Sozialforschung (2016), “Erhebungsinstrumente der IAB-BAMF-SOEP-Befragung von Geflüchteten 2016: Integrierter Personen- und Biografiefragebogen, Stichproben M3-M4.” SOEP Survey Papers (Series A): 362, Berlin: DIW/SOEP.

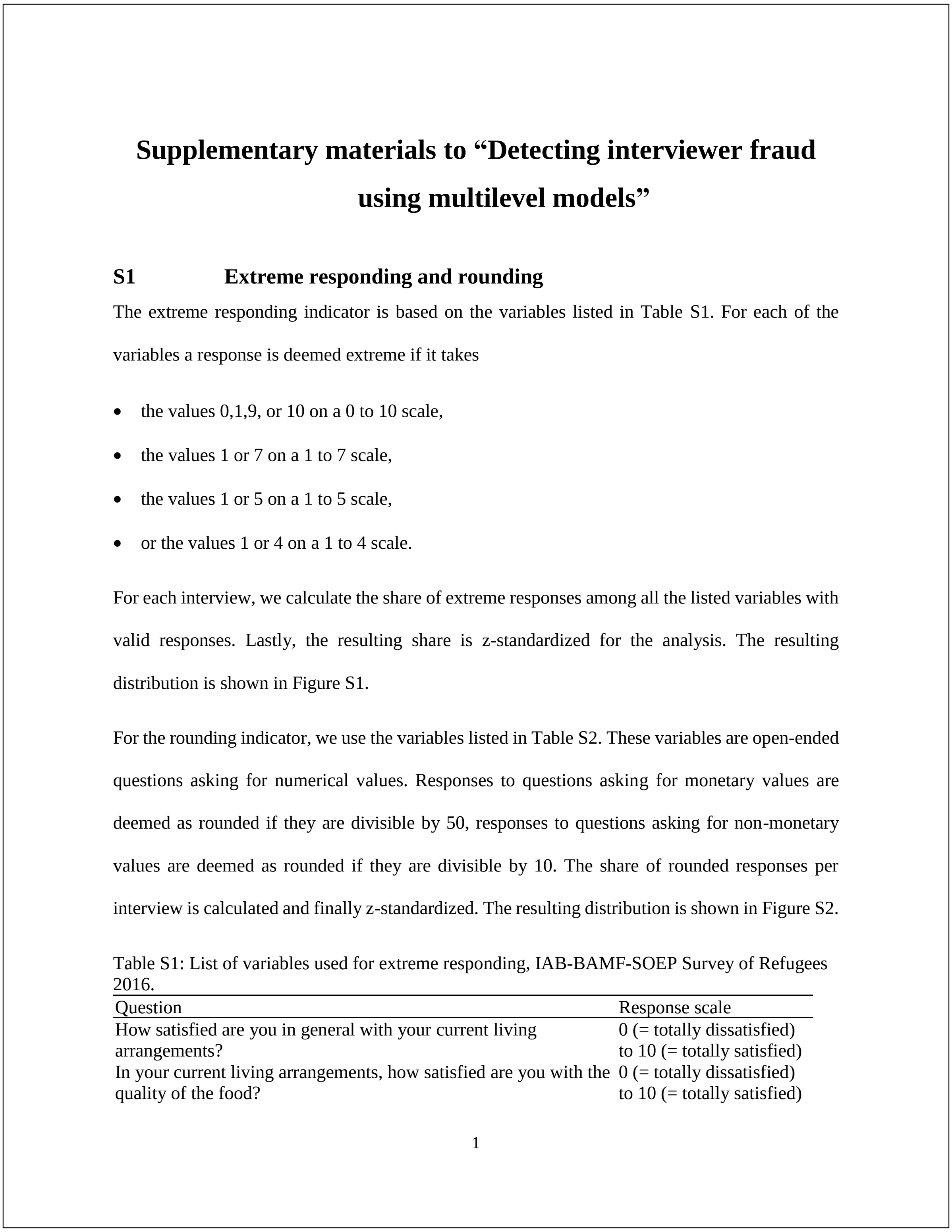
Weisberg, H. F. (2005), *The Total Survey Error Approach: A Guide to the New Science of Survey Research*, Chicago: The University of Chicago Press.

West, B. T., and Blom, A. G. (2017), “Explaining Interviewer Effects: A Research Synthesis,” *Journal of Survey Statistics and Methodology*, 5(2), 175–211.

Winker, P. (2016), “Assuring the Quality of Survey Data: Incentives, Detection and Documentation of Deviant Behavior,” *Statistical Journal of the IAOS*, 32(3), 295–303.

Yamamoto, K., and Lennon, M. L. (2018), “Understanding and Detecting Data Fabrication in Large-Scale Assessments,” *Quality Assurance in Education*, 26(2), 196–212.

Yan, T. (2008), “Nondifferentiation,” in *Encyclopedia of Survey Research Methods*, ed. P. J. Lavrakas, Thousand Oaks, CA: SAGE Publications, Inc, pp. 520–521.



Supplementary materials to “Detecting interviewer fraud using multilevel models”

S1 Extreme responding and rounding

The extreme responding indicator is based on the variables listed in Table S1. For each of the variables a response is deemed extreme if it takes

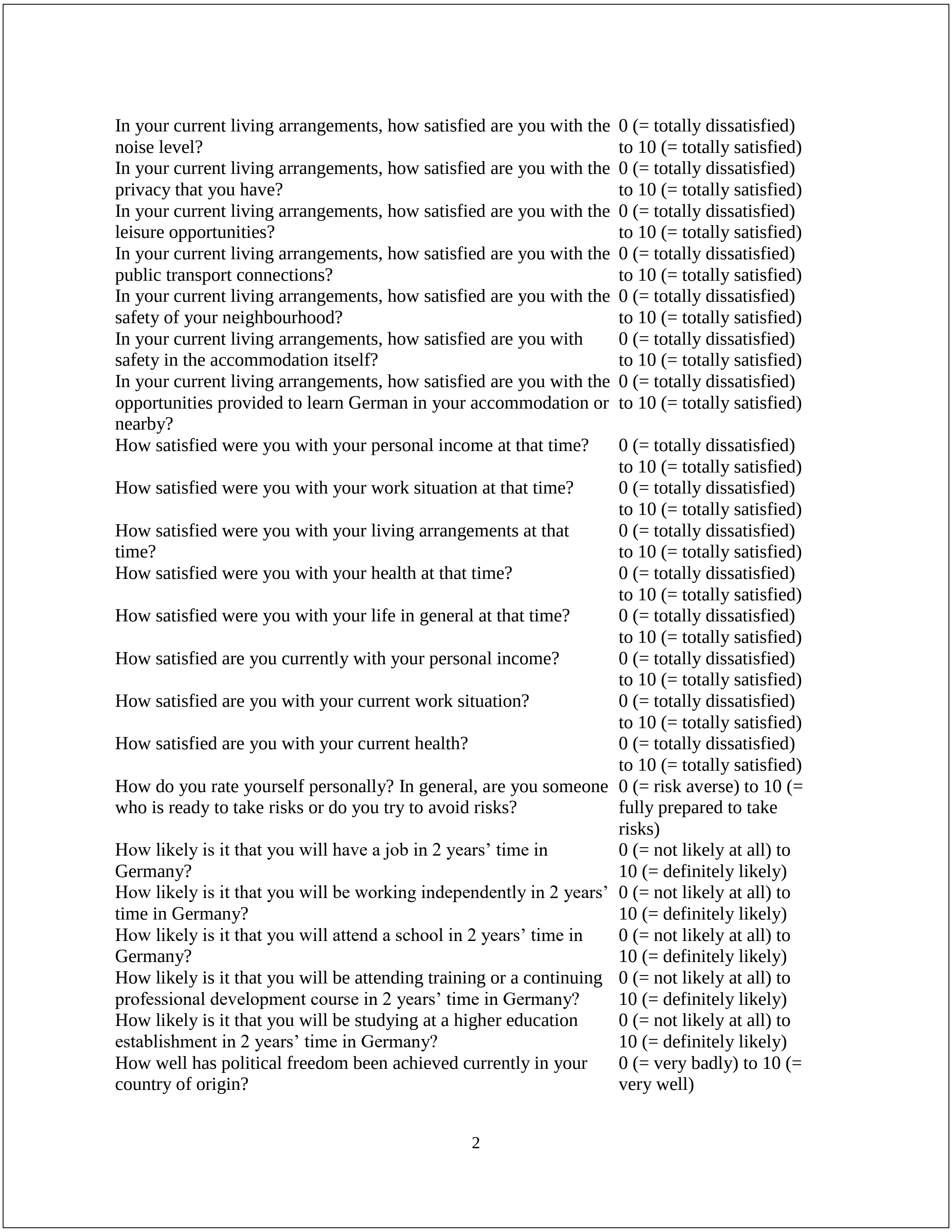
- the values 0,1,9, or 10 on a 0 to 10 scale,
- the values 1 or 7 on a 1 to 7 scale,
- the values 1 or 5 on a 1 to 5 scale,
- or the values 1 or 4 on a 1 to 4 scale.

For each interview, we calculate the share of extreme responses among all the listed variables with valid responses. Lastly, the resulting share is z-standardized for the analysis. The resulting distribution is shown in Figure S1.

For the rounding indicator, we use the variables listed in Table S2. These variables are open-ended questions asking for numerical values. Responses to questions asking for monetary values are deemed as rounded if they are divisible by 50, responses to questions asking for non-monetary values are deemed as rounded if they are divisible by 10. The share of rounded responses per interview is calculated and finally z-standardized. The resulting distribution is shown in Figure S2.

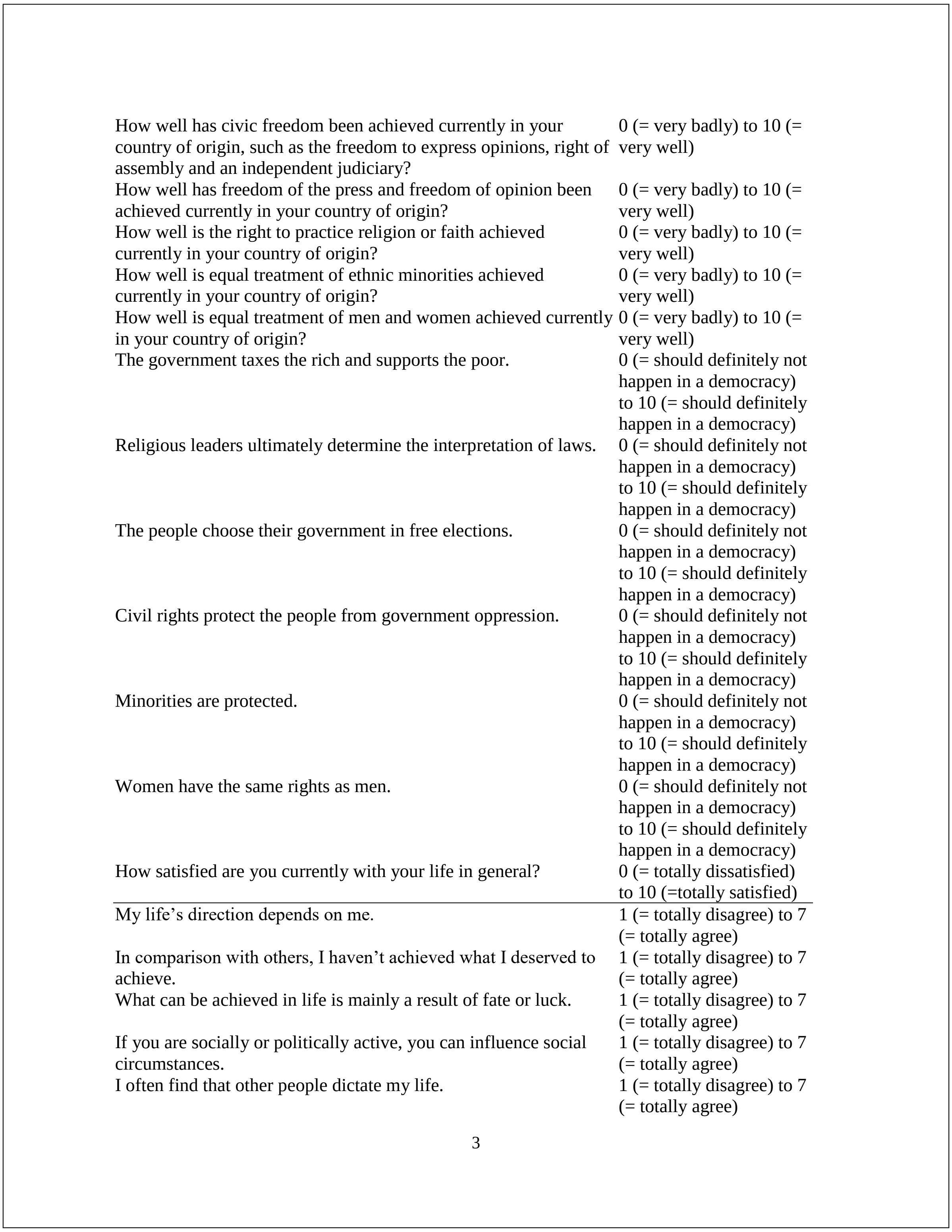
Table S1: List of variables used for extreme responding, IAB-BAMF-SOEP Survey of Refugees 2016.

Question	Response scale
How satisfied are you in general with your current living arrangements?	0 (= totally dissatisfied) to 10 (= totally satisfied)
In your current living arrangements, how satisfied are you with the quality of the food?	0 (= totally dissatisfied) to 10 (= totally satisfied)



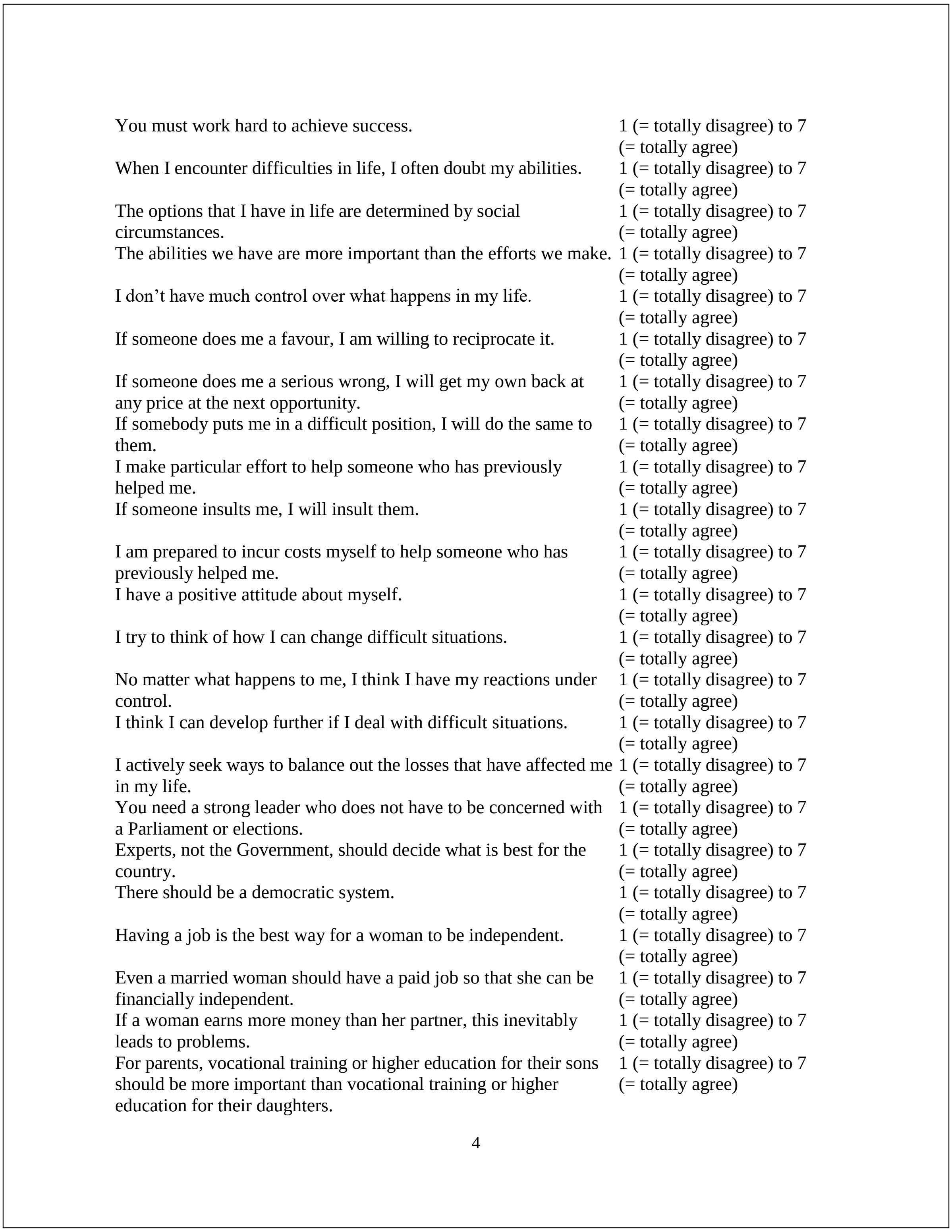
In your current living arrangements, how satisfied are you with the noise level?	0 (= totally dissatisfied) to 10 (= totally satisfied)
In your current living arrangements, how satisfied are you with the privacy that you have?	0 (= totally dissatisfied) to 10 (= totally satisfied)
In your current living arrangements, how satisfied are you with the leisure opportunities?	0 (= totally dissatisfied) to 10 (= totally satisfied)
In your current living arrangements, how satisfied are you with the public transport connections?	0 (= totally dissatisfied) to 10 (= totally satisfied)
In your current living arrangements, how satisfied are you with the safety of your neighbourhood?	0 (= totally dissatisfied) to 10 (= totally satisfied)
In your current living arrangements, how satisfied are you with safety in the accommodation itself?	0 (= totally dissatisfied) to 10 (= totally satisfied)
In your current living arrangements, how satisfied are you with the opportunities provided to learn German in your accommodation or nearby?	0 (= totally dissatisfied) to 10 (= totally satisfied)
How satisfied were you with your personal income at that time?	0 (= totally dissatisfied) to 10 (= totally satisfied)
How satisfied were you with your work situation at that time?	0 (= totally dissatisfied) to 10 (= totally satisfied)
How satisfied were you with your living arrangements at that time?	0 (= totally dissatisfied) to 10 (= totally satisfied)
How satisfied were you with your health at that time?	0 (= totally dissatisfied) to 10 (= totally satisfied)
How satisfied were you with your life in general at that time?	0 (= totally dissatisfied) to 10 (= totally satisfied)
How satisfied are you currently with your personal income?	0 (= totally dissatisfied) to 10 (= totally satisfied)
How satisfied are you with your current work situation?	0 (= totally dissatisfied) to 10 (= totally satisfied)
How satisfied are you with your current health?	0 (= totally dissatisfied) to 10 (= totally satisfied)
How do you rate yourself personally? In general, are you someone who is ready to take risks or do you try to avoid risks?	0 (= risk averse) to 10 (= fully prepared to take risks)
How likely is it that you will have a job in 2 years’ time in Germany?	0 (= not likely at all) to 10 (= definitely likely)
How likely is it that you will be working independently in 2 years’ time in Germany?	0 (= not likely at all) to 10 (= definitely likely)
How likely is it that you will attend a school in 2 years’ time in Germany?	0 (= not likely at all) to 10 (= definitely likely)
How likely is it that you will be attending training or a continuing professional development course in 2 years’ time in Germany?	0 (= not likely at all) to 10 (= definitely likely)
How likely is it that you will be studying at a higher education establishment in 2 years’ time in Germany?	0 (= not likely at all) to 10 (= definitely likely)
How well has political freedom been achieved currently in your country of origin?	0 (= very badly) to 10 (= very well)

2



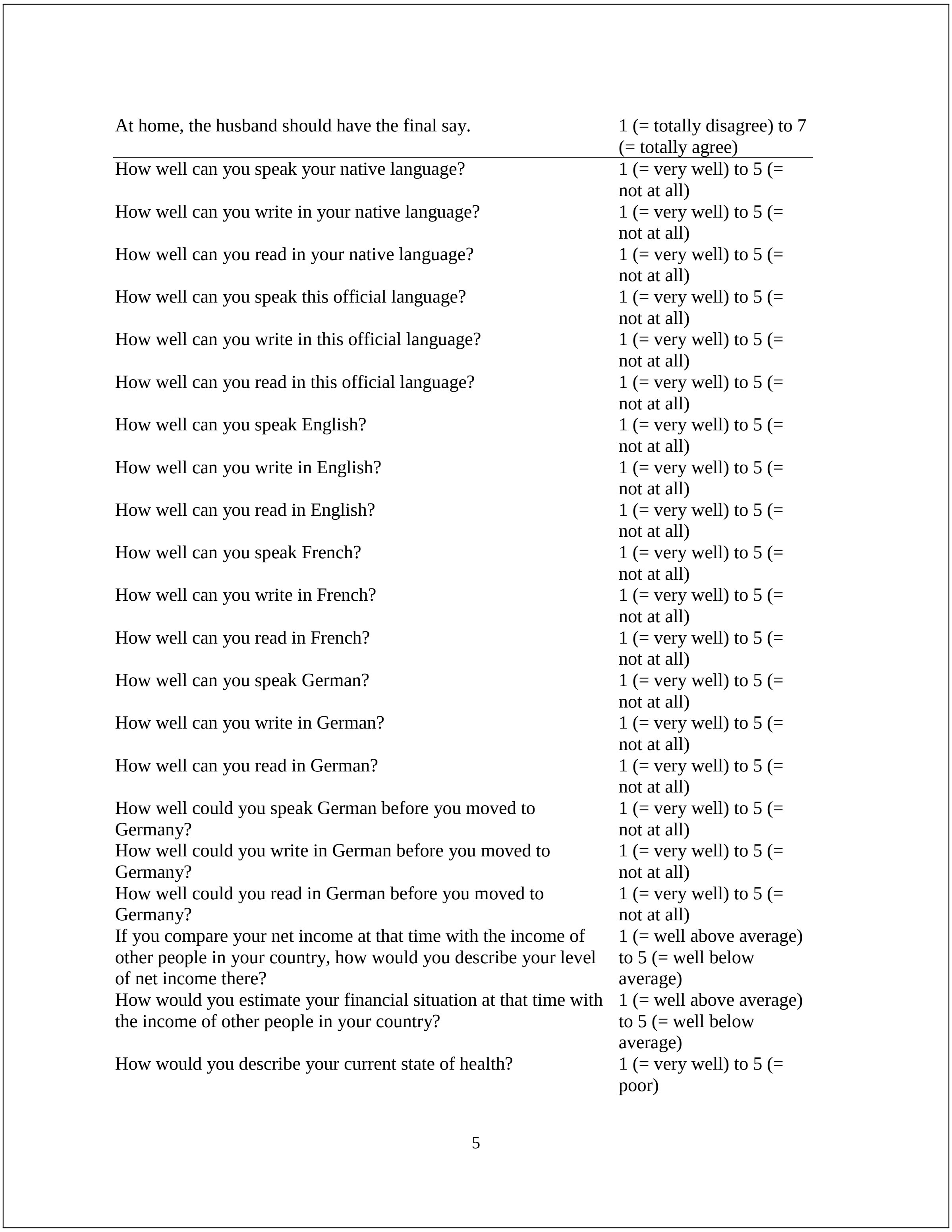
How well has civic freedom been achieved currently in your country of origin, such as the freedom to express opinions, right of assembly and an independent judiciary?	0 (= very badly) to 10 (= very well)
How well has freedom of the press and freedom of opinion been achieved currently in your country of origin?	0 (= very badly) to 10 (= very well)
How well is the right to practice religion or faith achieved currently in your country of origin?	0 (= very badly) to 10 (= very well)
How well is equal treatment of ethnic minorities achieved currently in your country of origin?	0 (= very badly) to 10 (= very well)
How well is equal treatment of men and women achieved currently in your country of origin?	0 (= very badly) to 10 (= very well)
The government taxes the rich and supports the poor.	0 (= should definitely not happen in a democracy) to 10 (= should definitely happen in a democracy)
Religious leaders ultimately determine the interpretation of laws.	0 (= should definitely not happen in a democracy) to 10 (= should definitely happen in a democracy)
The people choose their government in free elections.	0 (= should definitely not happen in a democracy) to 10 (= should definitely happen in a democracy)
Civil rights protect the people from government oppression.	0 (= should definitely not happen in a democracy) to 10 (= should definitely happen in a democracy)
Minorities are protected.	0 (= should definitely not happen in a democracy) to 10 (= should definitely happen in a democracy)
Women have the same rights as men.	0 (= should definitely not happen in a democracy) to 10 (= should definitely happen in a democracy)
How satisfied are you currently with your life in general?	0 (= totally dissatisfied) to 10 (=totally satisfied)
My life's direction depends on me.	1 (= totally disagree) to 7 (= totally agree)
In comparison with others, I haven't achieved what I deserved to achieve.	1 (= totally disagree) to 7 (= totally agree)
What can be achieved in life is mainly a result of fate or luck.	1 (= totally disagree) to 7 (= totally agree)
If you are socially or politically active, you can influence social circumstances.	1 (= totally disagree) to 7 (= totally agree)
I often find that other people dictate my life.	1 (= totally disagree) to 7 (= totally agree)

3



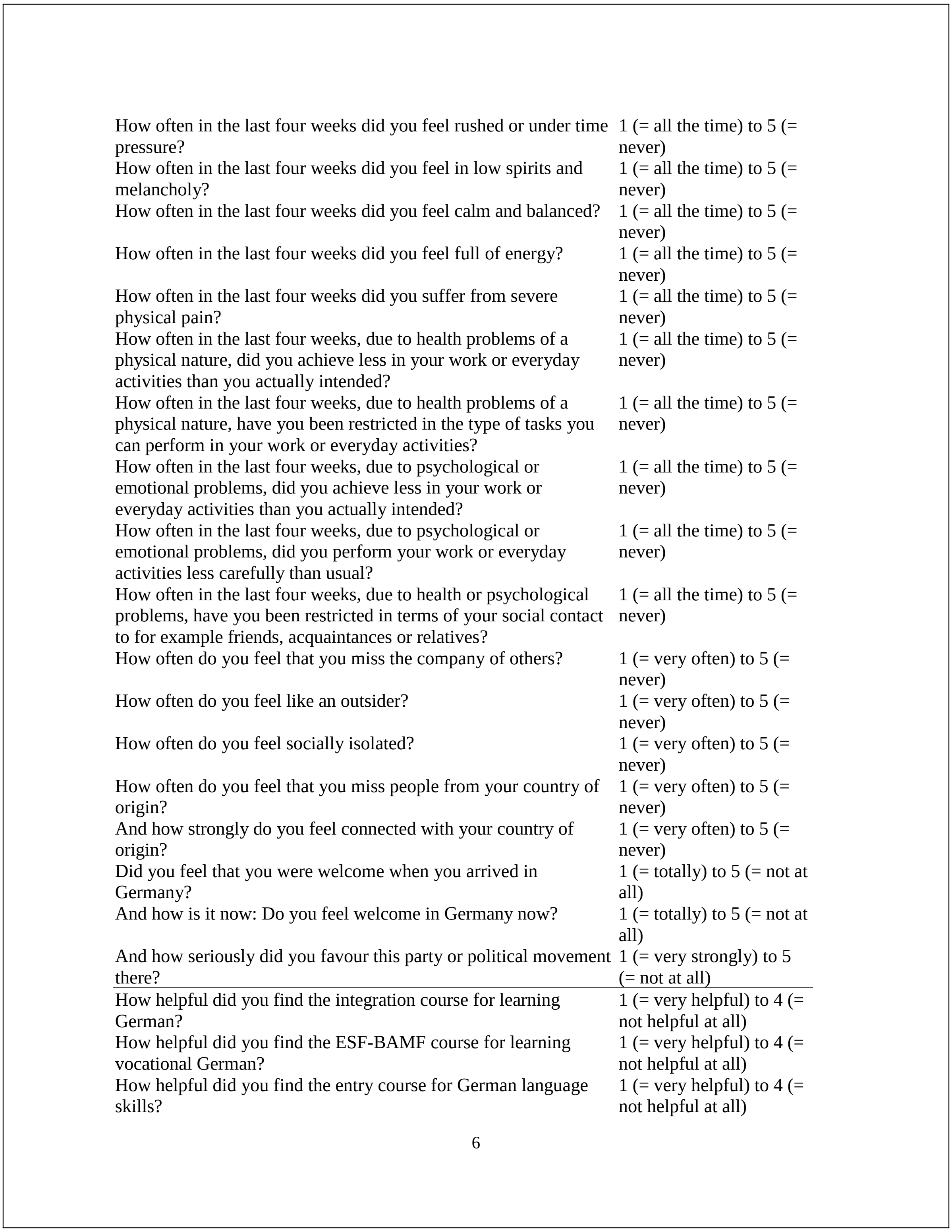
You must work hard to achieve success.	1 (= totally disagree) to 7 (= totally agree)
When I encounter difficulties in life, I often doubt my abilities.	1 (= totally disagree) to 7 (= totally agree)
The options that I have in life are determined by social circumstances.	1 (= totally disagree) to 7 (= totally agree)
The abilities we have are more important than the efforts we make.	1 (= totally disagree) to 7 (= totally agree)
I don’t have much control over what happens in my life.	1 (= totally disagree) to 7 (= totally agree)
If someone does me a favour, I am willing to reciprocate it.	1 (= totally disagree) to 7 (= totally agree)
If someone does me a serious wrong, I will get my own back at any price at the next opportunity.	1 (= totally disagree) to 7 (= totally agree)
If somebody puts me in a difficult position, I will do the same to them.	1 (= totally disagree) to 7 (= totally agree)
I make particular effort to help someone who has previously helped me.	1 (= totally disagree) to 7 (= totally agree)
If someone insults me, I will insult them.	1 (= totally disagree) to 7 (= totally agree)
I am prepared to incur costs myself to help someone who has previously helped me.	1 (= totally disagree) to 7 (= totally agree)
I have a positive attitude about myself.	1 (= totally disagree) to 7 (= totally agree)
I try to think of how I can change difficult situations.	1 (= totally disagree) to 7 (= totally agree)
No matter what happens to me, I think I have my reactions under control.	1 (= totally disagree) to 7 (= totally agree)
I think I can develop further if I deal with difficult situations.	1 (= totally disagree) to 7 (= totally agree)
I actively seek ways to balance out the losses that have affected me in my life.	1 (= totally disagree) to 7 (= totally agree)
You need a strong leader who does not have to be concerned with a Parliament or elections.	1 (= totally disagree) to 7 (= totally agree)
Experts, not the Government, should decide what is best for the country.	1 (= totally disagree) to 7 (= totally agree)
There should be a democratic system.	1 (= totally disagree) to 7 (= totally agree)
Having a job is the best way for a woman to be independent.	1 (= totally disagree) to 7 (= totally agree)
Even a married woman should have a paid job so that she can be financially independent.	1 (= totally disagree) to 7 (= totally agree)
If a woman earns more money than her partner, this inevitably leads to problems.	1 (= totally disagree) to 7 (= totally agree)
For parents, vocational training or higher education for their sons should be more important than vocational training or higher education for their daughters.	1 (= totally disagree) to 7 (= totally agree)

4



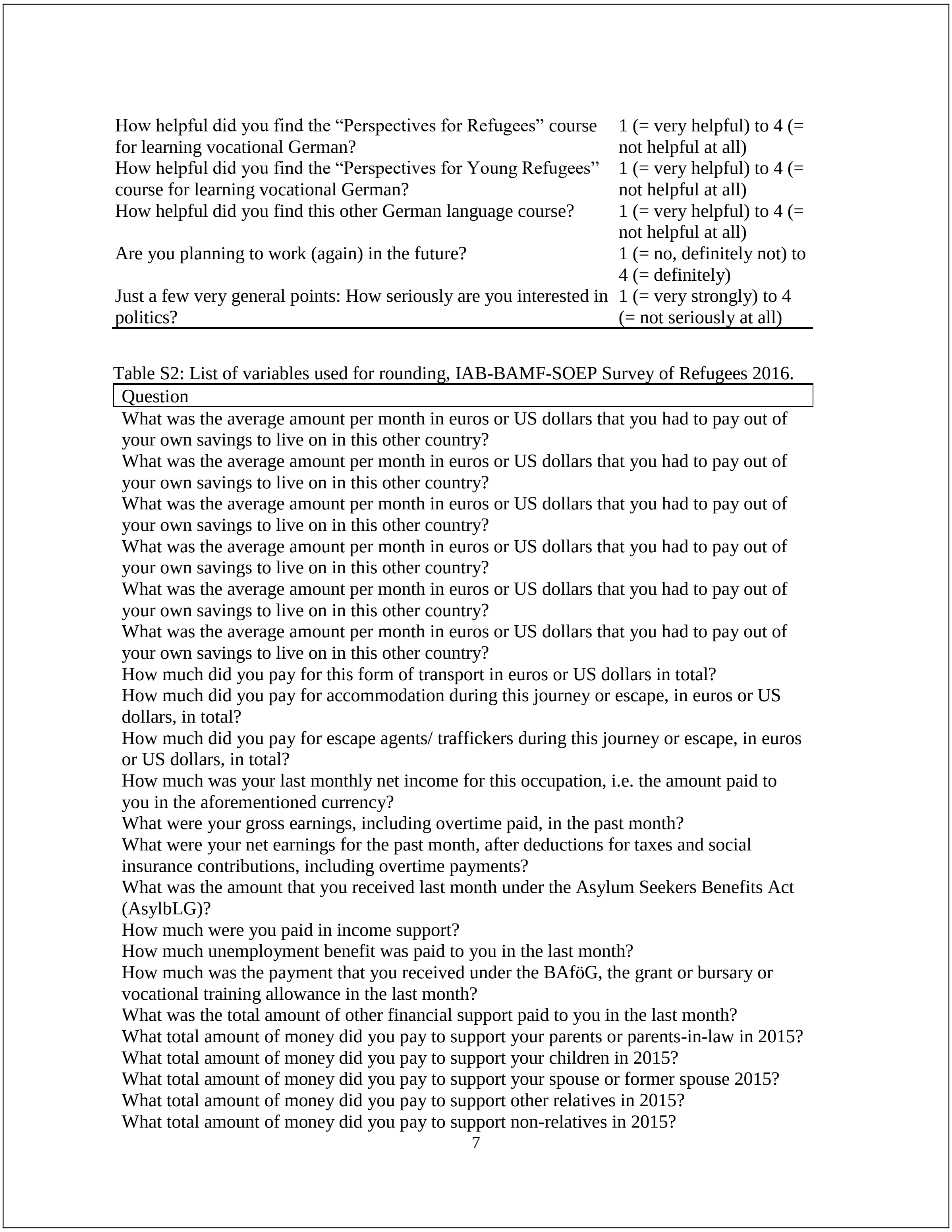
At home, the husband should have the final say.	1 (= totally disagree) to 7 (= totally agree)
How well can you speak your native language?	1 (= very well) to 5 (= not at all)
How well can you write in your native language?	1 (= very well) to 5 (= not at all)
How well can you read in your native language?	1 (= very well) to 5 (= not at all)
How well can you speak this official language?	1 (= very well) to 5 (= not at all)
How well can you write in this official language?	1 (= very well) to 5 (= not at all)
How well can you read in this official language?	1 (= very well) to 5 (= not at all)
How well can you speak English?	1 (= very well) to 5 (= not at all)
How well can you write in English?	1 (= very well) to 5 (= not at all)
How well can you read in English?	1 (= very well) to 5 (= not at all)
How well can you speak French?	1 (= very well) to 5 (= not at all)
How well can you write in French?	1 (= very well) to 5 (= not at all)
How well can you read in French?	1 (= very well) to 5 (= not at all)
How well can you speak German?	1 (= very well) to 5 (= not at all)
How well can you write in German?	1 (= very well) to 5 (= not at all)
How well can you read in German?	1 (= very well) to 5 (= not at all)
How well could you speak German before you moved to Germany?	1 (= very well) to 5 (= not at all)
How well could you write in German before you moved to Germany?	1 (= very well) to 5 (= not at all)
How well could you read in German before you moved to Germany?	1 (= very well) to 5 (= not at all)
If you compare your net income at that time with the income of other people in your country, how would you describe your level of net income there?	1 (= well above average) to 5 (= well below average)
How would you estimate your financial situation at that time with the income of other people in your country?	1 (= well above average) to 5 (= well below average)
How would you describe your current state of health?	1 (= very well) to 5 (= poor)

5



How often in the last four weeks did you feel rushed or under time pressure?	1 (= all the time) to 5 (= never)
How often in the last four weeks did you feel in low spirits and melancholy?	1 (= all the time) to 5 (= never)
How often in the last four weeks did you feel calm and balanced?	1 (= all the time) to 5 (= never)
How often in the last four weeks did you feel full of energy?	1 (= all the time) to 5 (= never)
How often in the last four weeks did you suffer from severe physical pain?	1 (= all the time) to 5 (= never)
How often in the last four weeks, due to health problems of a physical nature, did you achieve less in your work or everyday activities than you actually intended?	1 (= all the time) to 5 (= never)
How often in the last four weeks, due to health problems of a physical nature, have you been restricted in the type of tasks you can perform in your work or everyday activities?	1 (= all the time) to 5 (= never)
How often in the last four weeks, due to psychological or emotional problems, did you achieve less in your work or everyday activities than you actually intended?	1 (= all the time) to 5 (= never)
How often in the last four weeks, due to psychological or emotional problems, did you perform your work or everyday activities less carefully than usual?	1 (= all the time) to 5 (= never)
How often in the last four weeks, due to health or psychological problems, have you been restricted in terms of your social contact to for example friends, acquaintances or relatives?	1 (= all the time) to 5 (= never)
How often do you feel that you miss the company of others?	1 (= very often) to 5 (= never)
How often do you feel like an outsider?	1 (= very often) to 5 (= never)
How often do you feel socially isolated?	1 (= very often) to 5 (= never)
How often do you feel that you miss people from your country of origin?	1 (= very often) to 5 (= never)
And how strongly do you feel connected with your country of origin?	1 (= very often) to 5 (= never)
Did you feel that you were welcome when you arrived in Germany?	1 (= totally) to 5 (= not at all)
And how is it now: Do you feel welcome in Germany now?	1 (= totally) to 5 (= not at all)
And how seriously did you favour this party or political movement there?	1 (= very strongly) to 5 (= not at all)
How helpful did you find the integration course for learning German?	1 (= very helpful) to 4 (= not helpful at all)
How helpful did you find the ESF-BAMF course for learning vocational German?	1 (= very helpful) to 4 (= not helpful at all)
How helpful did you find the entry course for German language skills?	1 (= very helpful) to 4 (= not helpful at all)

6



How helpful did you find the “Perspectives for Refugees” course for learning vocational German?	1 (= very helpful) to 4 (= not helpful at all)
How helpful did you find the “Perspectives for Young Refugees” course for learning vocational German?	1 (= very helpful) to 4 (= not helpful at all)
How helpful did you find this other German language course?	1 (= very helpful) to 4 (= not helpful at all)
Are you planning to work (again) in the future?	1 (= no, definitely not) to 4 (= definitely)
Just a few very general points: How seriously are you interested in politics?	1 (= very strongly) to 4 (= not seriously at all)

Table S2: List of variables used for rounding, IAB-BAMF-SOEP Survey of Refugees 2016.

Question
What was the average amount per month in euros or US dollars that you had to pay out of your own savings to live on in this other country?
What was the average amount per month in euros or US dollars that you had to pay out of your own savings to live on in this other country?
What was the average amount per month in euros or US dollars that you had to pay out of your own savings to live on in this other country?
What was the average amount per month in euros or US dollars that you had to pay out of your own savings to live on in this other country?
What was the average amount per month in euros or US dollars that you had to pay out of your own savings to live on in this other country?
What was the average amount per month in euros or US dollars that you had to pay out of your own savings to live on in this other country?
How much did you pay for this form of transport in euros or US dollars in total?
How much did you pay for accommodation during this journey or escape, in euros or US dollars, in total?
How much did you pay for escape agents/ traffickers during this journey or escape, in euros or US dollars, in total?
How much was your last monthly net income for this occupation, i.e. the amount paid to you in the aforementioned currency?
What were your gross earnings, including overtime paid, in the past month?
What were your net earnings for the past month, after deductions for taxes and social insurance contributions, including overtime payments?
What was the amount that you received last month under the Asylum Seekers Benefits Act (AsylbLG)?
How much were you paid in income support?
How much unemployment benefit was paid to you in the last month?
How much was the payment that you received under the BAföG, the grant or bursary or vocational training allowance in the last month?
What was the total amount of other financial support paid to you in the last month?
What total amount of money did you pay to support your parents or parents-in-law in 2015?
What total amount of money did you pay to support your children in 2015?
What total amount of money did you pay to support your spouse or former spouse 2015?
What total amount of money did you pay to support other relatives in 2015?
What total amount of money did you pay to support non-relatives in 2015?

How many days did it take to travel from this country to Germany?
How many days did it take to travel from your country of birth to Germany?
How many days did you live in this accommodation?
How many people from your country of origin have you met since your arrival in Germany with whom you have regular contact?
How many German people have you met since your arrival in Germany with whom you have regular contact?
How many people from other countries have you met since your arrival in Germany with whom you have regular contact?

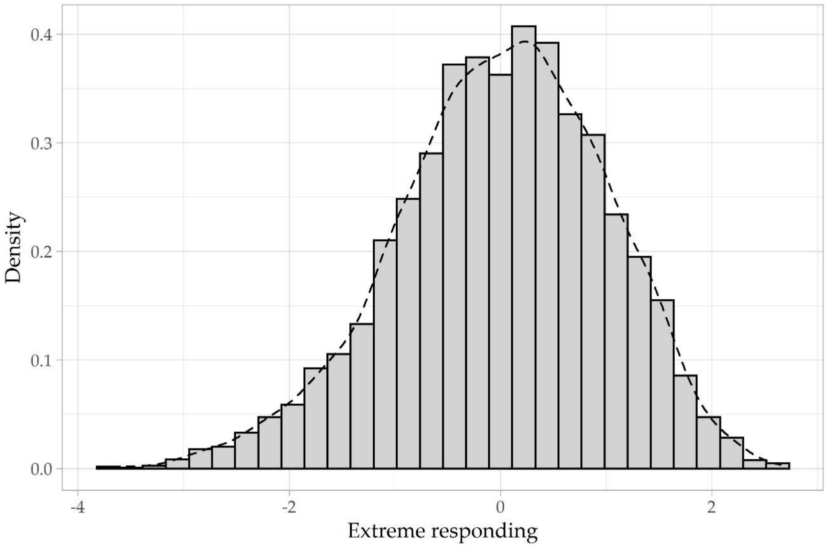


Figure S1: Distribution of extreme responding, IAB-BAMF-SOEP Survey of Refugees 2016.

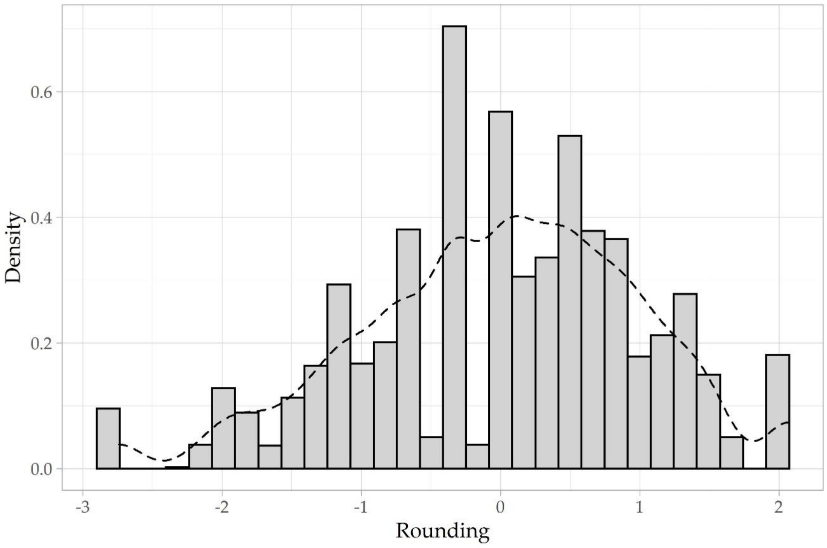


Figure S2: Distribution of rounding, IAB-BAMF-SOEP Survey of Refugees 2016.

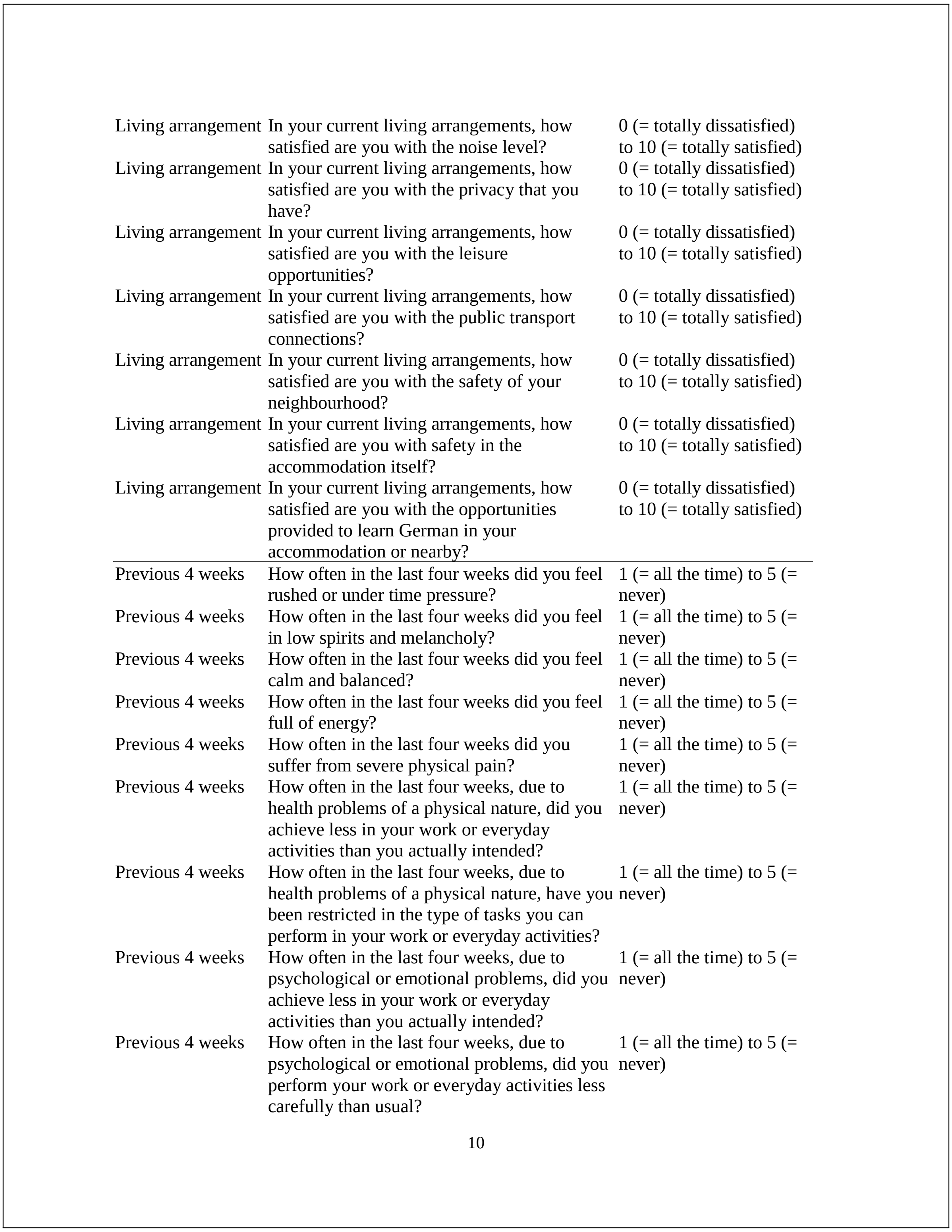
S2 Item batteries

Table S3: Item batteries, IAB-BAMF-SOEP Survey of Refugees 2016.

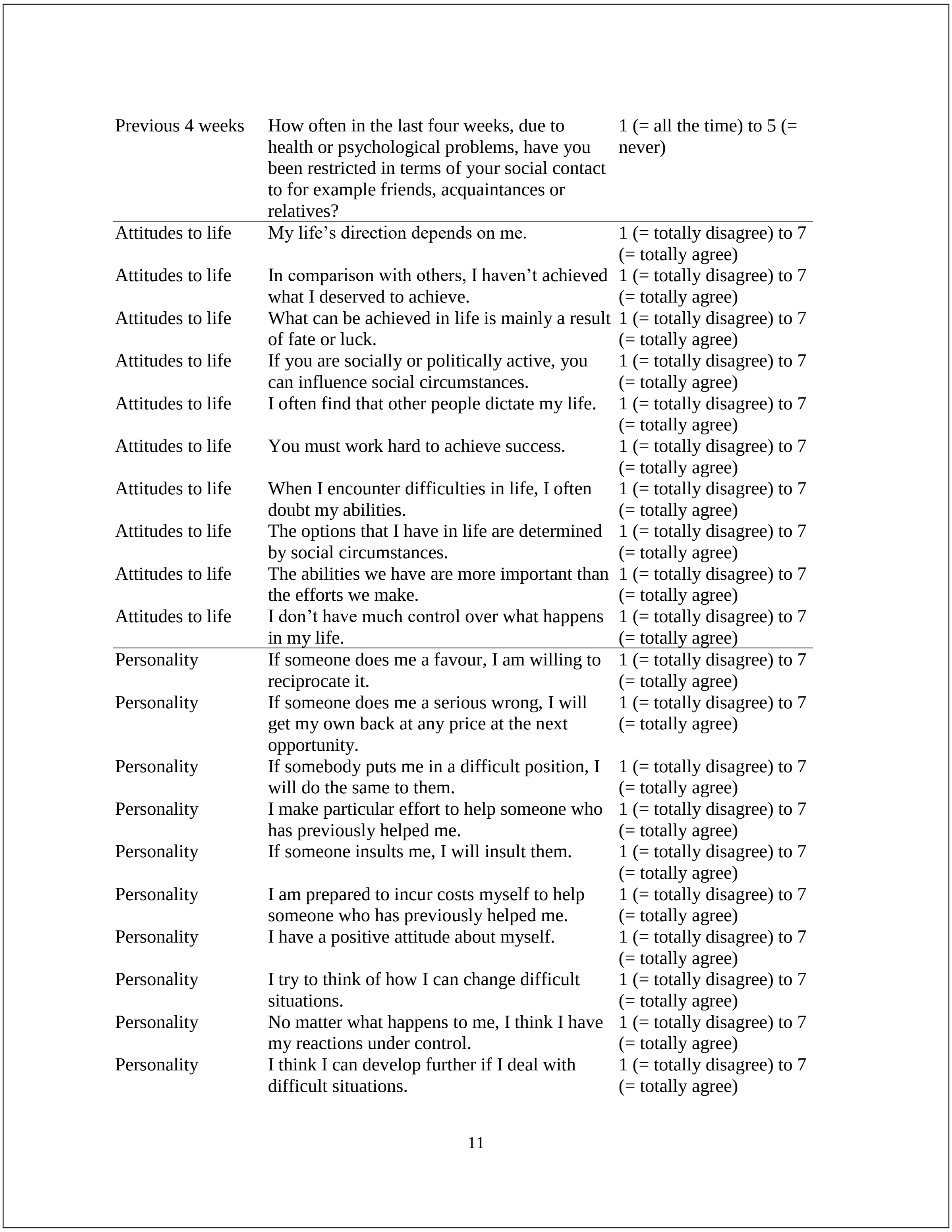
Topic	No. of items	Minimum of scale	Maximum of scale	N
Living arrangement	9	0 (= totally dissatisfied)	10 (= totally satisfied)	4,806
Previous 4 weeks	10	1 (= all the time)	5 (= never)	4,759
Attitudes to life	10	1 (= totally disagree)	7 (= totally agree)	4,575
Personality	11	1 (= totally disagree)	7 (= totally agree)	4,663
Situation in 2 years	5	0 (= not likely at all)	10 (= definitely likely)	4,656
Country of origin	6	0 (= very badly)	10 (= very well)	4,429
Democracy	6	0 (= should definitely not happen in a democracy)	10 (= should definitely happen in a democracy)	4,373
Woman	5	1 (= totally disagree)	7 (= totally agree)	4,558

Table S4: List of variables used for response differentiation, IAB-BAMF-SOEP Survey of Refugees 2016.

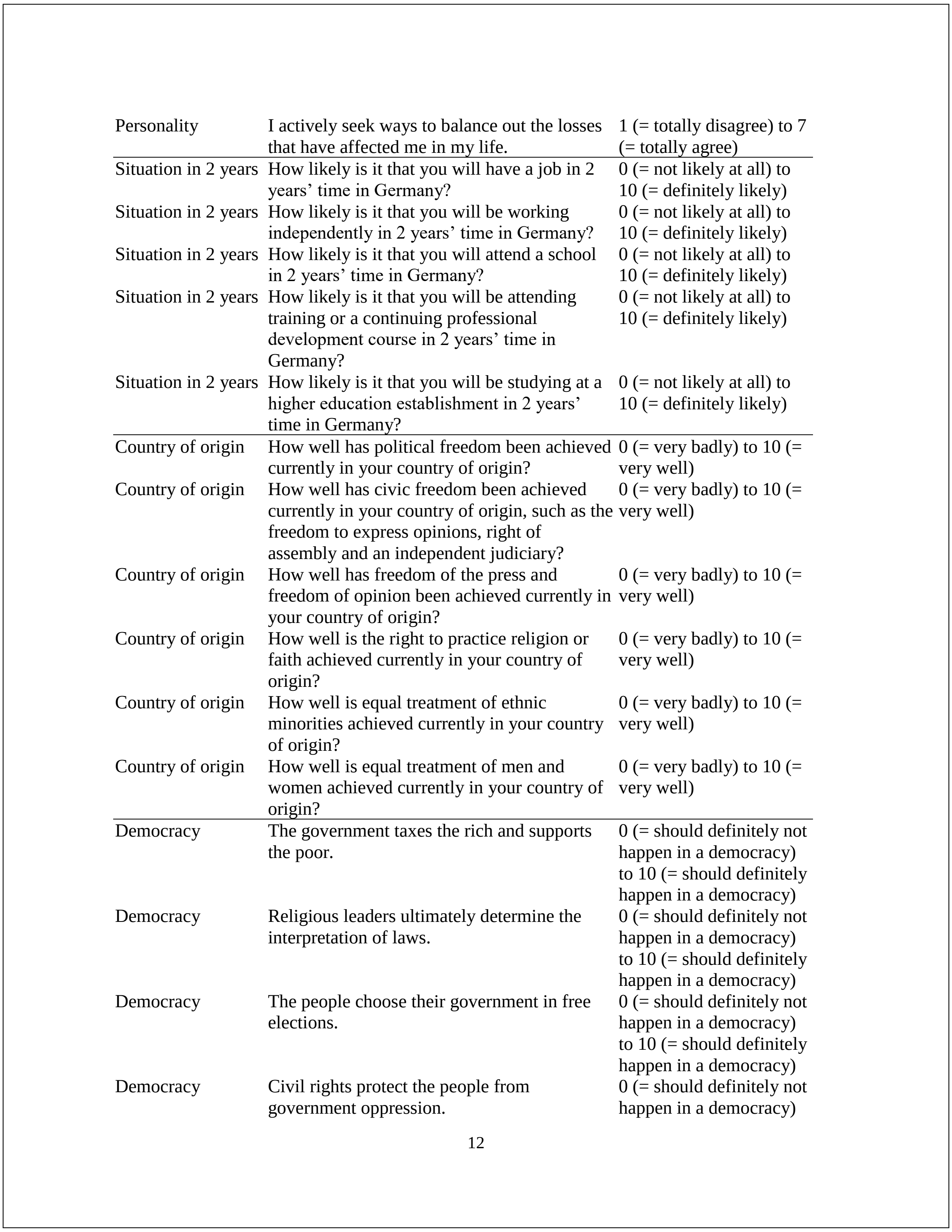
Battery	Question	Response scale
Living arrangement	How satisfied are you in general with your current living arrangements?	0 (= totally dissatisfied) to 10 (= totally satisfied)
Living arrangement	In your current living arrangements, how satisfied are you with the quality of the food?	0 (= totally dissatisfied) to 10 (= totally satisfied)



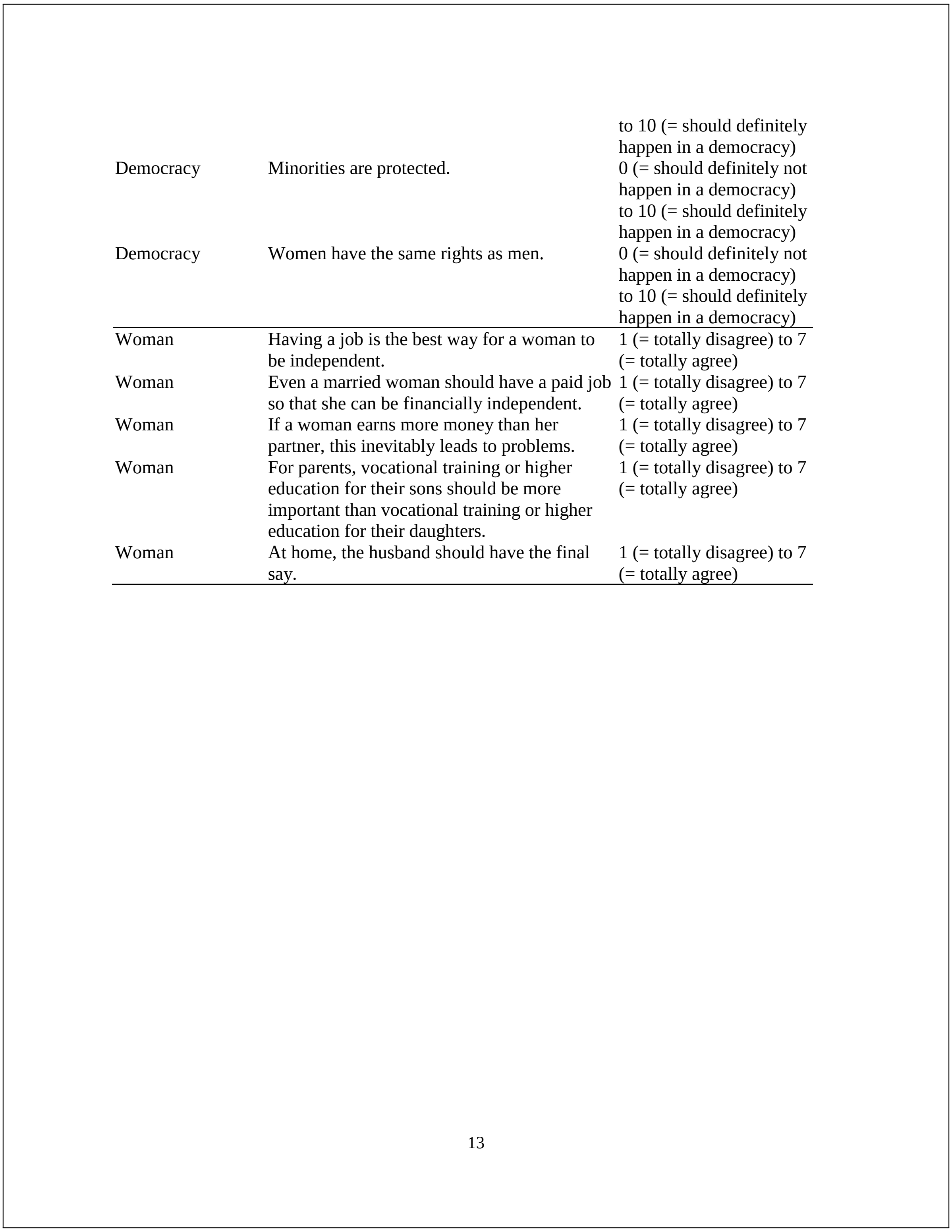
Living arrangement	In your current living arrangements, how satisfied are you with the noise level?	0 (= totally dissatisfied) to 10 (= totally satisfied)
Living arrangement	In your current living arrangements, how satisfied are you with the privacy that you have?	0 (= totally dissatisfied) to 10 (= totally satisfied)
Living arrangement	In your current living arrangements, how satisfied are you with the leisure opportunities?	0 (= totally dissatisfied) to 10 (= totally satisfied)
Living arrangement	In your current living arrangements, how satisfied are you with the public transport connections?	0 (= totally dissatisfied) to 10 (= totally satisfied)
Living arrangement	In your current living arrangements, how satisfied are you with the safety of your neighbourhood?	0 (= totally dissatisfied) to 10 (= totally satisfied)
Living arrangement	In your current living arrangements, how satisfied are you with safety in the accommodation itself?	0 (= totally dissatisfied) to 10 (= totally satisfied)
Living arrangement	In your current living arrangements, how satisfied are you with the opportunities provided to learn German in your accommodation or nearby?	0 (= totally dissatisfied) to 10 (= totally satisfied)
Previous 4 weeks	How often in the last four weeks did you feel rushed or under time pressure?	1 (= all the time) to 5 (= never)
Previous 4 weeks	How often in the last four weeks did you feel in low spirits and melancholy?	1 (= all the time) to 5 (= never)
Previous 4 weeks	How often in the last four weeks did you feel calm and balanced?	1 (= all the time) to 5 (= never)
Previous 4 weeks	How often in the last four weeks did you feel full of energy?	1 (= all the time) to 5 (= never)
Previous 4 weeks	How often in the last four weeks did you suffer from severe physical pain?	1 (= all the time) to 5 (= never)
Previous 4 weeks	How often in the last four weeks, due to health problems of a physical nature, did you achieve less in your work or everyday activities than you actually intended?	1 (= all the time) to 5 (= never)
Previous 4 weeks	How often in the last four weeks, due to health problems of a physical nature, have you been restricted in the type of tasks you can perform in your work or everyday activities?	1 (= all the time) to 5 (= never)
Previous 4 weeks	How often in the last four weeks, due to psychological or emotional problems, did you achieve less in your work or everyday activities than you actually intended?	1 (= all the time) to 5 (= never)
Previous 4 weeks	How often in the last four weeks, due to psychological or emotional problems, did you perform your work or everyday activities less carefully than usual?	1 (= all the time) to 5 (= never)



Previous 4 weeks	How often in the last four weeks, due to health or psychological problems, have you been restricted in terms of your social contact to for example friends, acquaintances or relatives?	1 (= all the time) to 5 (= never)
Attitudes to life	My life's direction depends on me.	1 (= totally disagree) to 7 (= totally agree)
Attitudes to life	In comparison with others, I haven't achieved what I deserved to achieve.	1 (= totally disagree) to 7 (= totally agree)
Attitudes to life	What can be achieved in life is mainly a result of fate or luck.	1 (= totally disagree) to 7 (= totally agree)
Attitudes to life	If you are socially or politically active, you can influence social circumstances.	1 (= totally disagree) to 7 (= totally agree)
Attitudes to life	I often find that other people dictate my life.	1 (= totally disagree) to 7 (= totally agree)
Attitudes to life	You must work hard to achieve success.	1 (= totally disagree) to 7 (= totally agree)
Attitudes to life	When I encounter difficulties in life, I often doubt my abilities.	1 (= totally disagree) to 7 (= totally agree)
Attitudes to life	The options that I have in life are determined by social circumstances.	1 (= totally disagree) to 7 (= totally agree)
Attitudes to life	The abilities we have are more important than the efforts we make.	1 (= totally disagree) to 7 (= totally agree)
Attitudes to life	I don't have much control over what happens in my life.	1 (= totally disagree) to 7 (= totally agree)
Personality	If someone does me a favour, I am willing to reciprocate it.	1 (= totally disagree) to 7 (= totally agree)
Personality	If someone does me a serious wrong, I will get my own back at any price at the next opportunity.	1 (= totally disagree) to 7 (= totally agree)
Personality	If somebody puts me in a difficult position, I will do the same to them.	1 (= totally disagree) to 7 (= totally agree)
Personality	I make particular effort to help someone who has previously helped me.	1 (= totally disagree) to 7 (= totally agree)
Personality	If someone insults me, I will insult them.	1 (= totally disagree) to 7 (= totally agree)
Personality	I am prepared to incur costs myself to help someone who has previously helped me.	1 (= totally disagree) to 7 (= totally agree)
Personality	I have a positive attitude about myself.	1 (= totally disagree) to 7 (= totally agree)
Personality	I try to think of how I can change difficult situations.	1 (= totally disagree) to 7 (= totally agree)
Personality	No matter what happens to me, I think I have my reactions under control.	1 (= totally disagree) to 7 (= totally agree)
Personality	I think I can develop further if I deal with difficult situations.	1 (= totally disagree) to 7 (= totally agree)



Personality	I actively seek ways to balance out the losses that have affected me in my life.	1 (= totally disagree) to 7 (= totally agree)
Situation in 2 years	How likely is it that you will have a job in 2 years’ time in Germany?	0 (= not likely at all) to 10 (= definitely likely)
Situation in 2 years	How likely is it that you will be working independently in 2 years’ time in Germany?	0 (= not likely at all) to 10 (= definitely likely)
Situation in 2 years	How likely is it that you will attend a school in 2 years’ time in Germany?	0 (= not likely at all) to 10 (= definitely likely)
Situation in 2 years	How likely is it that you will be attending training or a continuing professional development course in 2 years’ time in Germany?	0 (= not likely at all) to 10 (= definitely likely)
Situation in 2 years	How likely is it that you will be studying at a higher education establishment in 2 years’ time in Germany?	0 (= not likely at all) to 10 (= definitely likely)
Country of origin	How well has political freedom been achieved currently in your country of origin?	0 (= very badly) to 10 (= very well)
Country of origin	How well has civic freedom been achieved currently in your country of origin, such as the freedom to express opinions, right of assembly and an independent judiciary?	0 (= very badly) to 10 (= very well)
Country of origin	How well has freedom of the press and freedom of opinion been achieved currently in your country of origin?	0 (= very badly) to 10 (= very well)
Country of origin	How well is the right to practice religion or faith achieved currently in your country of origin?	0 (= very badly) to 10 (= very well)
Country of origin	How well is equal treatment of ethnic minorities achieved currently in your country of origin?	0 (= very badly) to 10 (= very well)
Country of origin	How well is equal treatment of men and women achieved currently in your country of origin?	0 (= very badly) to 10 (= very well)
Democracy	The government taxes the rich and supports the poor.	0 (= should definitely not happen in a democracy) to 10 (= should definitely happen in a democracy)
Democracy	Religious leaders ultimately determine the interpretation of laws.	0 (= should definitely not happen in a democracy) to 10 (= should definitely happen in a democracy)
Democracy	The people choose their government in free elections.	0 (= should definitely not happen in a democracy) to 10 (= should definitely happen in a democracy)
Democracy	Civil rights protect the people from government oppression.	0 (= should definitely not happen in a democracy)



		to 10 (= should definitely happen in a democracy)
Democracy	Minorities are protected.	0 (= should definitely not happen in a democracy) to 10 (= should definitely happen in a democracy)
Democracy	Women have the same rights as men.	0 (= should definitely not happen in a democracy) to 10 (= should definitely happen in a democracy)
Woman	Having a job is the best way for a woman to be independent.	1 (= totally disagree) to 7 (= totally agree)
Woman	Even a married woman should have a paid job so that she can be financially independent.	1 (= totally disagree) to 7 (= totally agree)
Woman	If a woman earns more money than her partner, this inevitably leads to problems.	1 (= totally disagree) to 7 (= totally agree)
Woman	For parents, vocational training or higher education for their sons should be more important than vocational training or higher education for their daughters.	1 (= totally disagree) to 7 (= totally agree)
Woman	At home, the husband should have the final say.	1 (= totally disagree) to 7 (= totally agree)

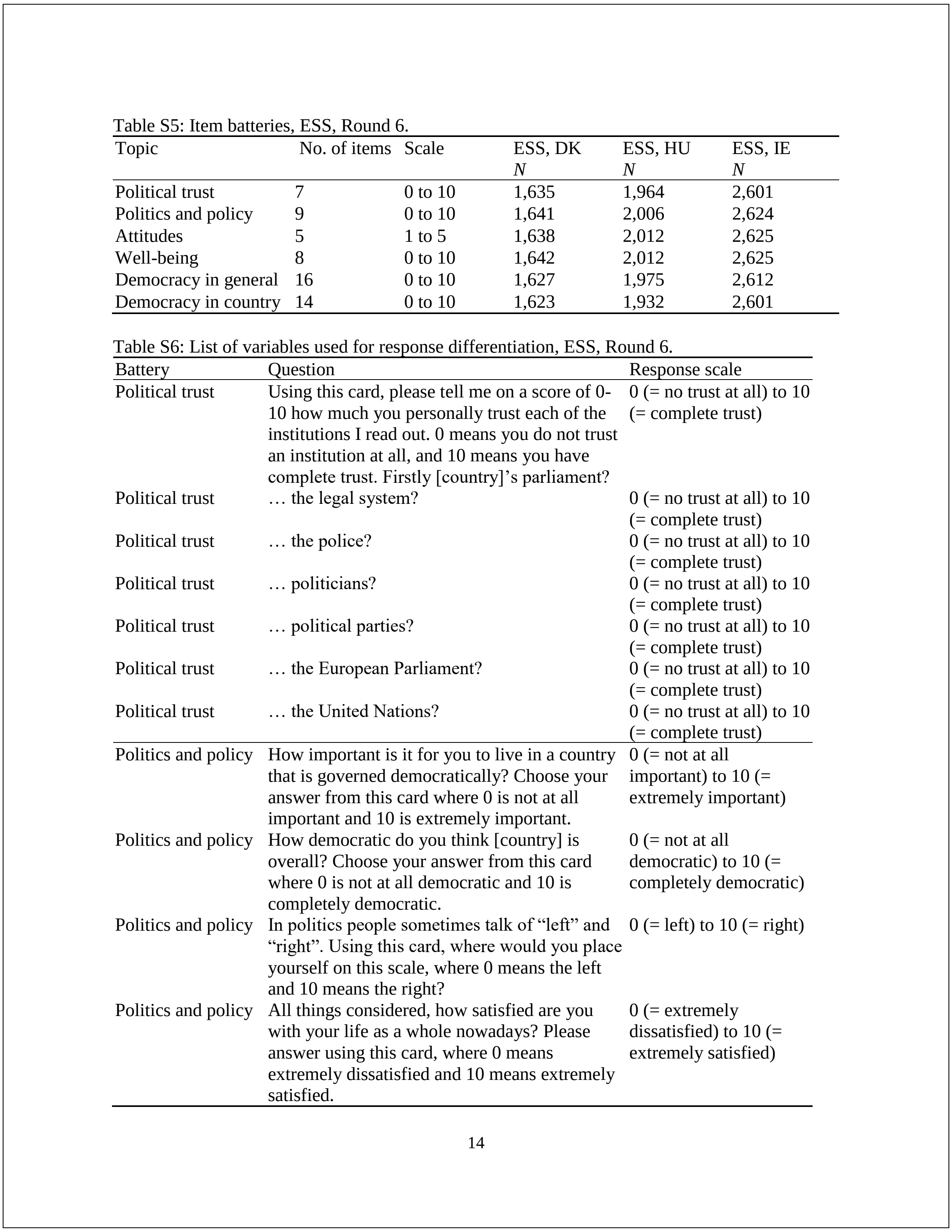
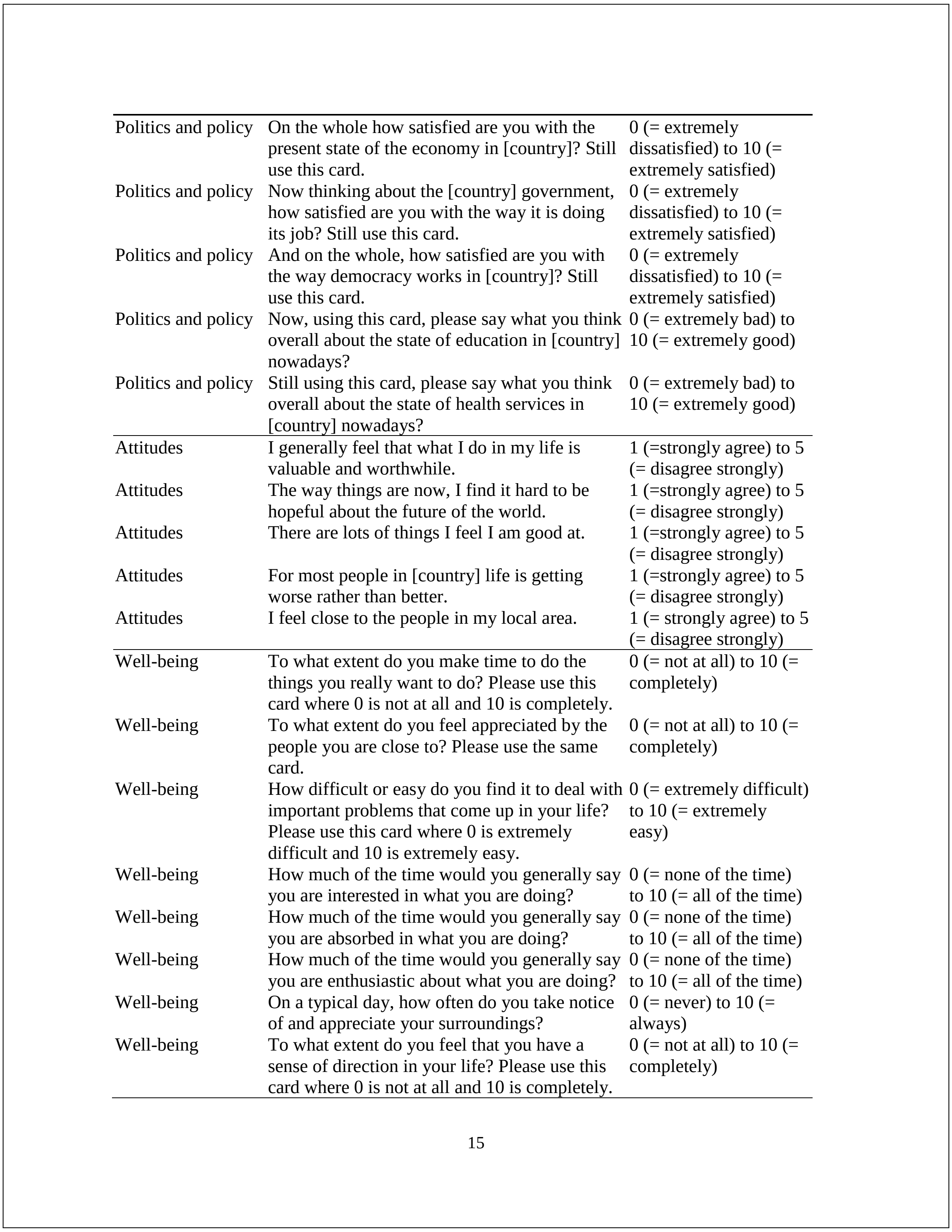


Table S5: Item batteries, ESS, Round 6.

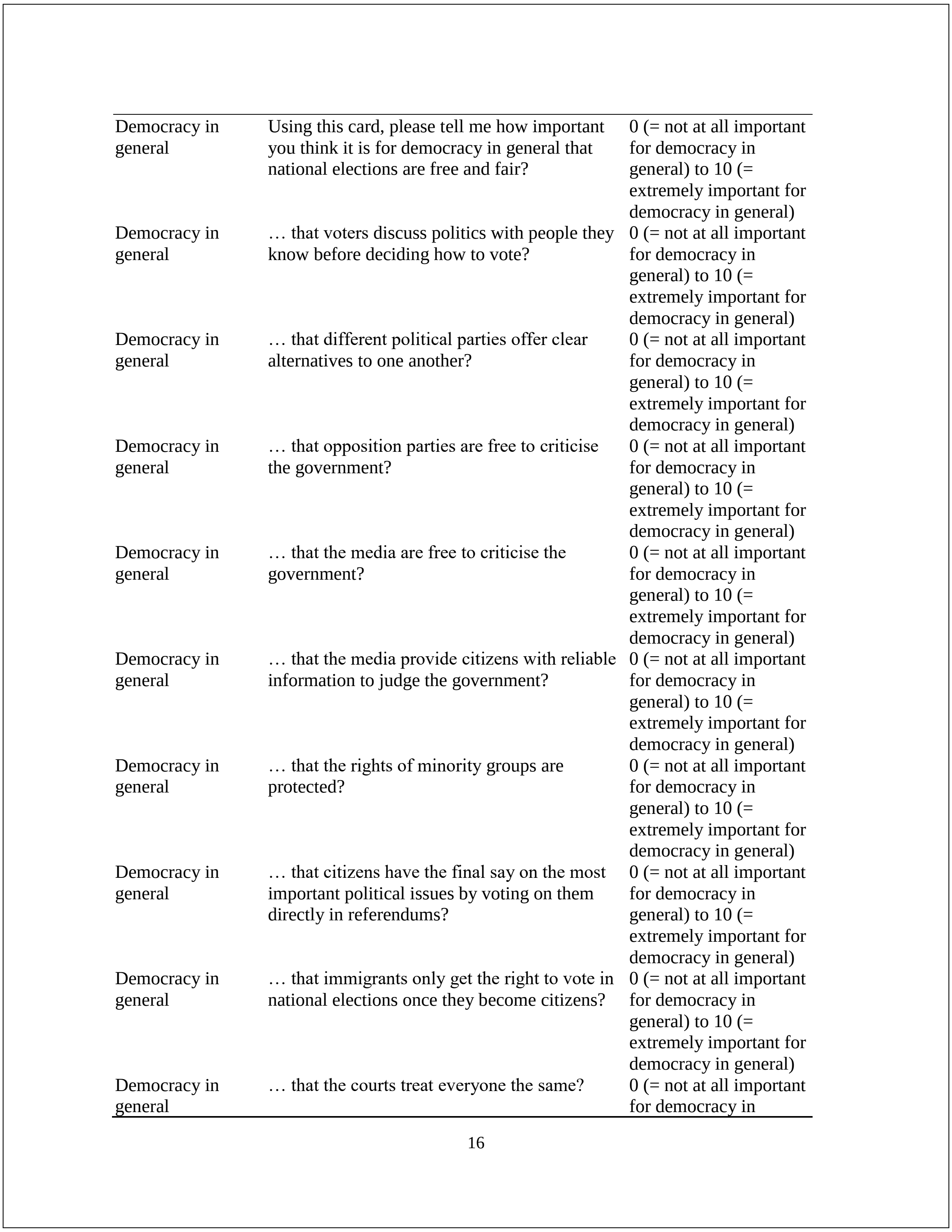
Topic	No. of items	Scale	ESS, DK *N*	ESS, HU *N*	ESS, IE *N*
Political trust	7	0 to 10	1,635	1,964	2,601
Politics and policy	9	0 to 10	1,641	2,006	2,624
Attitudes	5	1 to 5	1,638	2,012	2,625
Well-being	8	0 to 10	1,642	2,012	2,625
Democracy in general	16	0 to 10	1,627	1,975	2,612
Democracy in country	14	0 to 10	1,623	1,932	2,601

Table S6: List of variables used for response differentiation, ESS, Round 6.

Battery	Question	Response scale
Political trust	Using this card, please tell me on a score of 0-10 how much you personally trust each of the institutions I read out. 0 means you do not trust an institution at all, and 10 means you have complete trust. Firstly [country]'s parliament?	0 (= no trust at all) to 10 (= complete trust)
Political trust	… the legal system?	0 (= no trust at all) to 10 (= complete trust)
Political trust	… the police?	0 (= no trust at all) to 10 (= complete trust)
Political trust	… politicians?	0 (= no trust at all) to 10 (= complete trust)
Political trust	… political parties?	0 (= no trust at all) to 10 (= complete trust)
Political trust	… the European Parliament?	0 (= no trust at all) to 10 (= complete trust)
Political trust	… the United Nations?	0 (= no trust at all) to 10 (= complete trust)
Politics and policy	How important is it for you to live in a country that is governed democratically? Choose your answer from this card where 0 is not at all important and 10 is extremely important.	0 (= not at all important) to 10 (= extremely important)
Politics and policy	How democratic do you think [country] is overall? Choose your answer from this card where 0 is not at all democratic and 10 is completely democratic.	0 (= not at all democratic) to 10 (= completely democratic)
Politics and policy	In politics people sometimes talk of "left" and "right". Using this card, where would you place yourself on this scale, where 0 means the left and 10 means the right?	0 (= left) to 10 (= right)
Politics and policy	All things considered, how satisfied are you with your life as a whole nowadays? Please answer using this card, where 0 means extremely dissatisfied and 10 means extremely satisfied.	0 (= extremely dissatisfied) to 10 (= extremely satisfied)

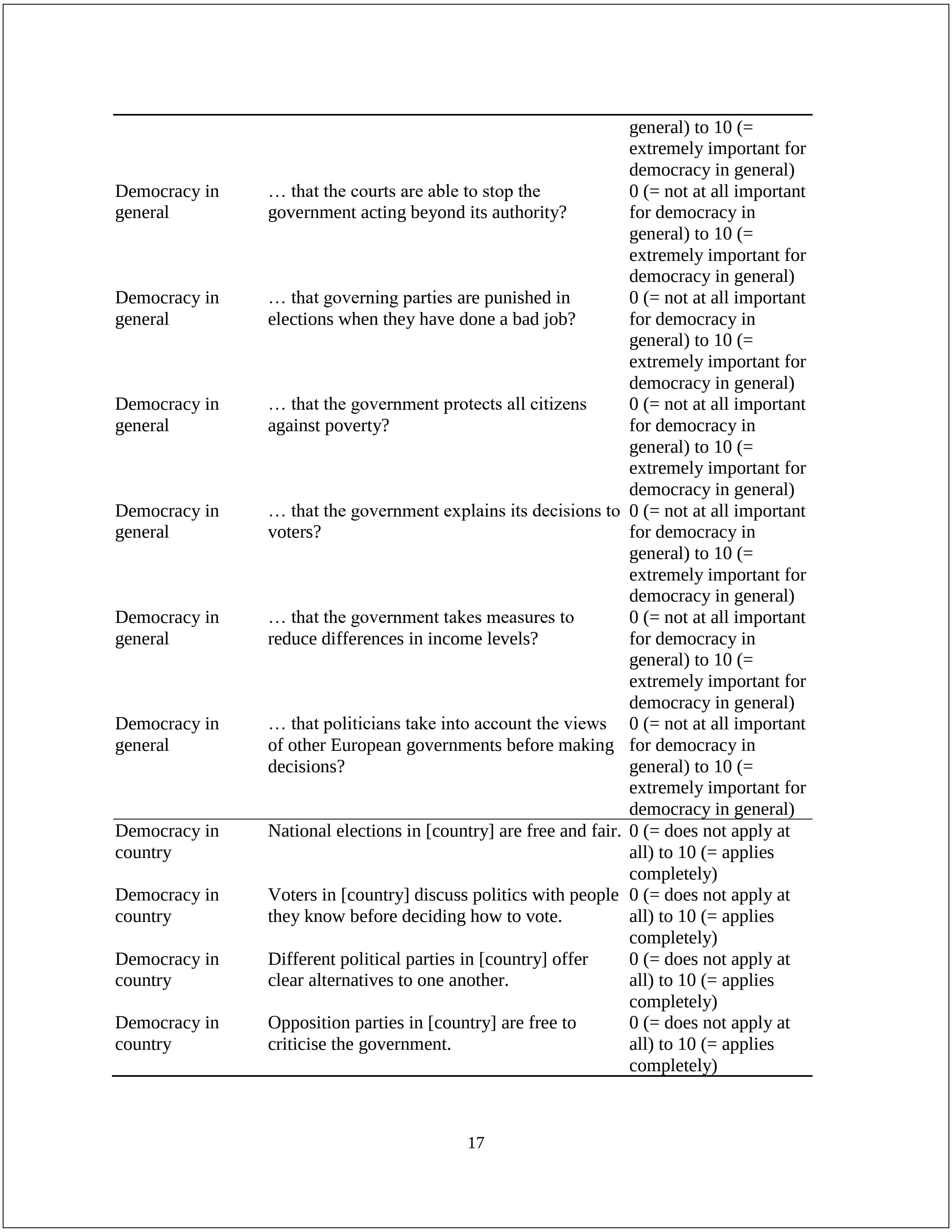


Politics and policy	On the whole how satisfied are you with the present state of the economy in [country]? Still use this card.	0 (= extremely dissatisfied) to 10 (= extremely satisfied)
Politics and policy	Now thinking about the [country] government, how satisfied are you with the way it is doing its job? Still use this card.	0 (= extremely dissatisfied) to 10 (= extremely satisfied)
Politics and policy	And on the whole, how satisfied are you with the way democracy works in [country]? Still use this card.	0 (= extremely dissatisfied) to 10 (= extremely satisfied)
Politics and policy	Now, using this card, please say what you think overall about the state of education in [country] nowadays?	0 (= extremely bad) to 10 (= extremely good)
Politics and policy	Still using this card, please say what you think overall about the state of health services in [country] nowadays?	0 (= extremely bad) to 10 (= extremely good)
Attitudes	I generally feel that what I do in my life is valuable and worthwhile.	1 (=strongly agree) to 5 (= disagree strongly)
Attitudes	The way things are now, I find it hard to be hopeful about the future of the world.	1 (=strongly agree) to 5 (= disagree strongly)
Attitudes	There are lots of things I feel I am good at.	1 (=strongly agree) to 5 (= disagree strongly)
Attitudes	For most people in [country] life is getting worse rather than better.	1 (=strongly agree) to 5 (= disagree strongly)
Attitudes	I feel close to the people in my local area.	1 (= strongly agree) to 5 (= disagree strongly)
Well-being	To what extent do you make time to do the things you really want to do? Please use this card where 0 is not at all and 10 is completely.	0 (= not at all) to 10 (= completely)
Well-being	To what extent do you feel appreciated by the people you are close to? Please use the same card.	0 (= not at all) to 10 (= completely)
Well-being	How difficult or easy do you find it to deal with important problems that come up in your life? Please use this card where 0 is extremely difficult and 10 is extremely easy.	0 (= extremely difficult) to 10 (= extremely easy)
Well-being	How much of the time would you generally say you are interested in what you are doing?	0 (= none of the time) to 10 (= all of the time)
Well-being	How much of the time would you generally say you are absorbed in what you are doing?	0 (= none of the time) to 10 (= all of the time)
Well-being	How much of the time would you generally say you are enthusiastic about what you are doing?	0 (= none of the time) to 10 (= all of the time)
Well-being	On a typical day, how often do you take notice of and appreciate your surroundings?	0 (= never) to 10 (= always)
Well-being	To what extent do you feel that you have a sense of direction in your life? Please use this card where 0 is not at all and 10 is completely.	0 (= not at all) to 10 (= completely)



Democracy in general	Using this card, please tell me how important you think it is for democracy in general that national elections are free and fair?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that voters discuss politics with people they know before deciding how to vote?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that different political parties offer clear alternatives to one another?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that opposition parties are free to criticise the government?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that the media are free to criticise the government?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that the media provide citizens with reliable information to judge the government?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that the rights of minority groups are protected?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that citizens have the final say on the most important political issues by voting on them directly in referendums?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that immigrants only get the right to vote in national elections once they become citizens?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that the courts treat everyone the same?	0 (= not at all important for democracy in

16



		general) to 10 (= extremely important for democracy in general)
Democracy in general	… that the courts are able to stop the government acting beyond its authority?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that governing parties are punished in elections when they have done a bad job?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that the government protects all citizens against poverty?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that the government explains its decisions to voters?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that the government takes measures to reduce differences in income levels?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in general	… that politicians take into account the views of other European governments before making decisions?	0 (= not at all important for democracy in general) to 10 (= extremely important for democracy in general)
Democracy in country	National elections in [country] are free and fair.	0 (= does not apply at all) to 10 (= applies completely)
Democracy in country	Voters in [country] discuss politics with people they know before deciding how to vote.	0 (= does not apply at all) to 10 (= applies completely)
Democracy in country	Different political parties in [country] offer clear alternatives to one another.	0 (= does not apply at all) to 10 (= applies completely)
Democracy in country	Opposition parties in [country] are free to criticise the government.	0 (= does not apply at all) to 10 (= applies completely)

Democracy in country	The media in [country] are free to criticize the government.	0 (= does not apply at all) to 10 (= applies completely)
Democracy in country	The media in [country] provide citizens with reliable information to judge the government.	0 (= does not apply at all) to 10 (= applies completely)
Democracy in country	The rights of minority groups in [country] are protected.	0 (= does not apply at all) to 10 (= applies completely)
Democracy in country	Citizens in [country] have the final say on the most important political issues by voting on them directly in referendums.	0 (= does not apply at all) to 10 (= applies completely)
Democracy in country	The courts in [country] treat everyone the same.	0 (= does not apply at all) to 10 (= applies completely)
Democracy in country	Governing parties in [country] are punished in elections when they have done a bad job.	0 (= does not apply at all) to 10 (= applies completely)
Democracy in country	The government in [country] protects all citizens against poverty.	0 (= does not apply at all) to 10 (= applies completely)
Democracy in country	The government in [country] explains its decisions to voters.	0 (= does not apply at all) to 10 (= applies completely)
Democracy in country	The government in [country] takes measures to reduce differences in income levels.	0 (= does not apply at all) to 10 (= applies completely)
Democracy in country	Politicians in [country] take into account the views of other European governments before making decisions.	0 (= does not apply at all) to 10 (= applies completely)

S3 Distribution of response differentiation

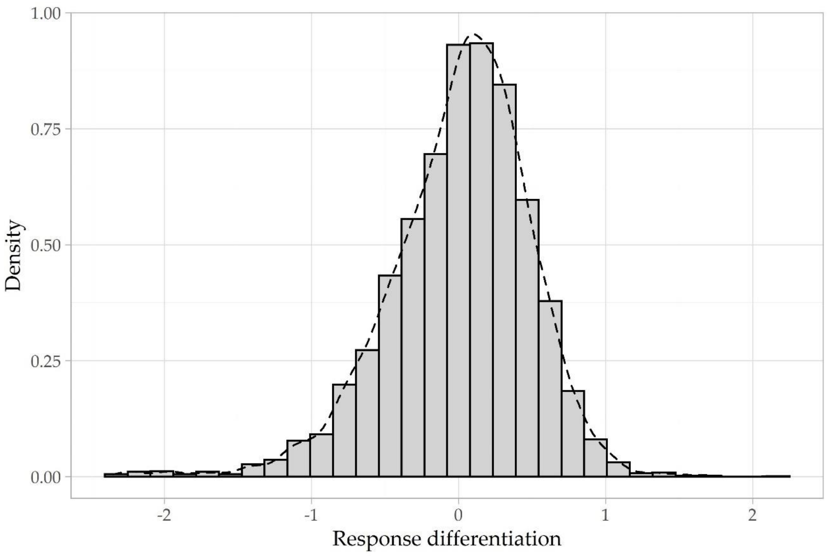


Figure S3: Distribution of response differentiation, IAB-BAMF-SOEP Survey of Refugees 2016.

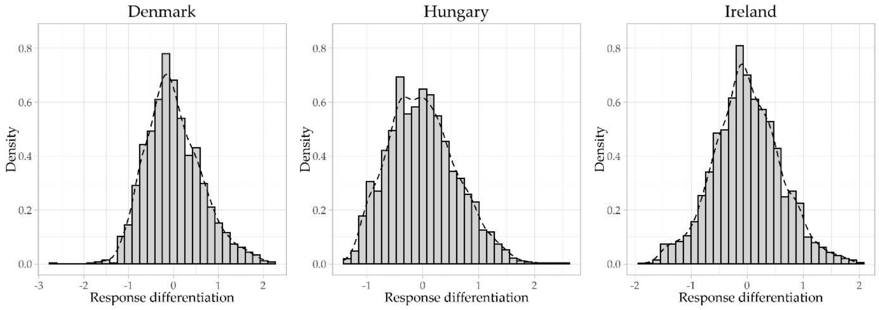


Figure S4: Distribution of response differentiation, ESS, Round 6.

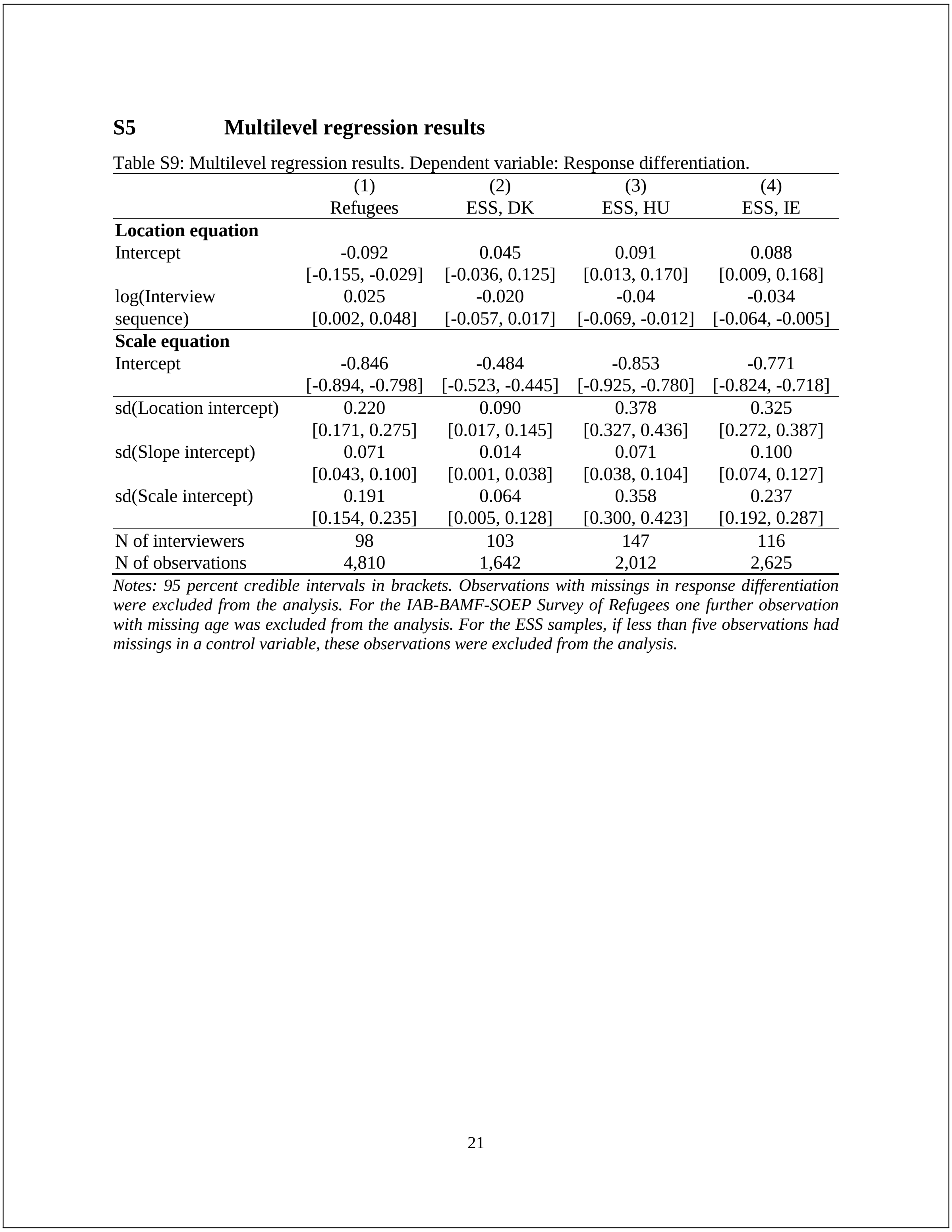
S4 Descriptive statistics of covariates

Table S7: Descriptive statistics, IAB-BAMF-SOEP Survey of Refugees 2016.

	Full sample	Without falsifiers
Variable	%/mean	%/mean
Age	33.512	33.590
Gender: Female	38.981	37.878
Gender: Male	61.019	62.122
Region: Rural	32.703	31.778
Region: Urban	67.297	68.222
Education: In school/Primary	35.052	36.981
Education: Secondary	38.773	36.241
Education: Post-secondary, tertiary and higher	18.773	19.377
Accommodation: Private	33.056	32.675
Accommodation: Shared	66.486	66.854
N	4,810	4,459
N of interviewers	98	95
N of regions	16	16

Table S8: Descriptive statistics, ESS, Round 6.

	ESS, DK	ESS, HU	ESS, IE
Variable	%/mean	%/mean	%/mean
Age	48.689	47.155	47.262
Gender: Female	49.574	55.219	52.343
Gender: Male	50.426	44.781	47.657
Education: Primary	9.440	3.827	13.257
Education: Secondary	49.330	72.465	46.400
Education: Post-secondary, tertiary and higher	41.230	23.410	39.733
Region: Rural	61.084	70.875	68.381
Region: Urban	38.916	29.125	31.619
N	1,642	2,012	2,625
N of interviewers	103	147	116
N of regions	5	20	8



S5 Multilevel regression results

Table S9: Multilevel regression results. Dependent variable: Response differentiation.

	(1) Refugees	(2) ESS, DK	(3) ESS, HU	(4) ESS, IE
Location equation				
Intercept	-0.092	0.045	0.091	0.088
	[-0.155, -0.029]	[-0.036, 0.125]	[0.013, 0.170]	[0.009, 0.168]
log(Interview	0.025	-0.020	-0.04	-0.034
sequence)	[0.002, 0.048]	[-0.057, 0.017]	[-0.069, -0.012]	[-0.064, -0.005]
Scale equation				
Intercept	-0.846	-0.484	-0.853	-0.771
	[-0.894, -0.798]	[-0.523, -0.445]	[-0.925, -0.780]	[-0.824, -0.718]
sd(Location intercept)	0.220	0.090	0.378	0.325
	[0.171, 0.275]	[0.017, 0.145]	[0.327, 0.436]	[0.272, 0.387]
sd(Slope intercept)	0.071	0.014	0.071	0.100
	[0.043, 0.100]	[0.001, 0.038]	[0.038, 0.104]	[0.074, 0.127]
sd(Scale intercept)	0.191	0.064	0.358	0.237
	[0.154, 0.235]	[0.005, 0.128]	[0.300, 0.423]	[0.192, 0.287]
N of interviewers	98	103	147	116
N of observations	4,810	1,642	2,012	2,625

Notes: 95 percent credible intervals in brackets. Observations with missings in response differentiation were excluded from the analysis. For the IAB-BAMF-SOEP Survey of Refugees one further observation with missing age was excluded from the analysis. For the ESS samples, if less than five observations had missings in a control variable, these observations were excluded from the analysis.

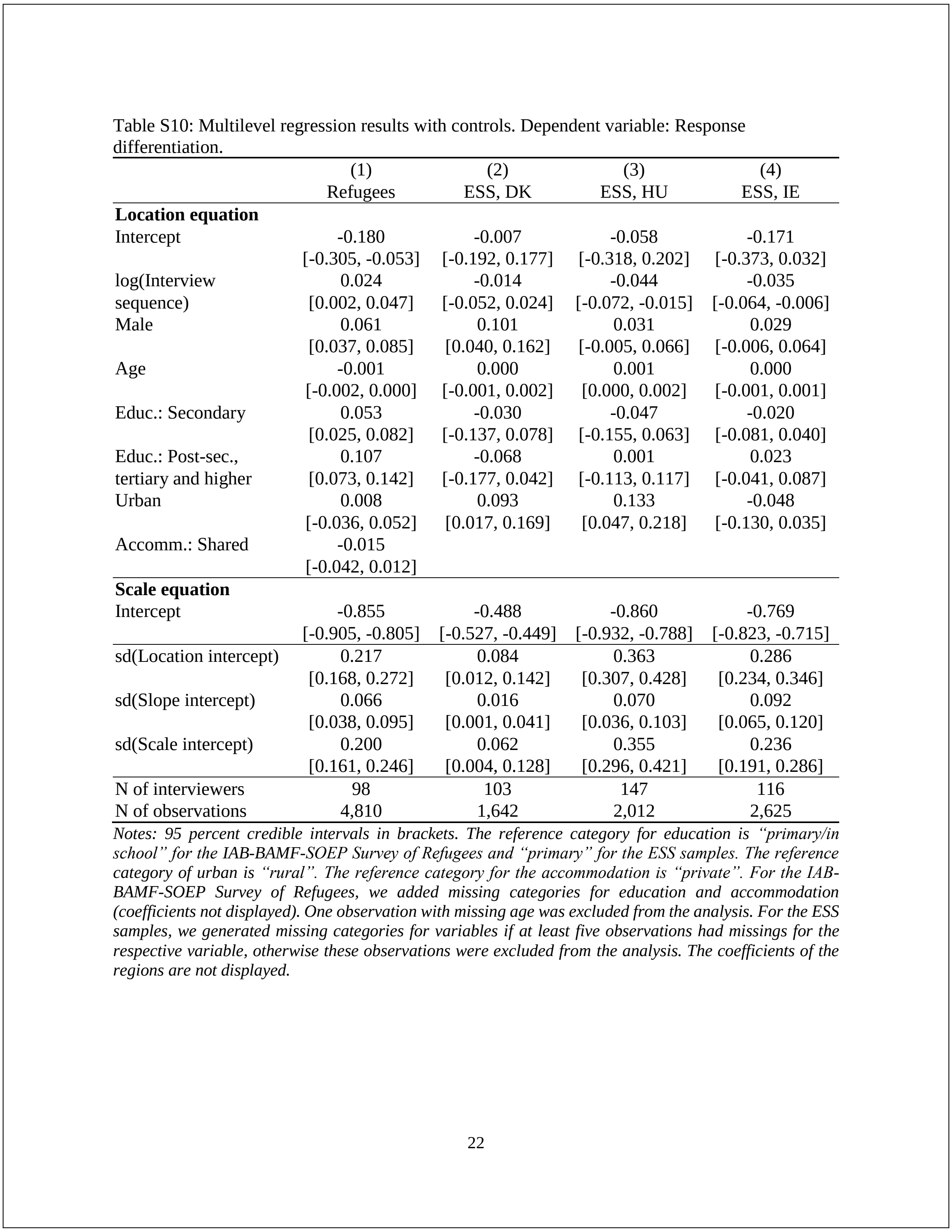


Table S10: Multilevel regression results with controls. Dependent variable: Response differentiation.

	(1) Refugees	(2) ESS, DK	(3) ESS, HU	(4) ESS, IE
Location equation				
Intercept	-0.180	-0.007	-0.058	-0.171
	[-0.305, -0.053]	[-0.192, 0.177]	[-0.318, 0.202]	[-0.373, 0.032]
log(Interview	0.024	-0.014	-0.044	-0.035
sequence)	[0.002, 0.047]	[-0.052, 0.024]	[-0.072, -0.015]	[-0.064, -0.006]
Male	0.061	0.101	0.031	0.029
	[0.037, 0.085]	[0.040, 0.162]	[-0.005, 0.066]	[-0.006, 0.064]
Age	-0.001	0.000	0.001	0.000
	[-0.002, 0.000]	[-0.001, 0.002]	[0.000, 0.002]	[-0.001, 0.001]
Educ.: Secondary	0.053	-0.030	-0.047	-0.020
	[0.025, 0.082]	[-0.137, 0.078]	[-0.155, 0.063]	[-0.081, 0.040]
Educ.: Post-sec.,	0.107	-0.068	0.001	0.023
tertiary and higher	[0.073, 0.142]	[-0.177, 0.042]	[-0.113, 0.117]	[-0.041, 0.087]
Urban	0.008	0.093	0.133	-0.048
	[-0.036, 0.052]	[0.017, 0.169]	[0.047, 0.218]	[-0.130, 0.035]
Accomm.: Shared	-0.015			
	[-0.042, 0.012]			
Scale equation				
Intercept	-0.855	-0.488	-0.860	-0.769
	[-0.905, -0.805]	[-0.527, -0.449]	[-0.932, -0.788]	[-0.823, -0.715]
sd(Location intercept)	0.217	0.084	0.363	0.286
	[0.168, 0.272]	[0.012, 0.142]	[0.307, 0.428]	[0.234, 0.346]
sd(Slope intercept)	0.066	0.016	0.070	0.092
	[0.038, 0.095]	[0.001, 0.041]	[0.036, 0.103]	[0.065, 0.120]
sd(Scale intercept)	0.200	0.062	0.355	0.236
	[0.161, 0.246]	[0.004, 0.128]	[0.296, 0.421]	[0.191, 0.286]
N of interviewers	98	103	147	116
N of observations	4,810	1,642	2,012	2,625

Notes: 95 percent credible intervals in brackets. The reference category for education is "primary/in school" for the IAB-BAMF-SOEP Survey of Refugees and "primary" for the ESS samples. The reference category of urban is "rural". The reference category for the accommodation is "private". For the IAB-BAMF-SOEP Survey of Refugees, we added missing categories for education and accommodation (coefficients not displayed). One observation with missing age was excluded from the analysis. For the ESS samples, we generated missing categories for variables if at least five observations had missings for the respective variable, otherwise these observations were excluded from the analysis. The coefficients of the regions are not displayed.

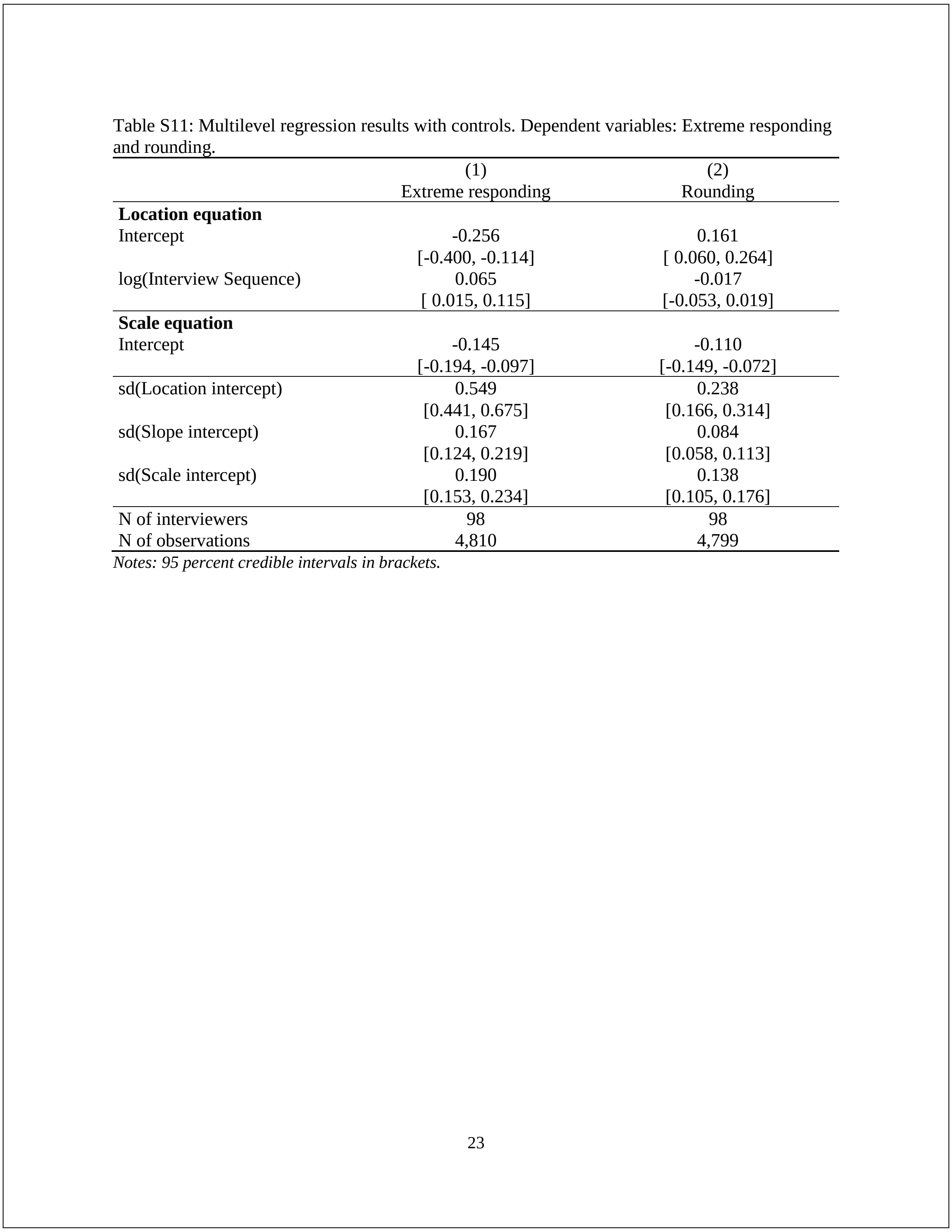


Table S11: Multilevel regression results with controls. Dependent variables: Extreme responding and rounding.

	(1) Extreme responding	(2) Rounding
Location equation		
Intercept	-0.256	0.161
	[-0.400, -0.114]	[0.060, 0.264]
log(Interview Sequence)	0.065	-0.017
	[0.015, 0.115]	[-0.053, 0.019]
Scale equation		
Intercept	-0.145	-0.110
	[-0.194, -0.097]	[-0.149, -0.072]
sd(Location intercept)	0.549	0.238
	[0.441, 0.675]	[0.166, 0.314]
sd(Slope intercept)	0.167	0.084
	[0.124, 0.219]	[0.058, 0.113]
sd(Scale intercept)	0.190	0.138
	[0.153, 0.234]	[0.105, 0.176]
N of interviewers	98	98
N of observations	4,810	4,799

Notes: 95 percent credible intervals in brackets.

S6 Interviewer effects for models with covariates

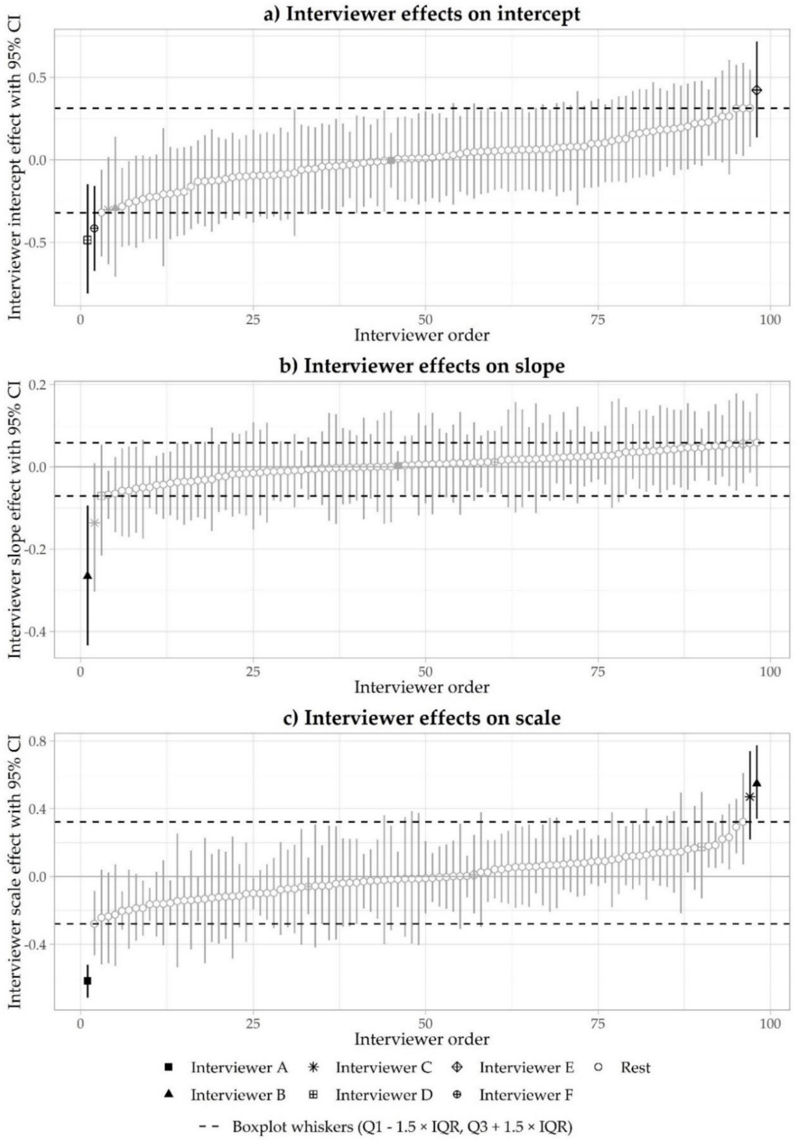


Figure S5: Interviewer effects on intercept, slope, and scale. Bolded interviewers have median posteriors exceeding the boxplot whiskers and credible intervals not including zero. Predictions are based on Model 1 in Table S10. IAB-BAMF-SOEP Survey of Refugees 2016.

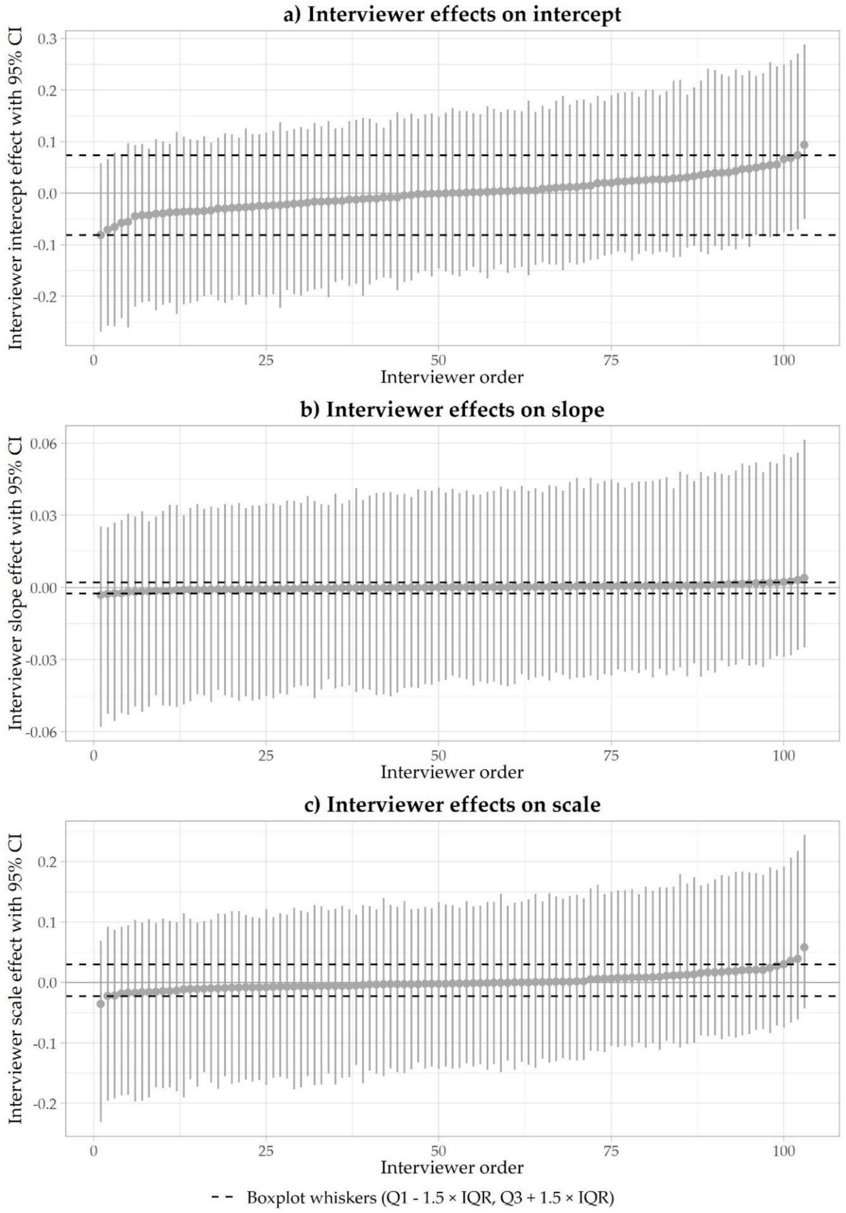


Figure S 6: Interviewer effects on intercept, slope, and scale. Predictions are based on Model 2 in Table S10. ESS Round 6, Denmark.

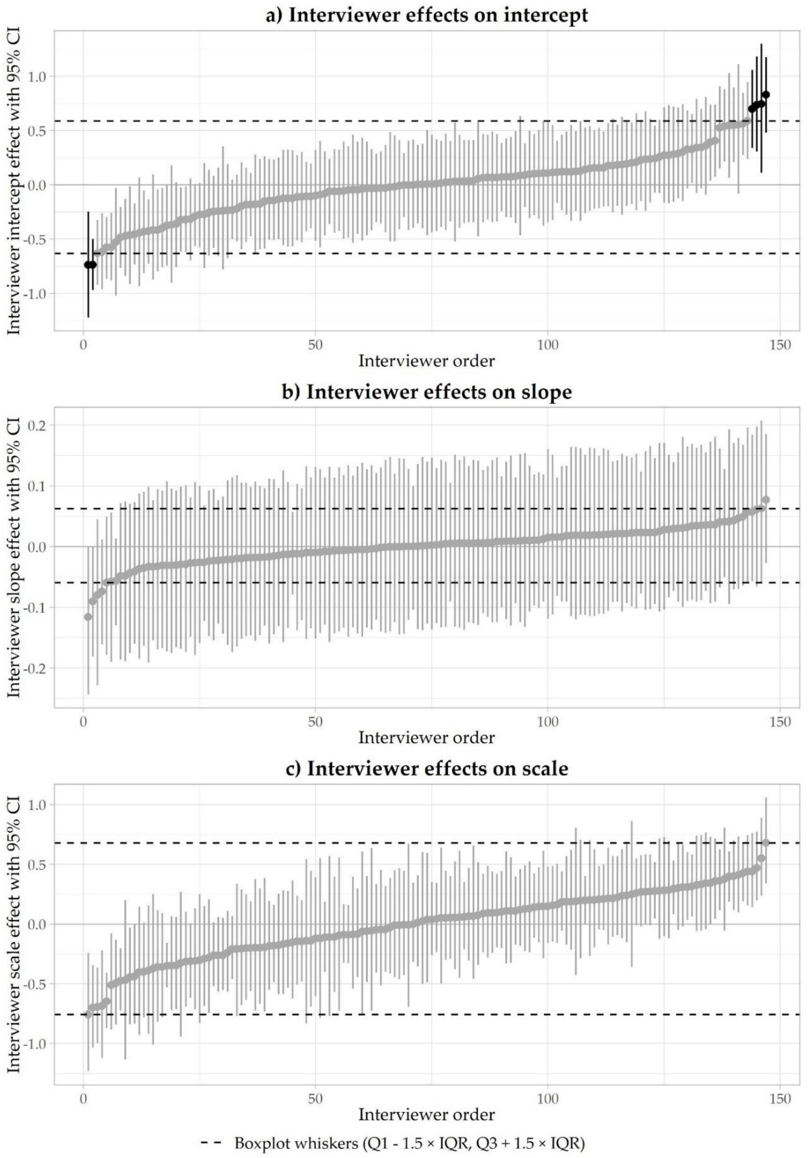


Figure S7: Interviewer effects on intercept, slope, and scale. Bolded interviewers have median posteriors exceeding the boxplot whiskers and credible intervals not including zero. Predictions are based on Model 3 in Table S10. ESS Round 6, Hungary.

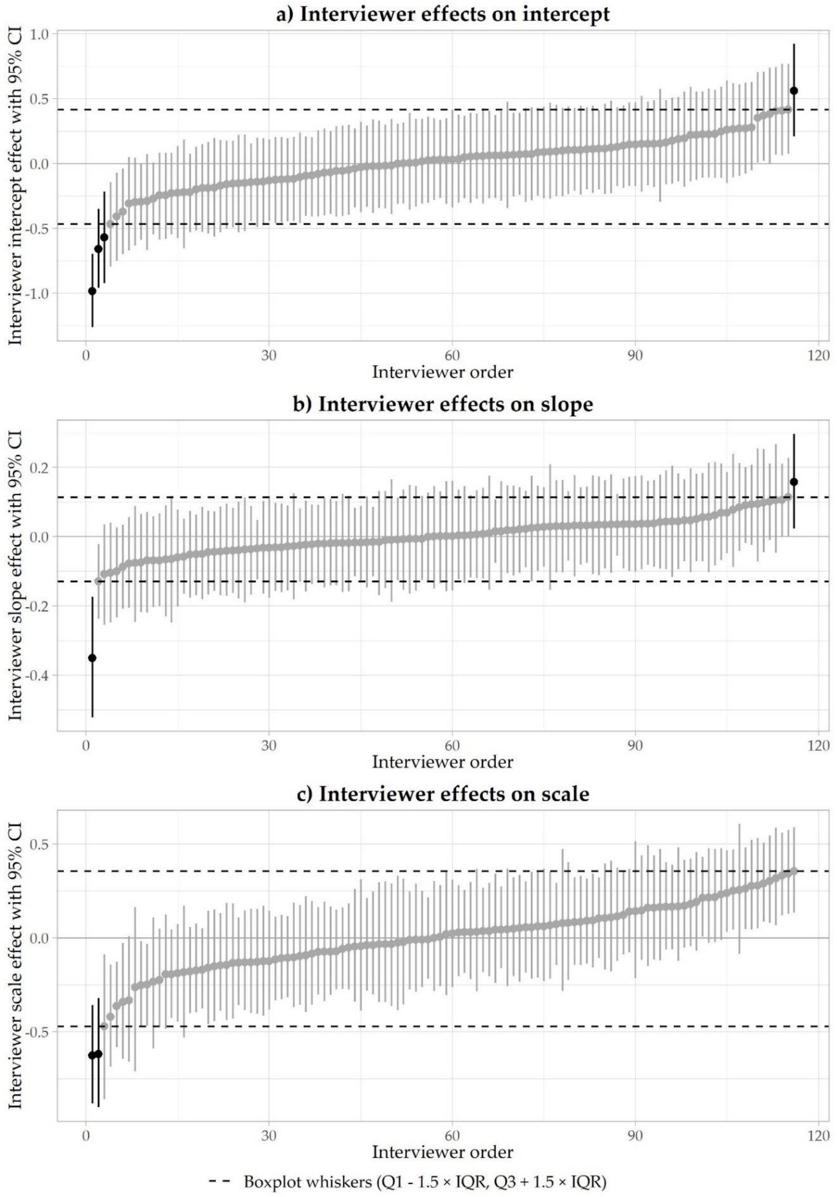


Figure S8: Interviewer effects on intercept, slope, and scale. Bolded interviewers have median posteriors exceeding the boxplot whiskers and credible intervals not including zero. Predictions are based on Model 4 in Table S10. ESS Round 6, Ireland.

27

S7 Interviewer effects for extreme responding and rounding

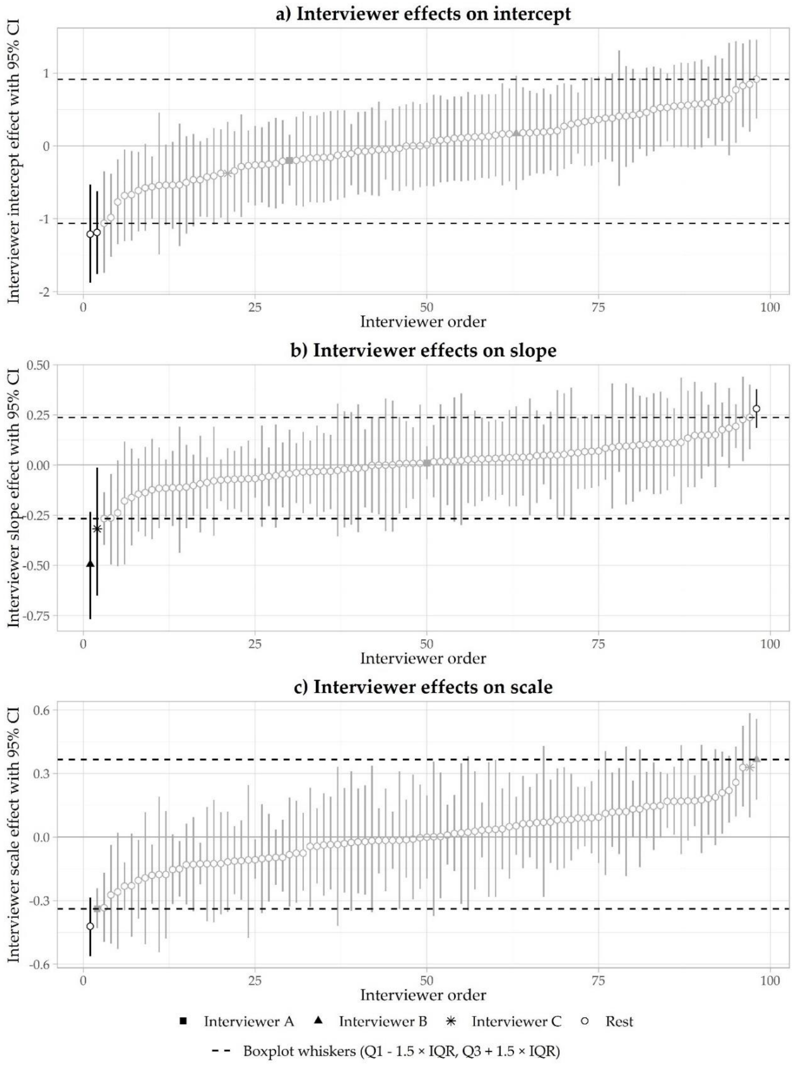


Figure S9: Interviewer effects on intercept, slope, and scale for extreme responding. Bolded interviewers have median posteriors exceeding the boxplot whiskers and credible intervals not including zero. Predictions are based on Model 1 in Table S11. IAB-BAMF-SOEP Survey of Refugees 2016.

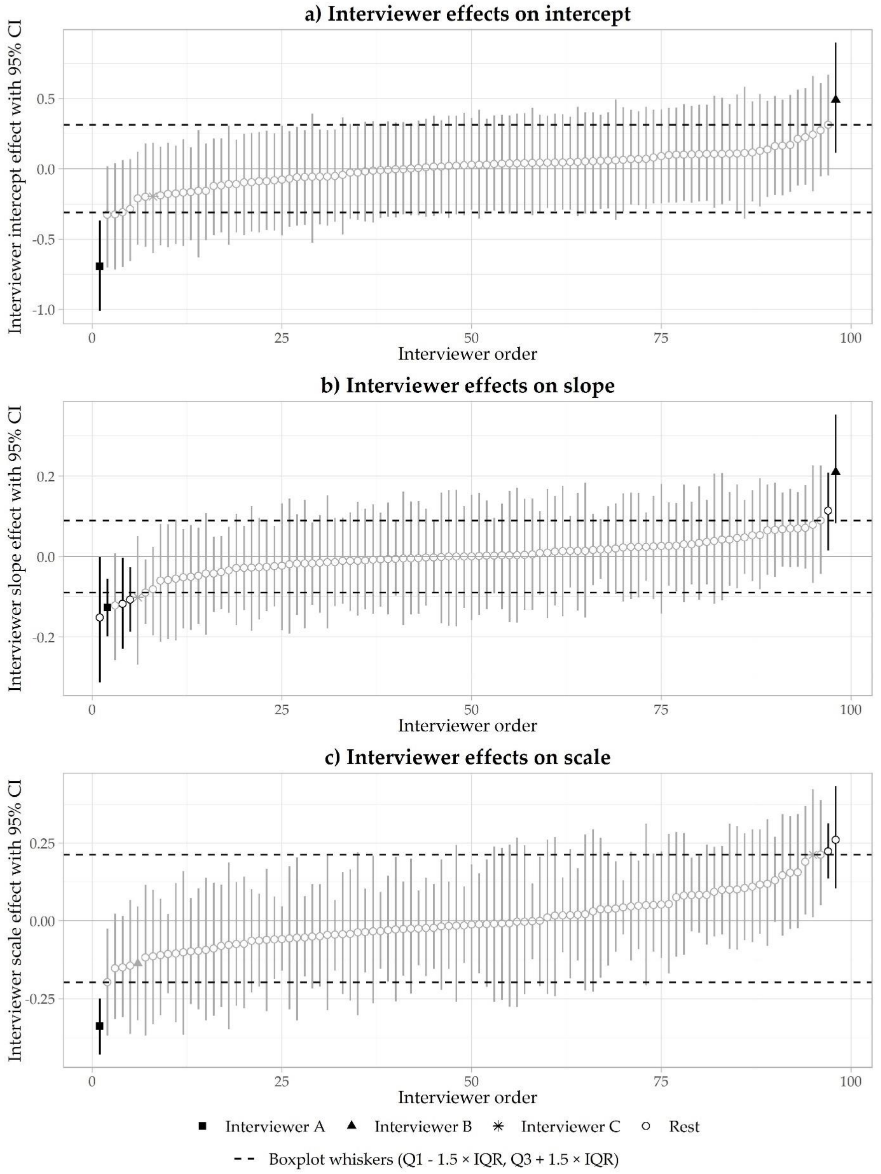


Figure S10: Interviewer effects on intercept, slope, and scale for rounding. Bolded interviewers have median posteriors exceeding the boxplot whiskers and credible intervals not including zero. Predictions are based on Model 2 in Table S11. IAB-BAMF-SOEP Survey of Refugees 2016.

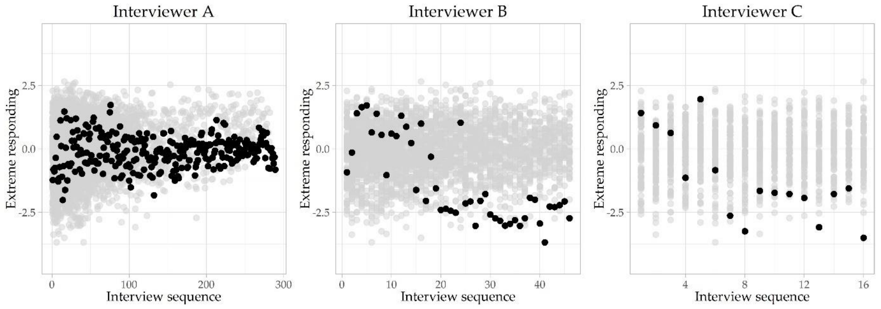


Figure S11: Development of extreme responding for verified falsifiers. Black dots correspond to the respective falsifier, grey dots correspond to the rest of the sample. IAB-BAMF-SOEP Survey of Refugees 2016.

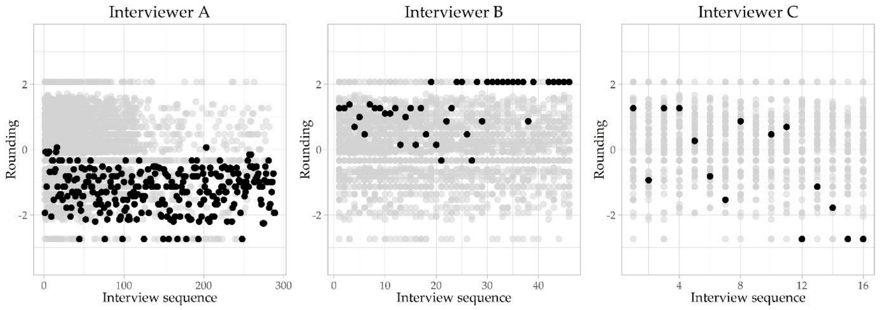


Figure S12: Development of rounding for verified falsifiers. Black dots correspond to the respective falsifier, grey dots correspond to the rest of the sample. IAB-BAMF-SOEP Survey of Refugees 2016.

S8 European Social Survey, results for Denmark and Hungary

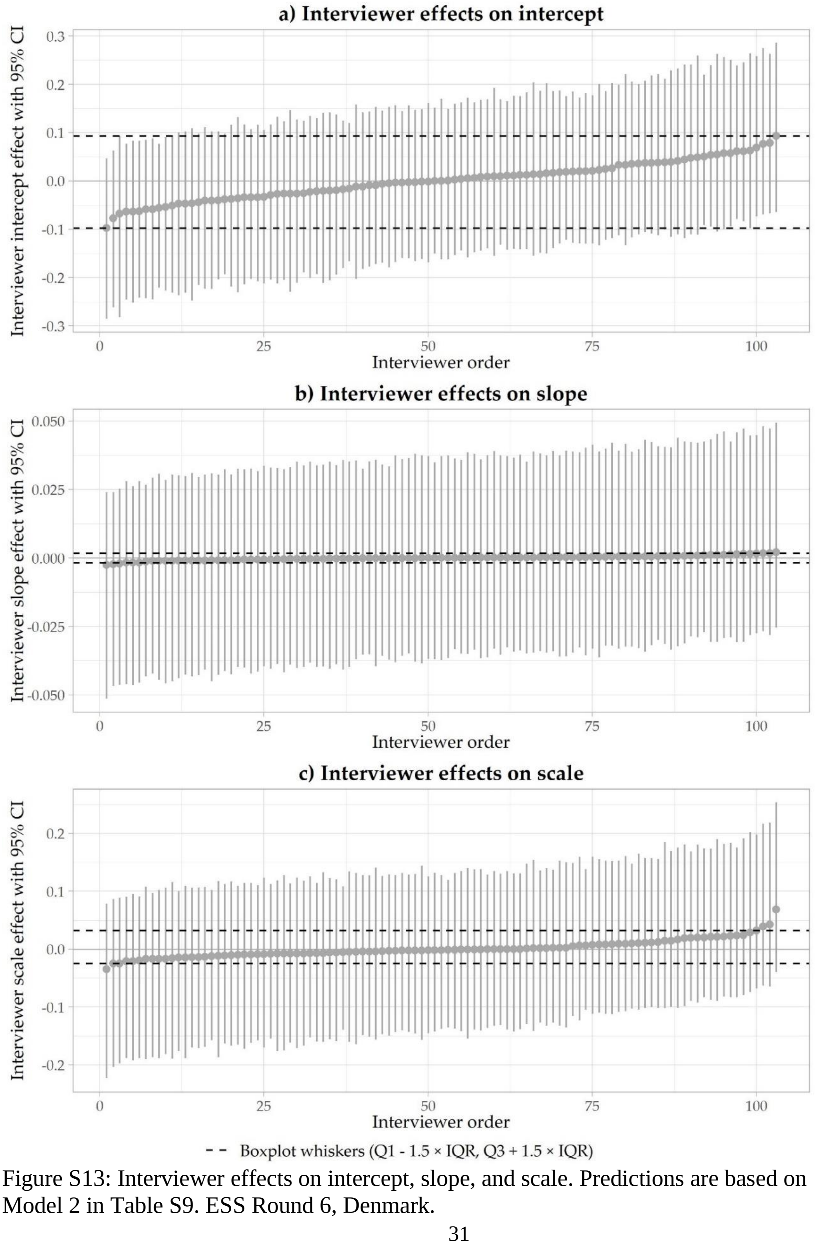


Figure S13: Interviewer effects on intercept, slope, and scale. Predictions are based on Model 2 in Table S9. ESS Round 6, Denmark.

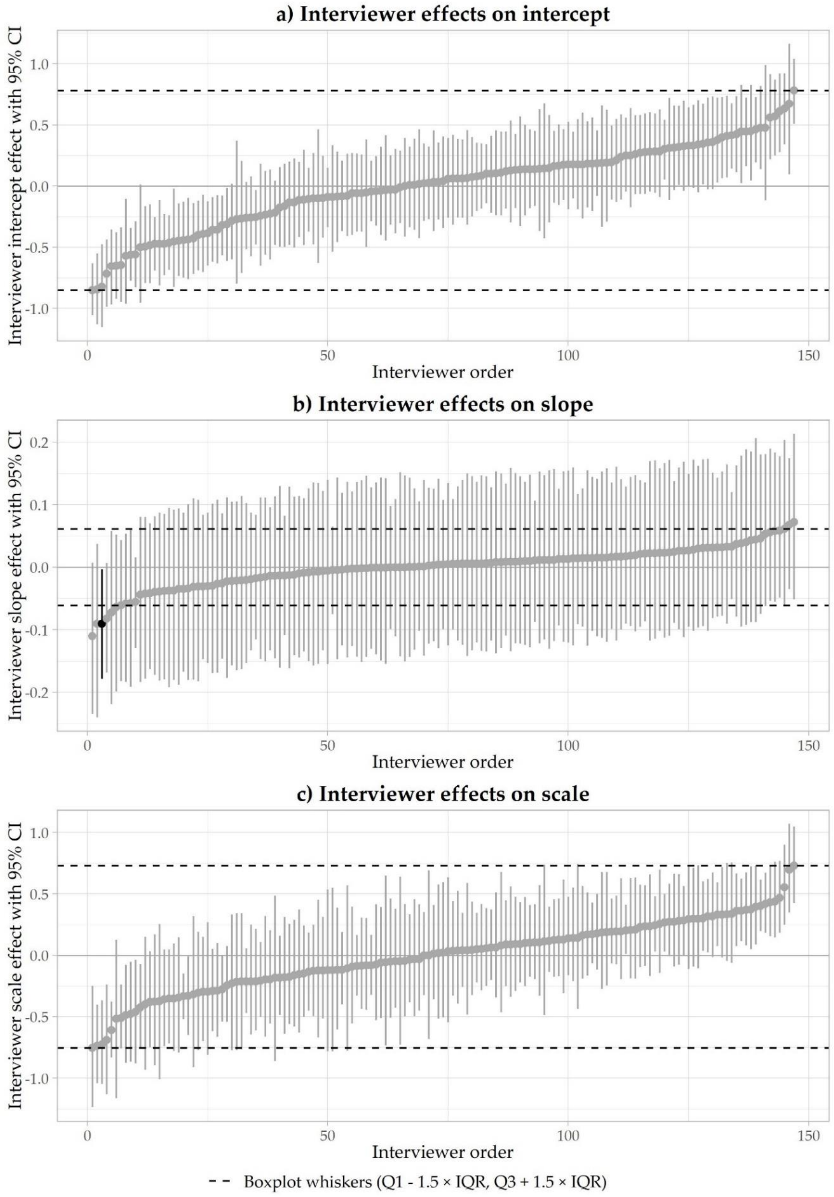


Figure S14: Interviewer effects on intercept, slope, and scale. Bolded interviewers have median posteriors exceeding the boxplot whiskers and credible intervals not including zero. Predictions are based on Model 3 in Table S9. ESS Round 6, Hungary.

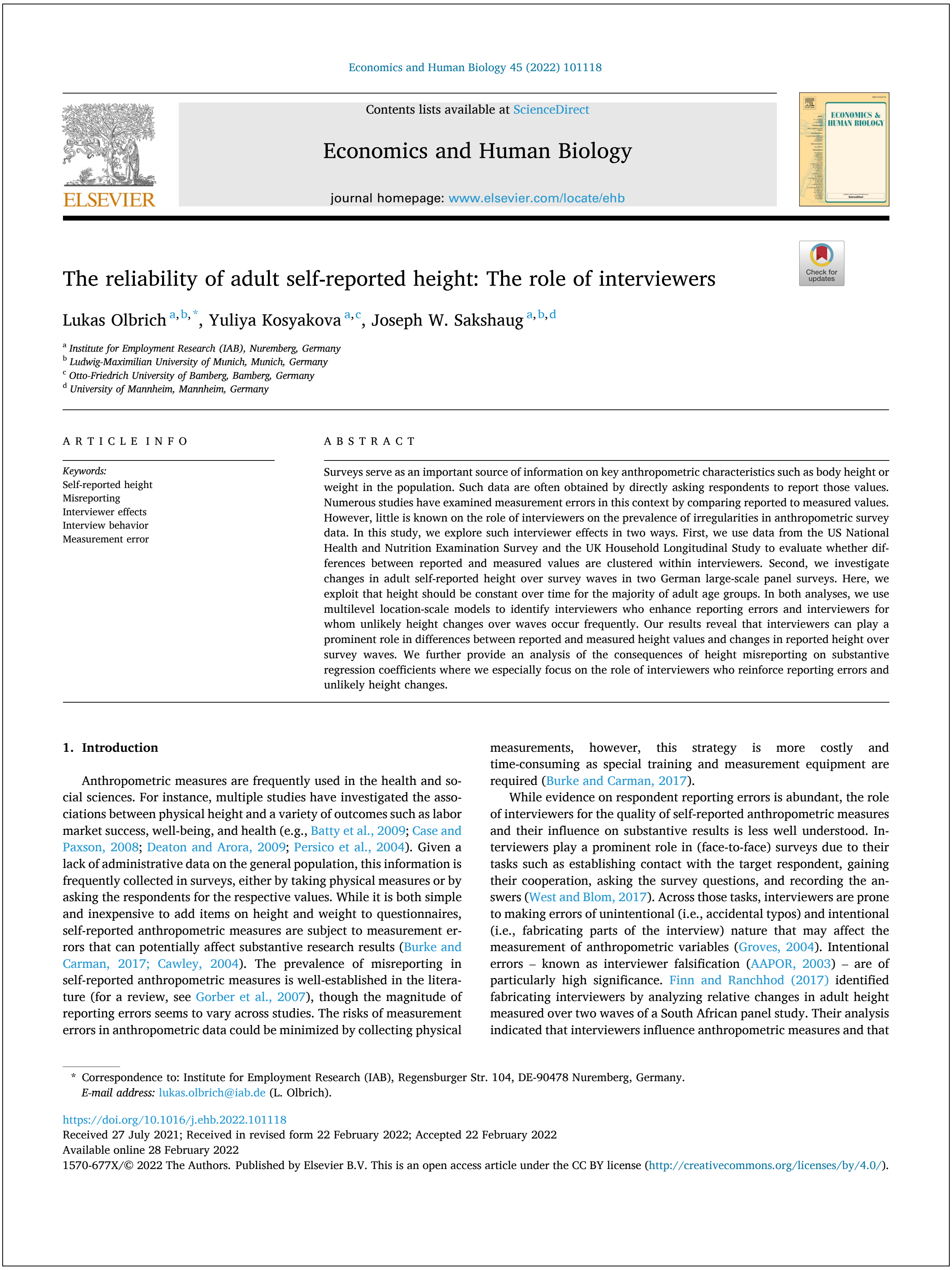
5 The reliability of adult self-reported height: The role of interviewers

Declaration of Contributions

Lukas Olbrich has completed the data analyses, drafted the paper, and revised the paper based upon Yuliya Kosyakova's and Joseph W. Sakshaug's comments.

Contributions of Co-authors

Yuliya Kosyakova has supervised the process of data analyses, reviewed, edited, and commented on various versions of the paper. Joseph W. Sakshaug has supervised the process of data analyses, reviewed, edited, and commented on various versions of the paper.



Economics and Human Biology 45 (2022) 101118

Contents lists available at ScienceDirect

Economics and Human Biology

journal homepage: www.elsevier.com/locate/ehb

The reliability of adult self-reported height: The role of interviewers

Lukas Olbrich [a,b,*], Yuliya Kosyakova [a,c], Joseph W. Sakshaug [a,b,d]

[a] *Institute for Employment Research (IAB), Nuremberg, Germany*
[b] *Ludwig-Maximilian University of Munich, Munich, Germany*
[c] *Otto-Friedrich University of Bamberg, Bamberg, Germany*
[d] *University of Mannheim, Mannheim, Germany*

ARTICLE INFO

Keywords:
Self-reported height
Misreporting
Interviewer effects
Interview behavior
Measurement error

ABSTRACT

Surveys serve as an important source of information on key anthropometric characteristics such as body height or weight in the population. Such data are often obtained by directly asking respondents to report those values. Numerous studies have examined measurement errors in this context by comparing reported to measured values. However, little is known on the role of interviewers on the prevalence of irregularities in anthropometric survey data. In this study, we explore such interviewer effects in two ways. First, we use data from the US National Health and Nutrition Examination Survey and the UK Household Longitudinal Study to evaluate whether differences between reported and measured values are clustered within interviewers. Second, we investigate changes in adult self-reported height over survey waves in two German large-scale panel surveys. Here, we exploit that height should be constant over time for the majority of adult age groups. In both analyses, we use multilevel location-scale models to identify interviewers who enhance reporting errors and interviewers for whom unlikely height changes over waves occur frequently. Our results reveal that interviewers can play a prominent role in differences between reported and measured height values and changes in reported height over survey waves. We further provide an analysis of the consequences of height misreporting on substantive regression coefficients where we especially focus on the role of interviewers who reinforce reporting errors and unlikely height changes.

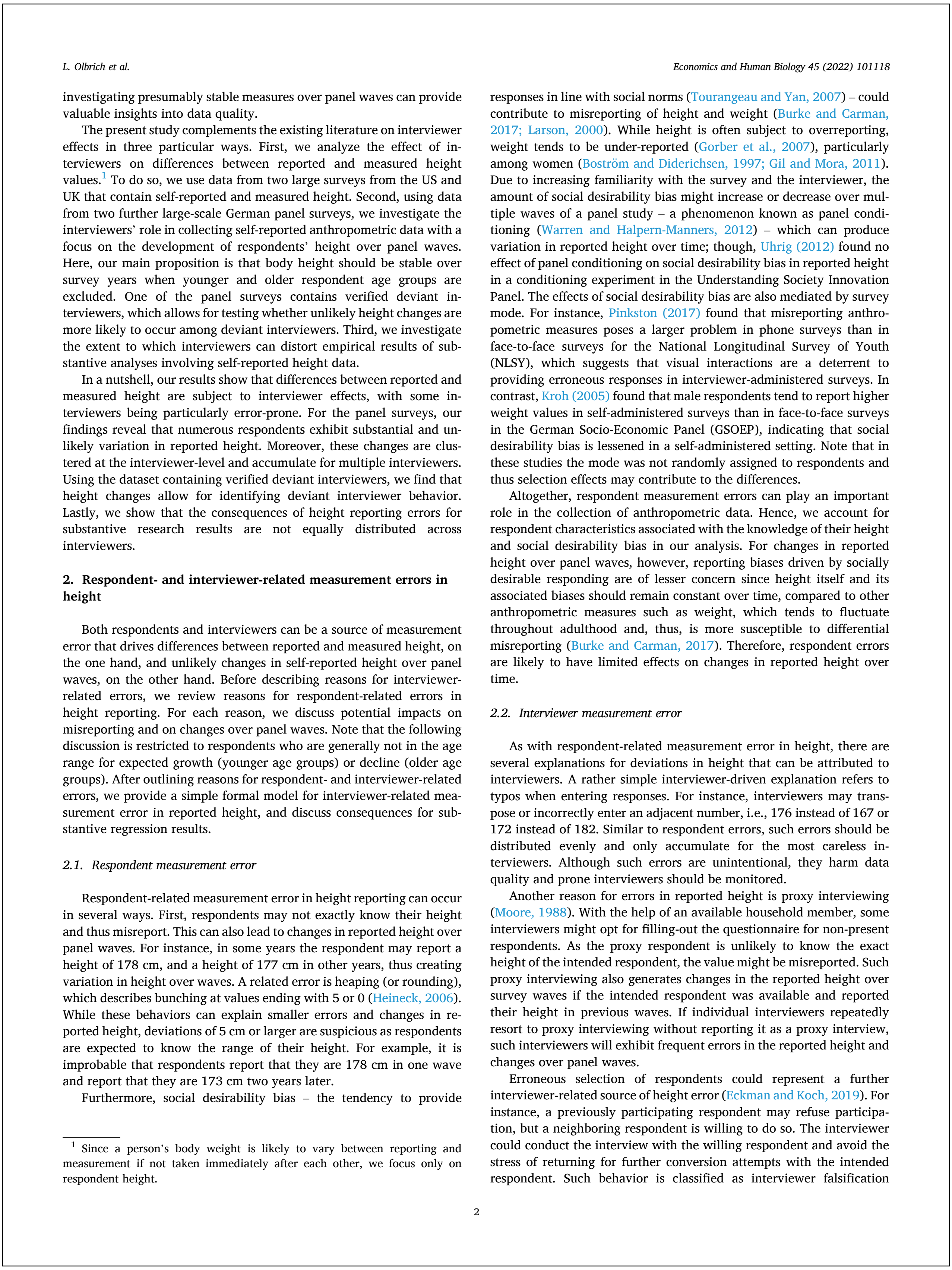
1. Introduction

Anthropometric measures are frequently used in the health and social sciences. For instance, multiple studies have investigated the associations between physical height and a variety of outcomes such as labor market success, well-being, and health (e.g., Batty et al., 2009; Case and Paxson, 2008; Deaton and Arora, 2009; Persico et al., 2004). Given a lack of administrative data on the general population, this information is frequently collected in surveys, either by taking physical measures or by asking the respondents for the respective values. While it is both simple and inexpensive to add items on height and weight to questionnaires, self-reported anthropometric measures are subject to measurement errors that can potentially affect substantive research results (Burke and Carman, 2017; Cawley, 2004). The prevalence of misreporting in self-reported anthropometric measures is well-established in the literature (for a review, see Gorber et al., 2007), though the magnitude of reporting errors seems to vary across studies. The risks of measurement errors in anthropometric data could be minimized by collecting physical measurements, however, this strategy is more costly and time-consuming as special training and measurement equipment are required (Burke and Carman, 2017).

While evidence on respondent reporting errors is abundant, the role of interviewers for the quality of self-reported anthropometric measures and their influence on substantive results is less well understood. Interviewers play a prominent role in (face-to-face) surveys due to their tasks such as establishing contact with the target respondent, gaining their cooperation, asking the survey questions, and recording the answers (West and Blom, 2017). Across those tasks, interviewers are prone to making errors of unintentional (i.e., accidental typos) and intentional (i.e., fabricating parts of the interview) nature that may affect the measurement of anthropometric variables (Groves, 2004). Intentional errors – known as interviewer falsification (AAPOR, 2003) – are of particularly high significance. Finn and Ranchhod (2017) identified fabricating interviewers by analyzing relative changes in adult height measured over two waves of a South African panel study. Their analysis indicated that interviewers influence anthropometric measures and that

[*] Correspondence to: Institute for Employment Research (IAB), Regensburger Str. 104, DE-90478 Nuremberg, Germany.
E-mail address: lukas.olbrich@iab.de (L. Olbrich).

https://doi.org/10.1016/j.ehb.2022.101118
Received 27 July 2021; Received in revised form 22 February 2022; Accepted 22 February 2022
Available online 28 February 2022



investigating presumably stable measures over panel waves can provide valuable insights into data quality.

The present study complements the existing literature on interviewer effects in three particular ways. First, we analyze the effect of interviewers on differences between reported and measured height values.[1] To do so, we use data from two large surveys from the US and UK that contain self-reported and measured height. Second, using data from two further large-scale German panel surveys, we investigate the interviewers' role in collecting self-reported anthropometric data with a focus on the development of respondents' height over panel waves. Here, our main proposition is that body height should be stable over survey years when younger and older respondent age groups are excluded. One of the panel surveys contains verified deviant interviewers, which allows for testing whether unlikely height changes are more likely to occur among deviant interviewers. Third, we investigate the extent to which interviewers can distort empirical results of substantive analyses involving self-reported height data.

In a nutshell, our results show that differences between reported and measured height are subject to interviewer effects, with some interviewers being particularly error-prone. For the panel surveys, our findings reveal that numerous respondents exhibit substantial and unlikely variation in reported height. Moreover, these changes are clustered at the interviewer-level and accumulate for multiple interviewers. Using the dataset containing verified deviant interviewers, we find that height changes allow for identifying deviant interviewer behavior. Lastly, we show that the consequences of height reporting errors for substantive research results are not equally distributed across interviewers.

2. Respondent- and interviewer-related measurement errors in height

Both respondents and interviewers can be a source of measurement error that drives differences between reported and measured height, on the one hand, and unlikely changes in self-reported height over panel waves, on the other hand. Before describing reasons for interviewer-related errors, we review reasons for respondent-related errors in height reporting. For each reason, we discuss potential impacts on misreporting and on changes over panel waves. Note that the following discussion is restricted to respondents who are generally not in the age range for expected growth (younger age groups) or decline (older age groups). After outlining reasons for respondent- and interviewer-related errors, we provide a simple formal model for interviewer-related measurement error in reported height, and discuss consequences for substantive regression results.

2.1. Respondent measurement error

Respondent-related measurement error in height reporting can occur in several ways. First, respondents may not exactly know their height and thus misreport. This can also lead to changes in reported height over panel waves. For instance, in some years the respondent may report a height of 178 cm, and a height of 177 cm in other years, thus creating variation in height over waves. A related error is heaping (or rounding), which describes bunching at values ending with 5 or 0 (Heineck, 2006). While these behaviors can explain smaller errors and changes in reported height, deviations of 5 cm or larger are suspicious as respondents are expected to know the range of their height. For example, it is improbable that respondents report that they are 178 cm in one wave and report that they are 173 cm two years later.

Furthermore, social desirability bias – the tendency to provide responses in line with social norms (Tourangeau and Yan, 2007) – could contribute to misreporting of height and weight (Burke and Carman, 2017; Larson, 2000). While height is often subject to overreporting, weight tends to be under-reported (Gorber et al., 2007), particularly among women (Boström and Diderichsen, 1997; Gil and Mora, 2011). Due to increasing familiarity with the survey and the interviewer, the amount of social desirability bias might increase or decrease over multiple waves of a panel study – a phenomenon known as panel conditioning (Warren and Halpern-Manners, 2012) – which can produce variation in reported height over time; though, Uhrig (2012) found no effect of panel conditioning on social desirability bias in reported height in a conditioning experiment in the Understanding Society Innovation Panel. The effects of social desirability bias are also mediated by survey mode. For instance, Pinkston (2017) found that misreporting anthropometric measures poses a larger problem in phone surveys than in face-to-face surveys for the National Longitudinal Survey of Youth (NLSY), which suggests that visual interactions are a deterrent to providing erroneous responses in interviewer-administered surveys. In contrast, Kroh (2005) found that male respondents tend to report higher weight values in self-administered surveys than in face-to-face surveys in the German Socio-Economic Panel (GSOEP), indicating that social desirability bias is lessened in a self-administered setting. Note that in these studies the mode was not randomly assigned to respondents and thus selection effects may contribute to the differences.

Altogether, respondent measurement errors can play an important role in the collection of anthropometric data. Hence, we account for respondent characteristics associated with the knowledge of their height and social desirability bias in our analysis. For changes in reported height over panel waves, however, reporting biases driven by socially desirable responding are of lesser concern since height itself and its associated biases should remain constant over time, compared to other anthropometric measures such as weight, which tends to fluctuate throughout adulthood and, thus, is more susceptible to differential misreporting (Burke and Carman, 2017). Therefore, respondent errors are likely to have limited effects on changes in reported height over time.

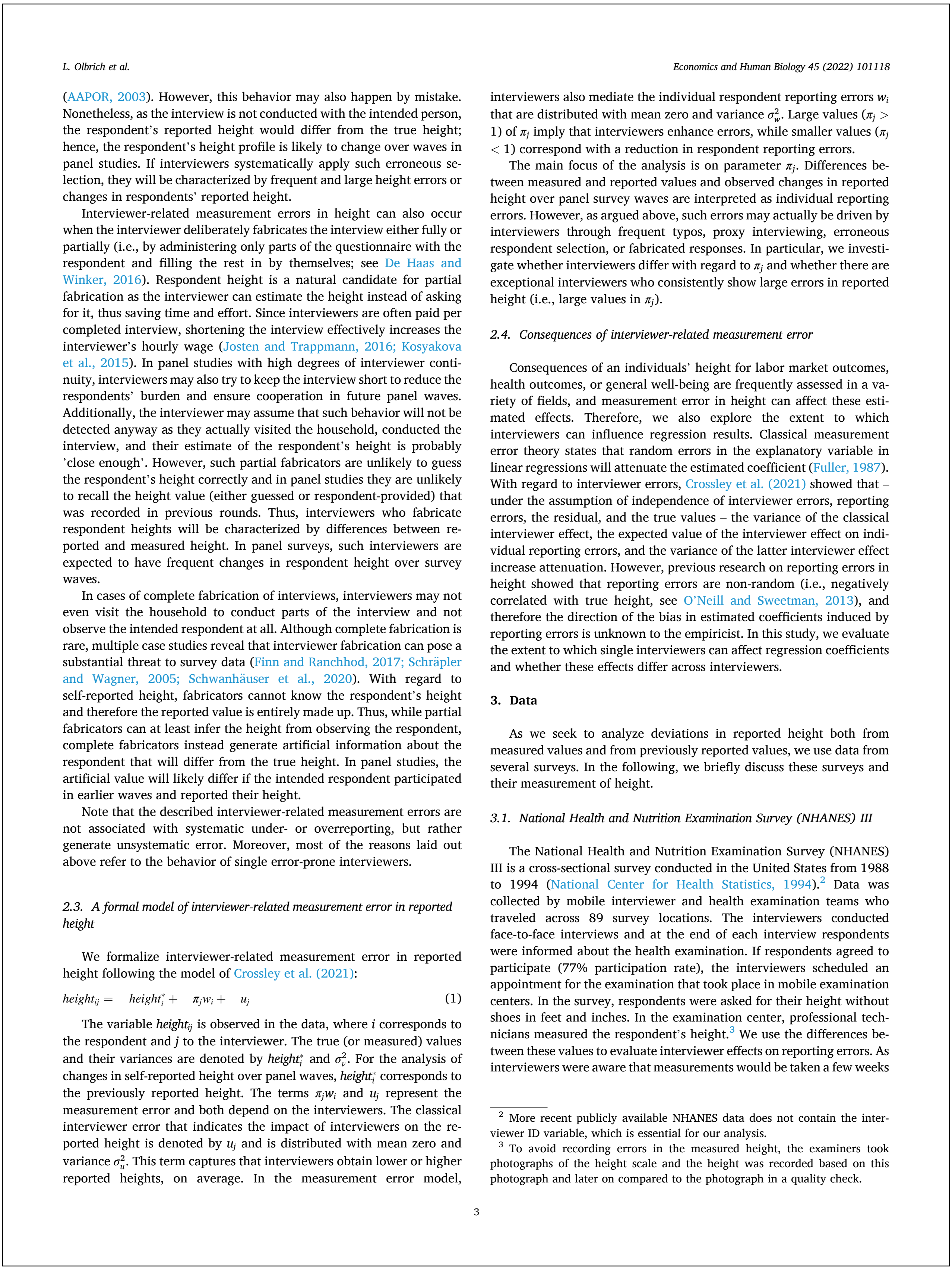
2.2. Interviewer measurement error

As with respondent-related measurement error in height, there are several explanations for deviations in height that can be attributed to interviewers. A rather simple interviewer-driven explanation refers to typos when entering responses. For instance, interviewers may transpose or incorrectly enter an adjacent number, i.e., 176 instead of 167 or 172 instead of 182. Similar to respondent errors, such errors should be distributed evenly and only accumulate for the most careless interviewers. Although such errors are unintentional, they harm data quality and prone interviewers should be monitored.

Another reason for errors in reported height is proxy interviewing (Moore, 1988). With the help of an available household member, some interviewers might opt for filling-out the questionnaire for non-present respondents. As the proxy respondent is unlikely to know the exact height of the intended respondent, the value might be misreported. Such proxy interviewing also generates changes in the reported height over survey waves if the intended respondent was available and reported their height in previous waves. If individual interviewers repeatedly resort to proxy interviewing without reporting it as a proxy interview, such interviewers will exhibit frequent errors in the reported height and changes over panel waves.

Erroneous selection of respondents could represent a further interviewer-related source of height error (Eckman and Koch, 2019). For instance, a previously participating respondent may refuse participation, but a neighboring respondent is willing to do so. The interviewer could conduct the interview with the willing respondent and avoid the stress of returning for further conversion attempts with the intended respondent. Such behavior is classified as interviewer falsification

[1] Since a person's body weight is likely to vary between reporting and measurement if not taken immediately after each other, we focus only on respondent height.



(AAPOR, 2003). However, this behavior may also happen by mistake. Nonetheless, as the interview is not conducted with the intended person, the respondent's reported height would differ from the true height; hence, the respondent's height profile is likely to change over waves in panel studies. If interviewers systematically apply such erroneous selection, they will be characterized by frequent and large height errors or changes in respondents' reported height.

Interviewer-related measurement errors in height can also occur when the interviewer deliberately fabricates the interview either fully or partially (i.e., by administering only parts of the questionnaire with the respondent and filling the rest in by themselves; see De Haas and Winker, 2016). Respondent height is a natural candidate for partial fabrication as the interviewer can estimate the height instead of asking for it, thus saving time and effort. Since interviewers are often paid per completed interview, shortening the interview effectively increases the interviewer's hourly wage (Josten and Trappmann, 2016; Kosyakova et al., 2015). In panel studies with high degrees of interviewer continuity, interviewers may also try to keep the interview short to reduce the respondents' burden and ensure cooperation in future panel waves. Additionally, the interviewer may assume that such behavior will not be detected anyway as they actually visited the household, conducted the interview, and their estimate of the respondent's height is probably 'close enough'. However, such partial fabricators are unlikely to guess the respondent's height correctly and in panel studies they are unlikely to recall the height value (either guessed or respondent-provided) that was recorded in previous rounds. Thus, interviewers who fabricate respondent heights will be characterized by differences between reported and measured height. In panel surveys, such interviewers are expected to have frequent changes in respondent height over survey waves.

In cases of complete fabrication of interviews, interviewers may not even visit the household to conduct parts of the interview and not observe the intended respondent at all. Although complete fabrication is rare, multiple case studies reveal that interviewer fabrication can pose a substantial threat to survey data (Finn and Ranchhod, 2017; Schräpler and Wagner, 2005; Schwanhäuser et al., 2020). With regard to self-reported height, fabricators cannot know the respondent's height and therefore the reported value is entirely made up. Thus, while partial fabricators can at least infer the height from observing the respondent, complete fabricators instead generate artificial information about the respondent that will differ from the true height. In panel studies, the artificial value will likely differ if the intended respondent participated in earlier waves and reported their height.

Note that the described interviewer-related measurement errors are not associated with systematic under- or overreporting, but rather generate unsystematic error. Moreover, most of the reasons laid out above refer to the behavior of single error-prone interviewers.

2.3. A formal model of interviewer-related measurement error in reported height

We formalize interviewer-related measurement error in reported height following the model of Crossley et al. (2021):

$$height_{ij} = \quad height_i^* + \quad \pi_j w_i + \quad u_j \tag{1}$$

The variable $height_{ij}$ is observed in the data, where i corresponds to the respondent and j to the interviewer. The true (or measured) values and their variances are denoted by $height_i^*$ and σ_v^2. For the analysis of changes in self-reported height over panel waves, $height_i^*$ corresponds to the previously reported height. The terms $\pi_j w_i$ and u_j represent the measurement error and both depend on the interviewers. The classical interviewer error that indicates the impact of interviewers on the reported height is denoted by u_j and is distributed with mean zero and variance σ_u^2. This term captures that interviewers obtain lower or higher reported heights, on average. In the measurement error model, interviewers also mediate the individual respondent reporting errors w_i that are distributed with mean zero and variance σ_w^2. Large values ($\pi_j >$ 1) of π_j imply that interviewers enhance errors, while smaller values (π_j $<$ 1) correspond with a reduction in respondent reporting errors.

The main focus of the analysis is on parameter π_j. Differences between measured and reported values and observed changes in reported height over panel survey waves are interpreted as individual reporting errors. However, as argued above, such errors may actually be driven by interviewers through frequent typos, proxy interviewing, erroneous respondent selection, or fabricated responses. In particular, we investigate whether interviewers differ with regard to π_j and whether there are exceptional interviewers who consistently show large errors in reported height (i.e., large values in π_j).

2.4. Consequences of interviewer-related measurement error

Consequences of an individuals' height for labor market outcomes, health outcomes, or general well-being are frequently assessed in a variety of fields, and measurement error in height can affect these estimated effects. Therefore, we also explore the extent to which interviewers can influence regression results. Classical measurement error theory states that random errors in the explanatory variable in linear regressions will attenuate the estimated coefficient (Fuller, 1987). With regard to interviewer errors, Crossley et al. (2021) showed that – under the assumption of independence of interviewer errors, reporting errors, the residual, and the true values – the variance of the classical interviewer effect, the expected value of the interviewer effect on individual reporting errors, and the variance of the latter interviewer effect increase attenuation. However, previous research on reporting errors in height showed that reporting errors are non-random (i.e., negatively correlated with true height, see O'Neill and Sweetman, 2013), and therefore the direction of the bias in estimated coefficients induced by reporting errors is unknown to the empiricist. In this study, we evaluate the extent to which single interviewers can affect regression coefficients and whether these effects differ across interviewers.

3. Data

As we seek to analyze deviations in reported height both from measured values and from previously reported values, we use data from several surveys. In the following, we briefly discuss these surveys and their measurement of height.

3.1. National Health and Nutrition Examination Survey (NHANES) III

The National Health and Nutrition Examination Survey (NHANES) III is a cross-sectional survey conducted in the United States from 1988 to 1994 (National Center for Health Statistics, 1994).[2] Data was collected by mobile interviewer and health examination teams who traveled across 89 survey locations. The interviewers conducted face-to-face interviews and at the end of each interview respondents were informed about the health examination. If respondents agreed to participate (77% participation rate), the interviewers scheduled an appointment for the examination that took place in mobile examination centers. In the survey, respondents were asked for their height without shoes in feet and inches. In the examination center, professional technicians measured the respondent's height.[3] We use the differences between these values to evaluate interviewer effects on reporting errors. As interviewers were aware that measurements would be taken a few weeks

[2] More recent publicly available NHANES data does not contain the interviewer ID variable, which is essential for our analysis.

[3] To avoid recording errors in the measured height, the examiners took photographs of the height scale and the height was recorded based on this photograph and later on compared to the photograph in a quality check.

Table 1
Descriptive statistics for each sample.

	N	N of int.	Avg. error (cm)	Avg. abs. error (cm)	% Abs. error > 5 cm
NHANES					
Female	5507	60	0.407	2.168	7.554
Male	4770	60	0.872	2.348	8.386
Total	10,277	60	0.623	2.251	7.940
UKHLS					
Female	3331	208	0.756	1.995	6.244
Male	2326	208	1.789	2.593	11.436
Total	5657	208	1.181	2.241	8.379
GSOEP					
Female	3961	195	-0.037	0.874	3.509
Male	3140	195	-0.071	0.957	3.726
Total	7101	195	-0.052	0.910	3.605
PASS					
Female	1127	82	-0.170	0.815	2.662
Male	920	82	0.023	0.940	3.913
Total	2047	82	-0.084	0.871	3.224

Notes: The average error is the average of the reported height minus the measured height. The average absolute error is the average of the absolute difference between the reported and measured height.

after the interview, deliberate fabrications of height values seem rather unlikely. In addition, the NHANES team recontacted roughly 10% of the respondents for verification to ensure that interviews were adequately conducted, and each questionnaire was checked for "error patterns" (National Center for Health Statistics, 1994, p. 31).

3.2. UK Household Longitudinal Study (UKHLS)

The UK Household Longitudinal Study (UKHLS) is a yearly household panel study that started in 2008 (University of Essex: Institute for Social and Economic Research, 2021). In the second wave, respondents of its predecessor, the British Household Panel Study (BHPS), were integrated in the UKHLS. In the first UKHLS wave, face-to-face interviewers asked respondents for their height and weight. In wave 2, a subsample of the general population sample respondents was selected for separate nurse visits that took place roughly 6 months after the interview (University of Essex: Institute for Social and Economic Research and National Centre for Social Research, 2014).[4] Professional nurses visited the respondents and took a variety of biomeasures such as height and weight, blood pressure, or the collection of blood samples (see Buck and McFall, 2012, for an overview). In this study, the main analysis variables are the self-reported height in wave 1 and the measured height collected after the wave 2 interview. Assuming absence of measurement errors in the nurse measures and absence of height changes between reporting and measurement, we can compare the self-reported height to the true measured height. Furthermore, wave 1 interviewers were likely unaware of the biomeasure collection after the wave 2 interview and thus might have been more careless as verification of collected values was not imminent. Note, however, that the UKHLS interviewers were controlled using a sophisticated monitoring system (Boreham et al., 2012), deeming extreme deviant behavior very unlikely.

3.3. German Socio-Economic Panel (GSOEP)

The German Socio-Economic Panel (GSOEP, DOI:10.5684/soep-core.v35) is a nationally representative, multi-mode household panel survey launched in 1984 (Goebel et al., 2019).[5] Since 2002, the GSOEP collects self-reported height and weight every two years. Our main variable of interest is the reported height. The exact wording of the question is "How tall are you? If you don't know, please estimate."[6] We use data from the most recent pair of height reports in 2016 and 2018. Due to the repeated collection of respondent height, it is possible to examine changes in reported height over several years for individual respondents.

To counteract such deviations, the GSOEP also provides an imputed and edited version of the height variable in the HEALTH data, a dataset that is part of GSOEP-Core and specifically provided to ease the analysis of health modules (SOEP Group, 2020). Missing values are simply replaced by the most recent reported height. Furthermore, "[i]t is assumed that for a two-year-period a change of body height of more than 10 cm is implausible if the values of the other observation years differ only in a range of at most 2 cm. Thus the respective information is imputed by the average of the other values of the respondent" (SOEP Group, 2020, p. 13). However, such editing has only been applied to 43 cases over all survey years (2002–2018) and all age ranges.

3.4. Panel Labour Market and Social Security (PASS)

The last dataset is the Panel Study Labour Market and Social Security (PASS, DOI: 10.5164/IAB.PASS-SUF0619.de.en.v2, an annual survey of households that receive unemployment benefits and households of German residents launched in 2007 (Trappmann et al., 2013, 2019). Interviews are conducted with computer-assisted telephone or computer-assisted personal interviewing mode (CATI and CAPI, respectively). During the field period of wave 15, two interviewers suspicious of deviant interviewing were identified based on paradata and survey data analysis, audio recordings of the interviews, and recontact procedures (Beste et al., 2021). The first interviewer (ID 154) had extremely short interview durations and recontacts indicated that the interviewer completely fabricated some interviews. The second interviewer (ID 109) also had implausible interview durations but recontacts did not yield evidence of complete fabrications. Analyses of previous waves suggested that interviewer 109 acquired habits of speeding through the questionnaire and responding to questions instead of posing them to the respondent. Due to workload limits in the PASS, the consequences for survey results of these deviant interviewers are negligible. However, these cases provide a rare opportunity to assess whether unlikely changes in the reported height over panel waves accumulate for verified deviant interviewers. We use data from waves 9 and 12, in which both deviant interviewers were active and the PASS questionnaire contained an item on the respondent's height.

3.5. Sample description

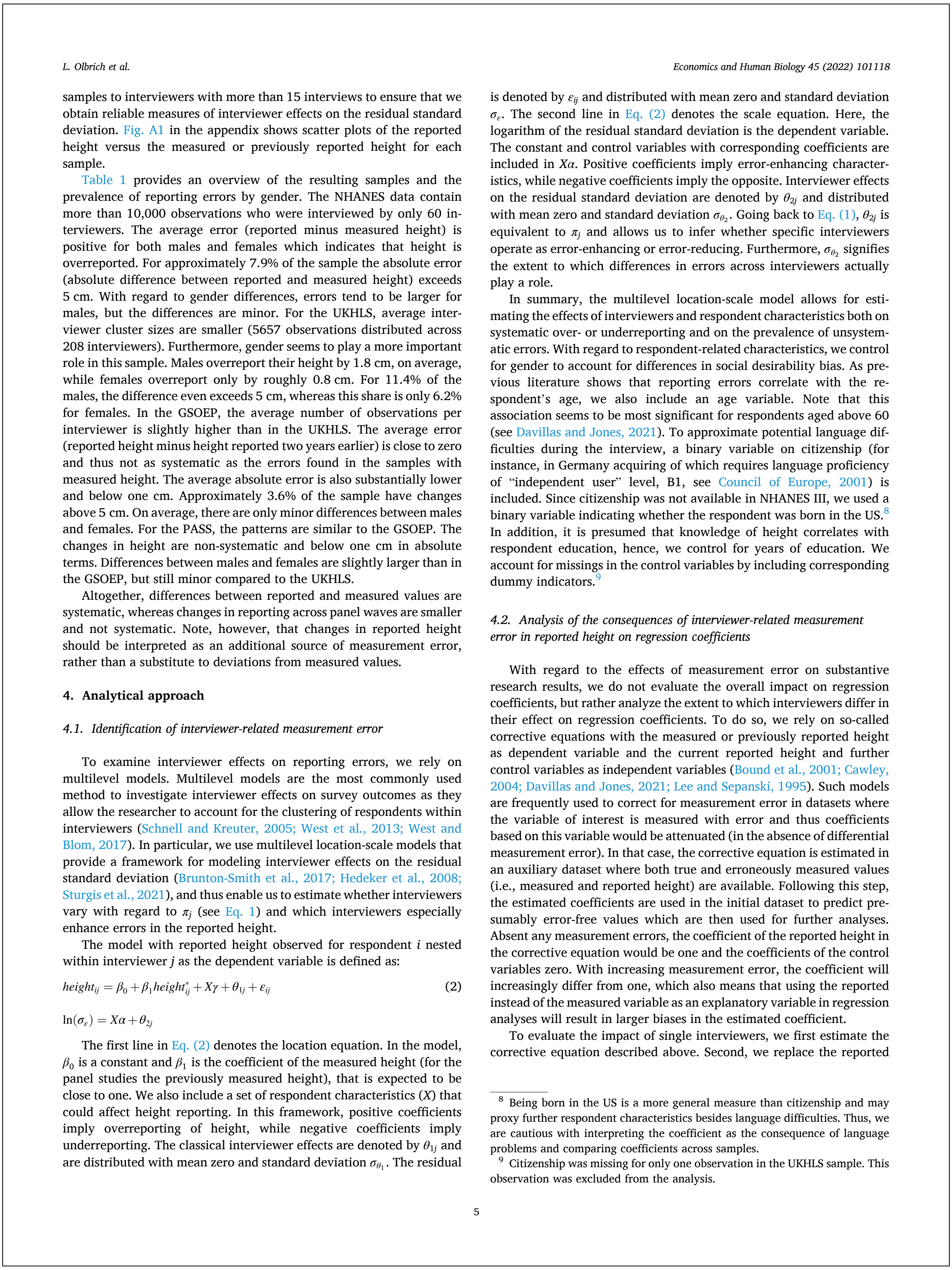
Across all datasets, we apply the same sample restrictions. Respondents for which either reported height or measured/previously reported height is missing are excluded from the analysis.[7] With regard to the interview mode, the analysis is restricted to face-to-face interviews with the intended respondent. All telephone interviews, self-administered interviews without interviewer presence, and proxy interviews are excluded. Respondents who are younger than 21 years of age and older than 60 are dropped to avoid natural growth and shrinkage effects (Case and Paxson, 2008; Fernihough and McGovern, 2015). We exclude respondents who reported height values below 130 cm and respondents with height deviations of 30 cm or higher to ensure that our results are not driven by outliers. Lastly, we restrict the

[4] BHPS respondents were sampled for nurse visits in wave 3. As we do not have data on self-reported height for these respondents, these data were not included in our analysis.

[5] More information on the GSOEP is available at https://www.diw.de/soep.

[6] Note that this formulation explicitly allows for providing imprecise responses. However, as argued above, such imprecision should occur within a limited range and randomly vary across interviewers.

[7] Interviewers who systematically produce item nonresponse for the height question will therefore not be identified by our analysis.



samples to interviewers with more than 15 interviews to ensure that we obtain reliable measures of interviewer effects on the residual standard deviation. Fig. A1 in the appendix shows scatter plots of the reported height versus the measured or previously reported height for each sample.

Table 1 provides an overview of the resulting samples and the prevalence of reporting errors by gender. The NHANES data contain more than 10,000 observations who were interviewed by only 60 interviewers. The average error (reported minus measured height) is positive for both males and females which indicates that height is overreported. For approximately 7.9% of the sample the absolute error (absolute difference between reported and measured height) exceeds 5 cm. With regard to gender differences, errors tend to be larger for males, but the differences are minor. For the UKHLS, average interviewer cluster sizes are smaller (5657 observations distributed across 208 interviewers). Furthermore, gender seems to play a more important role in this sample. Males overreport their height by 1.8 cm, on average, while females overreport only by roughly 0.8 cm. For 11.4% of the males, the difference even exceeds 5 cm, whereas this share is only 6.2% for females. In the GSOEP, the average number of observations per interviewer is slightly higher than in the UKHLS. The average error (reported height minus height reported two years earlier) is close to zero and thus not as systematic as the errors found in the samples with measured height. The average absolute error is also substantially lower and below one cm. Approximately 3.6% of the sample have changes above 5 cm. On average, there are only minor differences between males and females. For the PASS, the patterns are similar to the GSOEP. The changes in height are non-systematic and below one cm in absolute terms. Differences between males and females are slightly larger than in the GSOEP, but still minor compared to the UKHLS.

Altogether, differences between reported and measured values are systematic, whereas changes in reporting across panel waves are smaller and not systematic. Note, however, that changes in reported height should be interpreted as an additional source of measurement error, rather than a substitute to deviations from measured values.

4. Analytical approach

4.1. Identification of interviewer-related measurement error

To examine interviewer effects on reporting errors, we rely on multilevel models. Multilevel models are the most commonly used method to investigate interviewer effects on survey outcomes as they allow the researcher to account for the clustering of respondents within interviewers (Schnell and Kreuter, 2005; West et al., 2013; West and Blom, 2017). In particular, we use multilevel location-scale models that provide a framework for modeling interviewer effects on the residual standard deviation (Brunton-Smith et al., 2017; Hedeker et al., 2008; Sturgis et al., 2021), and thus enable us to estimate whether interviewers vary with regard to π_j (see Eq. 1) and which interviewers especially enhance errors in the reported height.

The model with reported height observed for respondent i nested within interviewer j as the dependent variable is defined as:

$$height_{ij} = \beta_0 + \beta_1 height^*_{ij} + X\gamma + \theta_{1j} + \varepsilon_{ij} \tag{2}$$

$$\ln(\sigma_\varepsilon) = X\alpha + \theta_{2j}$$

The first line in Eq. (2) denotes the location equation. In the model, β_0 is a constant and β_1 is the coefficient of the measured height (for the panel studies the previously measured height), that is expected to be close to one. We also include a set of respondent characteristics (X) that could affect height reporting. In this framework, positive coefficients imply overreporting of height, while negative coefficients imply underreporting. The classical interviewer effects are denoted by θ_{1j} and are distributed with mean zero and standard deviation σ_{θ_1}. The residual is denoted by ε_{ij} and distributed with mean zero and standard deviation σ_ε. The second line in Eq. (2) denotes the scale equation. Here, the logarithm of the residual standard deviation is the dependent variable. The constant and control variables with corresponding coefficients are included in $X\alpha$. Positive coefficients imply error-enhancing characteristics, while negative coefficients imply the opposite. Interviewer effects on the residual standard deviation are denoted by θ_{2j} and distributed with mean zero and standard deviation σ_{θ_2}. Going back to Eq. (1), θ_{2j} is equivalent to π_j and allows us to infer whether specific interviewers operate as error-enhancing or error-reducing. Furthermore, σ_{θ_2} signifies the extent to which differences in errors across interviewers actually play a role.

In summary, the multilevel location-scale model allows for estimating the effects of interviewers and respondent characteristics both on systematic over- or underreporting and on the prevalence of unsystematic errors. With regard to respondent-related characteristics, we control for gender to account for differences in social desirability bias. As previous literature shows that reporting errors correlate with the respondent's age, we also include an age variable. Note that this association seems to be most significant for respondents aged above 60 (see Davillas and Jones, 2021). To approximate potential language difficulties during the interview, a binary variable on citizenship (for instance, in Germany acquiring of which requires language proficiency of "independent user" level, B1, see Council of Europe, 2001) is included. Since citizenship was not available in NHANES III, we used a binary variable indicating whether the respondent was born in the US.[8] In addition, it is presumed that knowledge of height correlates with respondent education, hence, we control for years of education. We account for missings in the control variables by including corresponding dummy indicators.[9]

4.2. Analysis of the consequences of interviewer-related measurement error in reported height on regression coefficients

With regard to the effects of measurement error on substantive research results, we do not evaluate the overall impact on regression coefficients, but rather analyze the extent to which interviewers differ in their effect on regression coefficients. To do so, we rely on so-called corrective equations with the measured or previously reported height as dependent variable and the current reported height and further control variables as independent variables (Bound et al., 2001; Cawley, 2004; Davillas and Jones, 2021; Lee and Sepanski, 1995). Such models are frequently used to correct for measurement error in datasets where the variable of interest is measured with error and thus coefficients based on this variable would be attenuated (in the absence of differential measurement error). In that case, the corrective equation is estimated in an auxiliary dataset where both true and erroneously measured values (i.e., measured and reported height) are available. Following this step, the estimated coefficients are used in the initial dataset to predict presumably error-free values which are then used for further analyses. Absent any measurement errors, the coefficient of the reported height in the corrective equation would be one and the coefficients of the control variables zero. With increasing measurement error, the coefficient will increasingly differ from one, which also means that using the reported instead of the measured variable as an explanatory variable in regression analyses will result in larger biases in the estimated coefficient.

To evaluate the impact of single interviewers, we first estimate the corrective equation described above. Second, we replace the reported

[8] Being born in the US is a more general measure than citizenship and may proxy further respondent characteristics besides language difficulties. Thus, we are cautious with interpreting the coefficient as the consequence of language problems and comparing coefficients across samples.

[9] Citizenship was missing for only one observation in the UKHLS sample. This observation was excluded from the analysis.

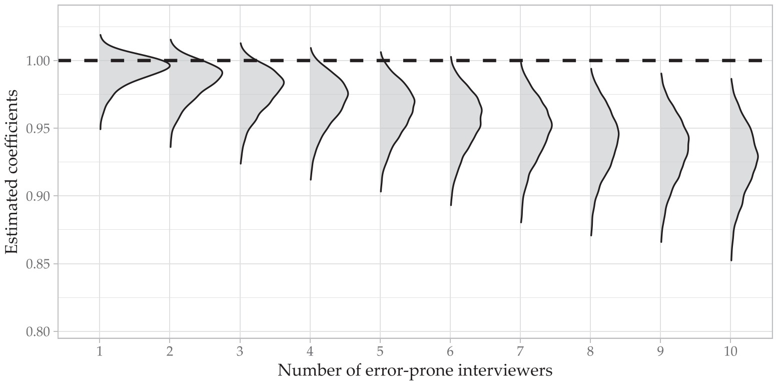


Fig. 1. Simulation results.

height of the interviewer under investigation with the measured or previously reported values while leaving the values of all other interviewers as is and fit the corrective equation again. Next, we calculate the difference between the resulting height coefficient and the height coefficient for the data without replacement. Lastly, as the change in the coefficient is also dependent on the interviewer's workload, we divide the difference by the workload. We repeat this for each interviewer separately. As a result, we obtain a measure of each interviewer's relative effect on the bias in regression coefficients.

4.2.1. *Simulation analysis*

To demonstrate the potential impact of error-prone interviewers on regression estimates, we provide a simple simulation analysis. To do so, we rely on a basic framework with measurement error in the explanatory variable. We assume a dataset consisting of 50 interviewers conducting 2500 interviews in total. The explanatory variable x_i is distributed with mean 167.75 and standard deviation 9.64 (measured height values taken from the NHANES sample) and the true regression equation is $y_i = 0.5 + 1x_i + \nu_i$, where ν_i follows a normal distribution with mean zero and a standard deviation of three.

We evaluate the impact of error-prone interviewers by simulating random error with mean zero in xi for single interviewers and re-estimating the regression. We do that for a varying number of error-prone interviewers (1–10). In every repetition, each interviewer's share of interviews is determined by drawing a value from a uniform distribution that is then rescaled to sum up to one across all interviewers. The standard deviation of the random error is drawn from a uniform distribution between one and ten for each error-prone interviewer (plausible values based on the multilevel analysis). These parameters depict that an interviewer's influence depends on the size of the random error and the interviewer's workload. For each number of error-prone interviewers we run 10,000 replications. The results are depicted in Fig. 1. The y-axis denotes the estimated coefficients, the x-axis denotes the number of error-prone interviewers, and the dashed horizontal line depicts the true coefficient.

The simulation results emphasize the uncertainty associated with the consequences of error-prone interviewers. For example, for the case of three error-prone interviewers, the bias can be as large as 7.5%, while it is below 2.5% for the majority of samples. With increasing numbers of error-prone interviewers, this uncertainty steadily increases. Hence, in some samples error-prone interviewers have little impact as their workload is small and their errors are negligible. However, in some cases interviewers with large workloads substantially exacerbate errors which can heavily impact regression coefficients.

5. Results

We fit the multilevel location-scale models presented in section Analytical approach using Markov Chain Monte Carlo Methods. To do so, we rely on the R-package brms (Bürkner, 2018) that was developed to use Stan (Carpenter et al., 2017) via R (R Core Team, 2020). We fit four chains with 16,000 iterations and a burn-in period of 8000 draws for each model. As priors we use the default non-informative priors for coefficients and half student-t priors with 3 degrees of freedom and a minimal scale parameter of 2.5 for the intercepts and the standard deviations of the interviewer effects. Convergence of the chains was assessed by $\widehat{R}$ (Gelman et al., 2013).

5.1. *Correlates of reporting errors*

Table 2 reports posterior means and 95% credible intervals of the covariates in the multilevel location-scale model for each sample. The location equation section denotes the extent to which the covariates contribute to systematic over- or underreporting of height. The scale equation denotes whether the covariates are associated with higher or lower residual variation.

With regard to the interviewers, the multilevel models indicate clustering effects both in the location and scale equations. Interviewer effects in the location equation are larger for the NHANES and UKHLS, which shows that interviewers play a lesser role for systematic differences between self-reported values. For the scale equation, we find evidence for a larger heterogeneity of interviewer scale effects in the GSOEP and PASS. Before investigating these effects in more detail in the next section, we briefly summarize the associations of respondent characteristics with height misreporting.

First, we discuss the results for the location equation. As expected, the coefficient for the measured or previously reported height is close to one across all samples. We also find that males are subject to more overreporting than females across all samples. In line with the descriptive results, this difference is more pronounced for the NHANES and the UKHLS. For the age groups, the results are less consistent. The coefficients provide evidence that compared to the reference group of 21–30 year-olds, the oldest age group is overreporting their height in the NHANES and UKHLS. For the GSOEP and PASS, age does not seem to play a role for reporting errors. Years of education are not associated with over- or underreporting across all samples. Lastly, citizenship or being born in the respective country correlates with underreporting in the NHANES, while we find no evidence for such a pattern in the UKHLS and GSOEP and only little evidence of overreporting in the PASS.

For the scale equation, positive coefficients denote higher residual

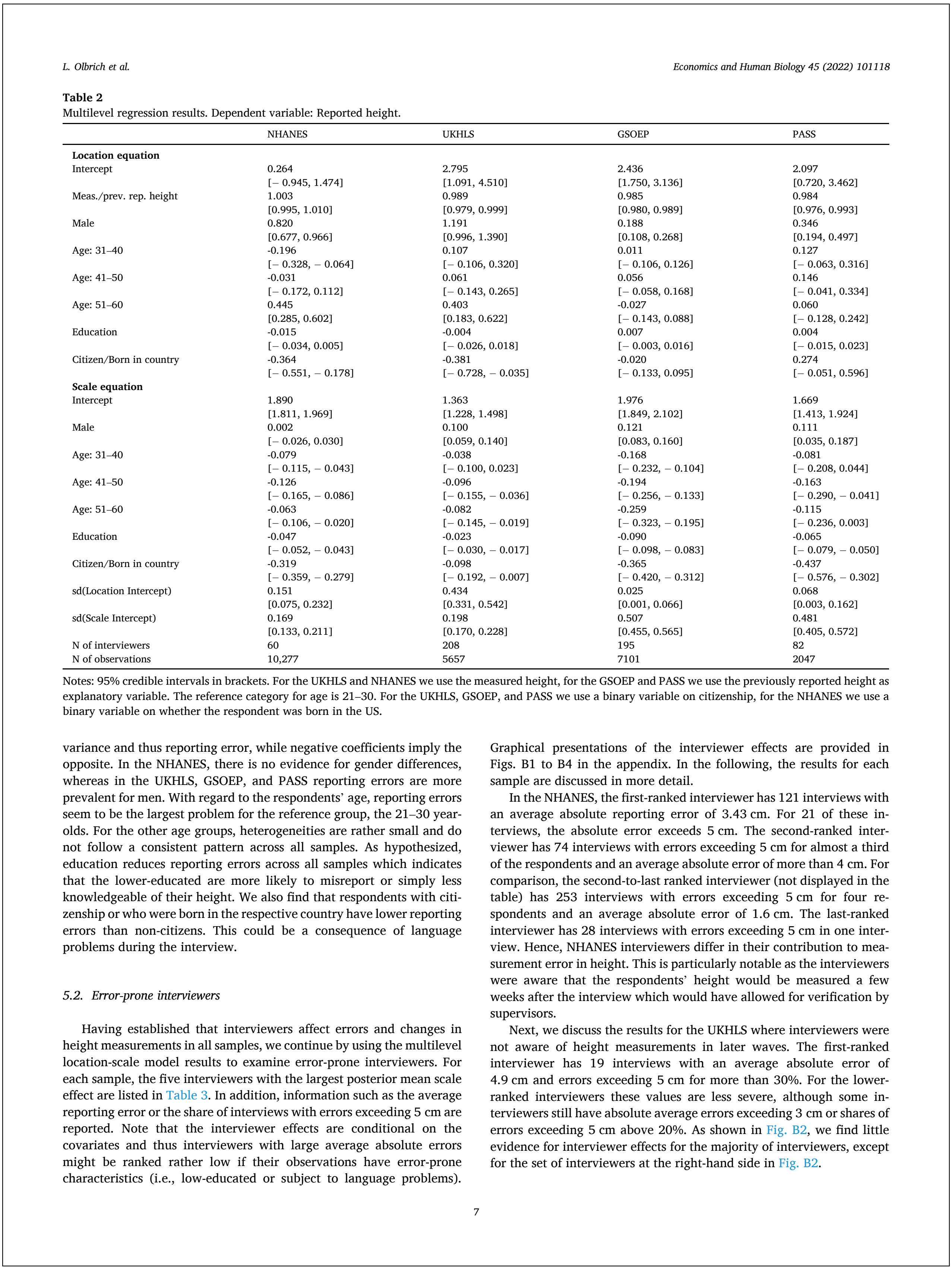


Table 2
Multilevel regression results. Dependent variable: Reported height.

	NHANES	UKHLS	GSOEP	PASS
Location equation				
Intercept	0.264	2.795	2.436	2.097
	[− 0.945, 1.474]	[1.091, 4.510]	[1.750, 3.136]	[0.720, 3.462]
Meas./prev. rep. height	1.003	0.989	0.985	0.984
	[0.995, 1.010]	[0.979, 0.999]	[0.980, 0.989]	[0.976, 0.993]
Male	0.820	1.191	0.188	0.346
	[0.677, 0.966]	[0.996, 1.390]	[0.108, 0.268]	[0.194, 0.497]
Age: 31–40	-0.196	0.107	0.011	0.127
	[− 0.328, − 0.064]	[− 0.106, 0.320]	[− 0.106, 0.126]	[− 0.063, 0.316]
Age: 41–50	-0.031	0.061	0.056	0.146
	[− 0.172, 0.112]	[− 0.143, 0.265]	[− 0.058, 0.168]	[− 0.041, 0.334]
Age: 51–60	0.445	0.403	-0.027	0.060
	[0.285, 0.602]	[0.183, 0.622]	[− 0.143, 0.088]	[− 0.128, 0.242]
Education	-0.015	-0.004	0.007	0.004
	[− 0.034, 0.005]	[− 0.026, 0.018]	[− 0.003, 0.016]	[− 0.015, 0.023]
Citizen/Born in country	-0.364	-0.381	-0.020	0.274
	[− 0.551, − 0.178]	[− 0.728, − 0.035]	[− 0.133, 0.095]	[− 0.051, 0.596]
Scale equation				
Intercept	1.890	1.363	1.976	1.669
	[1.811, 1.969]	[1.228, 1.498]	[1.849, 2.102]	[1.413, 1.924]
Male	0.002	0.100	0.121	0.111
	[− 0.026, 0.030]	[0.059, 0.140]	[0.083, 0.160]	[0.035, 0.187]
Age: 31–40	-0.079	-0.038	-0.168	-0.081
	[− 0.115, − 0.043]	[− 0.100, 0.023]	[− 0.232, − 0.104]	[− 0.208, 0.044]
Age: 41–50	-0.126	-0.096	-0.194	-0.163
	[− 0.165, − 0.086]	[− 0.155, − 0.036]	[− 0.256, − 0.133]	[− 0.290, − 0.041]
Age: 51–60	-0.063	-0.082	-0.259	-0.115
	[− 0.106, − 0.020]	[− 0.145, − 0.019]	[− 0.323, − 0.195]	[− 0.236, 0.003]
Education	-0.047	-0.023	-0.090	-0.065
	[− 0.052, − 0.043]	[− 0.030, − 0.017]	[− 0.098, − 0.083]	[− 0.079, − 0.050]
Citizen/Born in country	-0.319	-0.098	-0.365	-0.437
	[− 0.359, − 0.279]	[− 0.192, − 0.007]	[− 0.420, − 0.312]	[− 0.576, − 0.302]
sd(Location Intercept)	0.151	0.434	0.025	0.068
	[0.075, 0.232]	[0.331, 0.542]	[0.001, 0.066]	[0.003, 0.162]
sd(Scale Intercept)	0.169	0.198	0.507	0.481
	[0.133, 0.211]	[0.170, 0.228]	[0.455, 0.565]	[0.405, 0.572]
N of interviewers	60	208	195	82
N of observations	10,277	5657	7101	2047

Notes: 95% credible intervals in brackets. For the UKHLS and NHANES we use the measured height, for the GSOEP and PASS we use the previously reported height as explanatory variable. The reference category for age is 21–30. For the UKHLS, GSOEP, and PASS we use a binary variable on citizenship, for the NHANES we use a binary variable on whether the respondent was born in the US.

variance and thus reporting error, while negative coefficients imply the opposite. In the NHANES, there is no evidence for gender differences, whereas in the UKHLS, GSOEP, and PASS reporting errors are more prevalent for men. With regard to the respondents' age, reporting errors seem to be the largest problem for the reference group, the 21–30 year-olds. For the other age groups, heterogeneities are rather small and do not follow a consistent pattern across all samples. As hypothesized, education reduces reporting errors across all samples which indicates that the lower-educated are more likely to misreport or simply less knowledgeable of their height. We also find that respondents with citizenship or who were born in the respective country have lower reporting errors than non-citizens. This could be a consequence of language problems during the interview.

5.2. *Error-prone interviewers*

Having established that interviewers affect errors and changes in height measurements in all samples, we continue by using the multilevel location-scale model results to examine error-prone interviewers. For each sample, the five interviewers with the largest posterior mean scale effect are listed in Table 3. In addition, information such as the average reporting error or the share of interviews with errors exceeding 5 cm are reported. Note that the interviewer effects are conditional on the covariates and thus interviewers with large average absolute errors might be ranked rather low if their observations have error-prone characteristics (i.e., low-educated or subject to language problems). Graphical presentations of the interviewer effects are provided in Figs. B1 to B4 in the appendix. In the following, the results for each sample are discussed in more detail.

In the NHANES, the first-ranked interviewer has 121 interviews with an average absolute reporting error of 3.43 cm. For 21 of these interviews, the absolute error exceeds 5 cm. The second-ranked interviewer has 74 interviews with errors exceeding 5 cm for almost a third of the respondents and an average absolute error of more than 4 cm. For comparison, the second-to-last ranked interviewer (not displayed in the table) has 253 interviews with errors exceeding 5 cm for four respondents and an average absolute error of 1.6 cm. The last-ranked interviewer has 28 interviews with errors exceeding 5 cm in one interview. Hence, NHANES interviewers differ in their contribution to measurement error in height. This is particularly notable as the interviewers were aware that the respondents' height would be measured a few weeks after the interview which would have allowed for verification by supervisors.

Next, we discuss the results for the UKHLS where interviewers were not aware of height measurements in later waves. The first-ranked interviewer has 19 interviews with an average absolute error of 4.9 cm and errors exceeding 5 cm for more than 30%. For the lower-ranked interviewers these values are less severe, although some interviewers still have absolute average errors exceeding 3 cm or shares of errors exceeding 5 cm above 20%. As shown in Fig. B2, we find little evidence for interviewer effects for the majority of interviewers, except for the set of interviewers at the right-hand side in Fig. B2.

Table 3
Interviewers with highest scale effects.

Rank	ID	N	Scale effect	Avg. abs. error (cm)	Max. error (cm)	% Abs. error > 5 cm
NHANES						
1	78	121	0.4036	3.43	26.38	17.36
2	82	74	0.3574	4.15	22.12	29.73
3	17	141	0.2544	2.86	17.64	14.89
4	86	71	0.2414	3.13	15.08	18.31
5	31	65	0.2380	2.70	15.52	13.85
UKHLS						
1	50	19	0.5271	4.85	21.90	31.58
2	584	20	0.5028	3.34	26.74	5.00
3	303	16	0.4734	3.32	18.10	18.75
4	680	33	0.4406	3.73	15.02	24.24
5	537	49	0.4397	2.22	25.04	6.12
GSOEP						
1	408	18	1.2960	2.67	22.00	11.11
2	397	22	1.2700	1.27	20.00	4.55
3	331	47	1.1943	4.09	22.00	27.66
4	97	32	1.0686	1.53	20.00	9.38
5	416	30	1.0277	1.10	10.00	10.00
PASS						
1	217	23	0.8742	1.65	10.00	8.70
2	154	20	0.8613	3.40	10.00	35.00
3	88	16	0.8448	2.38	11.00	18.75
4	113	20	0.7553	1.25	14.00	10.00
5	241	17	0.7486	1.71	13.00	11.76

Notes: The scale effects denote the mean posteriors of the interviewer scale effects estimated in the multilevel location-scale models.

Table 4
Regression results. Dependent variable: Measured/previously reported height.

	NHANES	UKHLS	GSOEP	PASS
Intercept	26.767 ***	21.538 ***	9.155 ***	7.284 ***
	(0.909)	(1.170)	(0.812)	(1.030)
Reported height	0.821 ***	0.859 ***	0.944 ***	0.958 ***
	(0.006)	(0.007)	(0.005)	(0.006)
Male	2.029 ***	1.004 ***	0.786 ***	0.367 ***
	(0.105)	(0.127)	(0.089)	(0.123)
Age: 31–40	0.150 **	-0.256 **	-0.032	-0.152
	(0.072)	(0.110)	(0.098)	(0.157)
Age: 41–50	-0.078	-0.274 ***	-0.034	-0.290 *
	(0.078)	(0.104)	(0.097)	(0.154)
Age: 51–60	-0.518 ***	-0.796 ***	-0.096	-0.172
	(0.084)	(0.115)	(0.096)	(0.151)
Education	0.085 ***	0.049 ***	0.013	0.031 *
	(0.011)	(0.012)	(0.009)	(0.016)
Citizen/Born in country	1.104 ***	0.548 ***	0.185 **	-0.338
	(0.090)	(0.180)	(0.086)	(0.207)
Observations	10,277	5657	7101	2047
Adjusted R^2	0.914	0.925	0.948	0.959

Notes: Heteroskedasticity-robust standard-errors in parentheses.
Signif. Codes: *** : 0.01, **: 0.05, * : 0.1.

In the GSOEP, the heterogeneity across interviewers is substantially larger. For 85 of 195 interviewers, the absolute error never exceeds 5 cm and for 135 interviewers the average absolute error is below 1 cm. In that case, interviewers with rather large errors are particularly worrisome. The first and second-ranked interviewers' effect is mainly driven by extreme reporting errors. For the first-ranked interviewer 2 of 18 observations have absolute errors of 22 and 19 cm, respectively. For the second-ranked interviewer a single respondent with an absolute error of 20 cm drives the effect. While such errors are of course problematic, such isolated incidents may as well be driven by simple data entry errors. For the third-ranked interviewer this is not the case. The average absolute error is above 4 cm and the absolute error exceeds 5 cm for more than 25%. In the full sample these values are 0.9 cm and 3.6%. These differences suggest that interviewer 331 systematically drives unlikely changes in reported height.

Similar to the GSOEP, a large share of the PASS interviewers have no absolute errors exceeding 5 cm (N = 43; 52.4%) and average absolute errors below 1 cm (N = 53; 64.6%). We are particularly interested in the verified deviant interviewers 154 and 109. Interviewer 154 is ranked second and the height changes exceed 5 cm in 35% of the cases, and the average absolute error is 3.4 cm. These values clearly deviate from the sample averages. Thus, deviant interviewer 154 is characterized by frequent height changes which is in line with the hypothesis that (partial) falsifiers are unable to recover the height reported in previous waves. To the contrary, interviewer 109 is ranked 72nd with an average absolute error below 0.6 cm. Given that this interviewer presumably reinforced deviant habits over time, this indicates that the height question was not subject to deviant behavior in waves 9 and 12. For the other high-ranked interviewers, both the shares of large errors and the average errors are lower than for interviewer 154 and thus a lesser reason for concern.

In summary, the analysis of the NHANES and UKHLS data shows that there are several interviewers who are particularly prone to errors in height reporting. Furthermore, the analysis of the GSOEP and PASS data demonstrated that single interviewers also exacerbate changes in height reporting that should not occur. For the PASS data, we even find that deviant interviewer behavior can lead to frequent changes in height and that deviant interviewers can even be detected by such behavior. With regard to the analyses of the other samples, this implies that inadequate interviewer behavior might explain misreporting or frequent changes in reported height.

5.3. Effects of height errors on substantive regression estimates

In this section, we evaluate whether interviewers differ in their contribution to bias in regression coefficients using corrective equations. Table 4 reports the results for the corrective equations for each sample. The estimated coefficients show that the height coefficients are attenuated across all samples. This attenuation is larger for the NHANES and UKHLS. Several of the control variables associated with reporting errors are statistically significant (see Section 5.1). The results of the analysis procedure described in Section 4.2 are depicted in Fig. 2. The x-axis denotes the rank of the interviewers by their scale effect taken from the results of the multilevel location-scale models (in ascending order). The y-axis shows the average change per respondent in the coefficient of the reported height when the respective interviewer's values are replaced by the measured or previously reported height, i.e., what happens on average when the reported value of one observation of the respective interviewer is replaced with the measured value. For example, for the NHANES this corresponds to the change in the coefficient when one out of 10,277 values is replaced. Positive values depict a reduction in attenuation. A local linear regression line was added to the graphs to show whether the relative coefficient change varies across the interviewer ranks.

In the NHANES, the replacement of reported with measured values reduces attenuation for every interviewer. However, the reduction is not constant across all interviewers. For the interviewers with ranks above 40, the average changes in the coefficient are steadily increasing, which is also denoted by the positive slope of the local linear regression line. At the same time, the variation of average changes is increasing, indicating that the higher-ranked interviewers are heterogeneous in their error pattern. For the UKHLS, several interviewers have slightly negative values, but for the vast majority of interviewers the changes are positive. Similar to the NHANES, the average coefficient change is homogeneously distributed for most of the interviewers and substantially increases for the highest-ranked interviewers. The steep slope of the regression line shows that the highest-ranked interviewers have larger effects on the coefficient than the remaining interviewers. This provides evidence that some interviewers enhance bias in regression coefficients induced by measurement error. Next, we turn to the GSOEP and PASS, where reported height values are replaced by previously reported

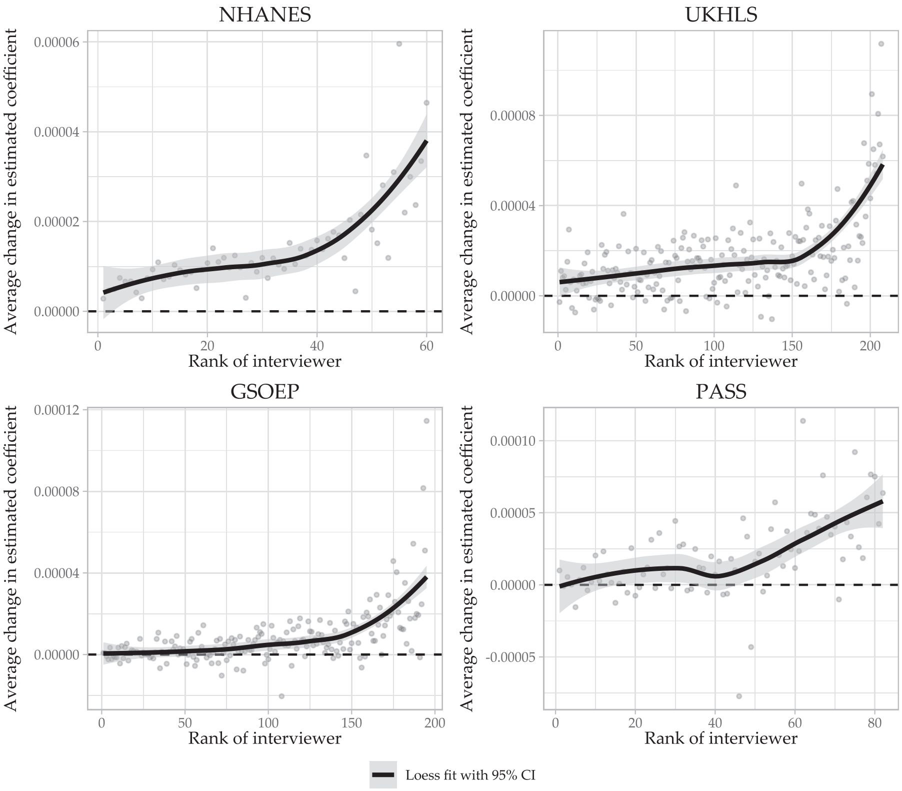


Fig. 2. Interviewer effects on regression estimates.

values. In the GSOEP, we find a similar pattern as for the NHANES and UKHLS. For the first 150 interviewers, the changes are relatively small and distributed around zero. For the higher-ranked interviewers, the changes and their variation increase. For the PASS, we also observe an increase in the average change and the heterogeneity of changes. Hence, the coefficient attenuation driven by changes in height reporting is not equally distributed across interviewers.

Our results imply that particularly error-prone interviewers can have – compared to the remaining interviewers – larger effects on regression coefficients in all samples. By depicting the relevance of the interviewer ranking derived from the multilevel location-scale models, this analysis also shows that the multilevel results can be used to target suspicious interviewers for further controls to limit their effects on research results.

6. Discussion

Anthropometric data are of crucial importance in the health and social sciences. Given the lack of administrative data for the general population, anthropometric data are often collected in surveys by directly asking respondents to report their values. While respondent misreporting of such data is well-documented in the literature, the role of the interviewers who collect these measures has not been extensively examined. Both sources of error could lead to false conclusions in health and social research, for instance, with regard to the impact of anthropometric measures on individual life-course outcomes. Using data from four large-scale surveys (NHANES, UKHLS, GSOEP, PASS) and multilevel models, we examined two types of errors: differences between reported and measured height and unlikely changes in adult self-reported height over time. After providing multiple theoretical explanations for the causes of interviewer effects on erroneous height reporting, we first assessed empirically whether these errors are associated with interviewers. Using multilevel model results, we also examined whether height reporting errors allow for identifying particularly error-prone interviewers. Lastly, we evaluated whether the effects of height reporting errors on substantive research results are equally distributed across interviewers. This is a novel contribution and important in understanding the role of interviewer-related measurement error on self-reported anthropometric data and the consequent effect of bias in substantive data analyses.

Our results revealed that interviewers affect both differences between reported and measured height and unlikely changes in adult self-reported height over panel survey waves. Our results further indicated that these errors accumulate for several interviewers. As one of the datasets contained falsified interviews, we were able to demonstrate that height changes can accumulate for deviant interviewers. Moreover, we found that the effects of height reporting errors on regression coefficients are not equally distributed across interviewers. A simulation analysis of the effects of error-prone interviewers showed that in many cases error-prone interviewers will have little consequence on regression coefficients as their workloads and error sizes are small. However, if the error-prone interviewers' share of interviews increases or the size of their errors increases, substantive research results will be increasingly biased. In practice, interviewer effects on additional analysis variables might further enhance bias in estimated parameters. This is particularly relevant for studies using the body mass index, where both components – height and weight – can be subject to interviewer effects.

All surveys in our analysis had interviewer monitoring procedures in place. In surveys with no (or poor) monitoring procedures, the effects and consequences of interviewers documented in this article would likely be reinforced. This calls for putting limits on interviewer workloads and thorough interviewer training and monitoring to minimize the prevalence and impact of error-prone interviewing. Note that monitoring systems can be quite inexpensive and easily automated (for instance by comparing previously reported height to the reported height in the current wave), and thus the benefits of monitoring interviewers will outweigh costs in most cases. Furthermore, data producers should provide interviewer IDs to researchers to enable testing and accounting for interviewer effects in their analyses.

Our analysis is subject to several limitations. While our approach flagged a verified deviant interviewer, we were unable to identify the exact reasons why frequent height errors occurred for other error-prone interviewers. Practitioners may shed more light on this issue by applying the analysis during the fieldwork period and following up with flagged interviewers. Second, we were unable to fully disentangle respondent and interviewer effects. It is possible that some interviewers were predominantly assigned to presumably difficult respondents characterized by erroneous height reporting. However, interviewers usually work in geographic clusters and are not assigned to respondents solely based on their skills and the target respondents' interview difficulty. Thus, it seems unlikely that the patterns observed for error-prone interviewers were fully driven by respondents. Lastly, homogeneity of respondents within and heterogeneity across geographic regions regarding characteristics shaping reporting errors might affect our results. However, in our case, respondents from specific regions where error-prone interviewers work would have to be particularly poor at reporting their height. Given that we control for several socio-demographics and that previous research found minor effects of regions on data quality measures or reporting error (e.g., Meyer and Mittag, 2019; Sturgis et al., 2021), regional effects are unlikely to explain the prevalence of error-prone interviewers.

Our analysis exhibited a straight-forward approach to exploit height reporting to conduct data quality control procedures. The multilevel location-scale model may also be applied to further survey variables which can be compared to validation data such as register data or recontact interviews. Future research on interviewers' role in panel studies may delve into whether interviewer effects drive variation in less-stable items and whether changes in interviewer assignment increase response variation over waves. Developing further tools to ensure high data quality of panel surveys is especially relevant as such databases have consistently large numbers of users in a variety of research fields.

CRediT authorship contribution statement

Lukas Olbrich has completed the data analyses, drafted the paper, and revised the paper based upon Yuliya Kosyakova's and Joseph W. Sakshaug's comments. Yuliya Kosyakova has supervised the process of data analyses, reviewed, edited, and commented on various versions of the paper. Joseph W. Sakshaug has supervised the process of data analyses, reviewed, edited, and commented on various versions of the paper.

Acknowledgements

Lukas Olbrich gratefully acknowledges financial support from the Graduate Programme of the Institute for Employment Research (IAB) and the University of Erlangen-Nuremberg (GradAB). Joseph W. Sakshaug gratefully acknowledges financial support from the Ludwig-Maximilian University of Munich. We are grateful to two anonymous reviewers and the editor for very helpful comments.

Appendix

See Fig. A1 and Fig. B1, Fig. B2, Fig. B3, Fig. B4.

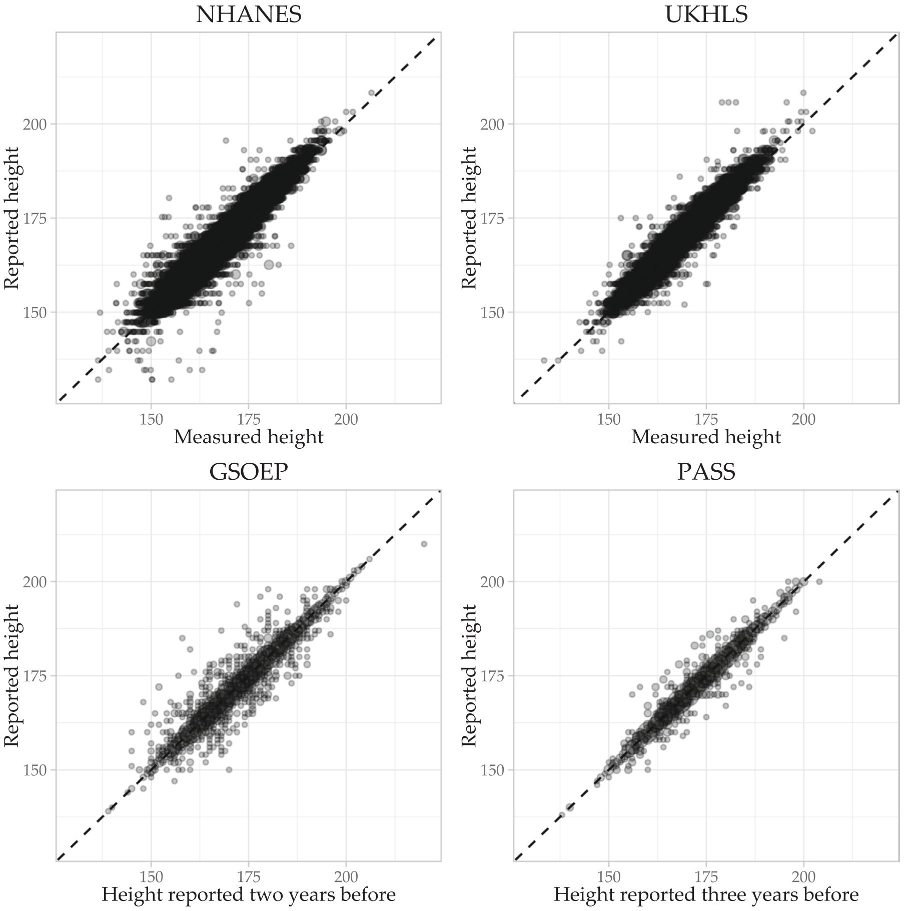


Fig. A1. Reported height versus measured/previously reported height (in cm).

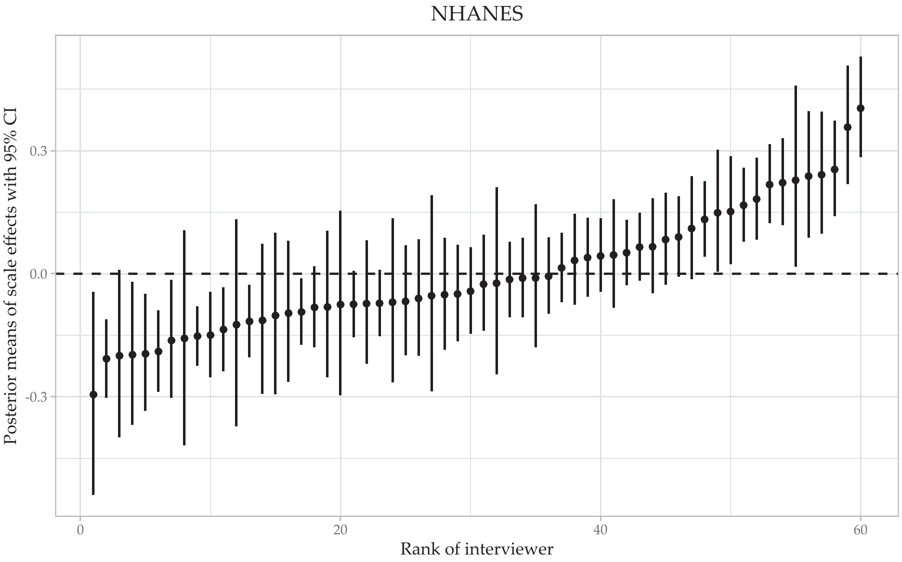


Fig. B1. Interviewer effects on scale, NHANES.

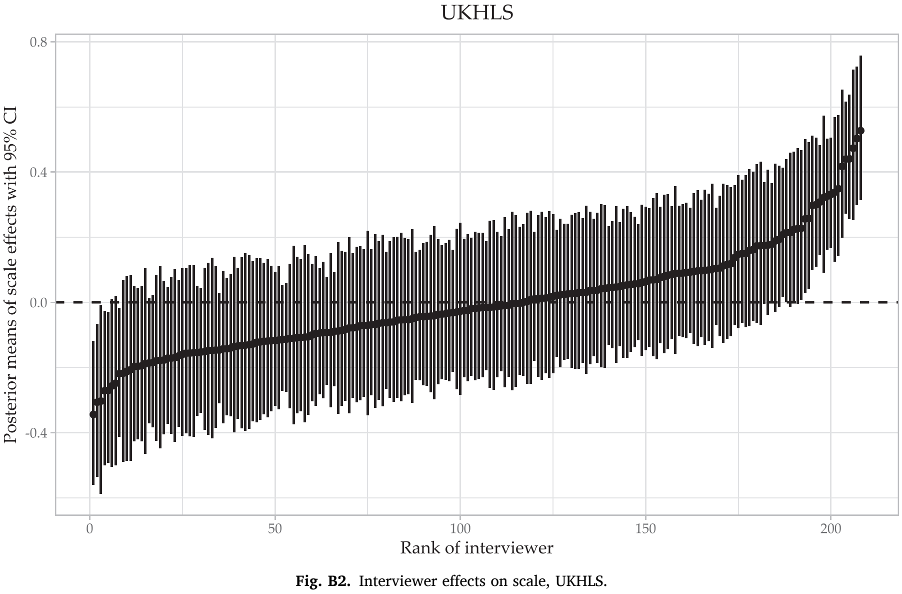


Fig. B2. Interviewer effects on scale, UKHLS.

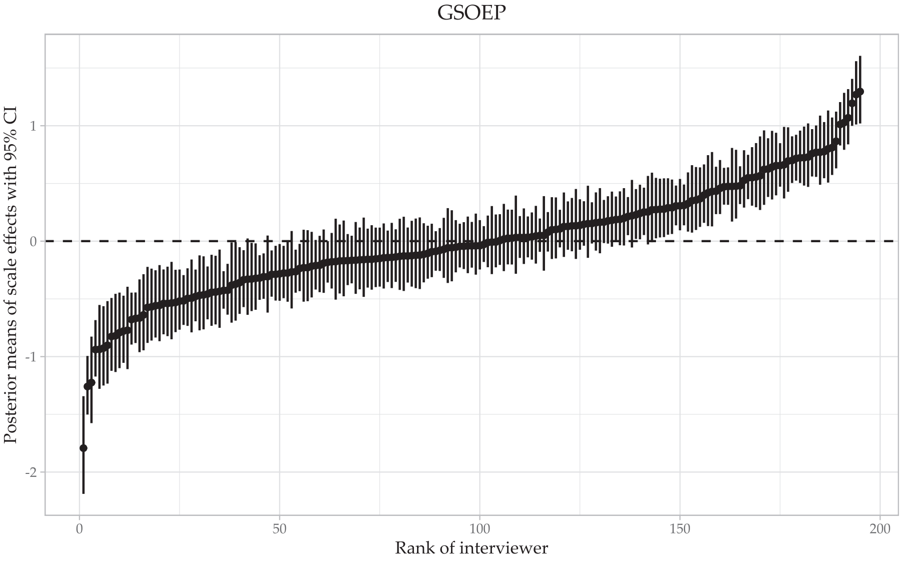


Fig. B3. Interviewer effects on scale, GSOEP.

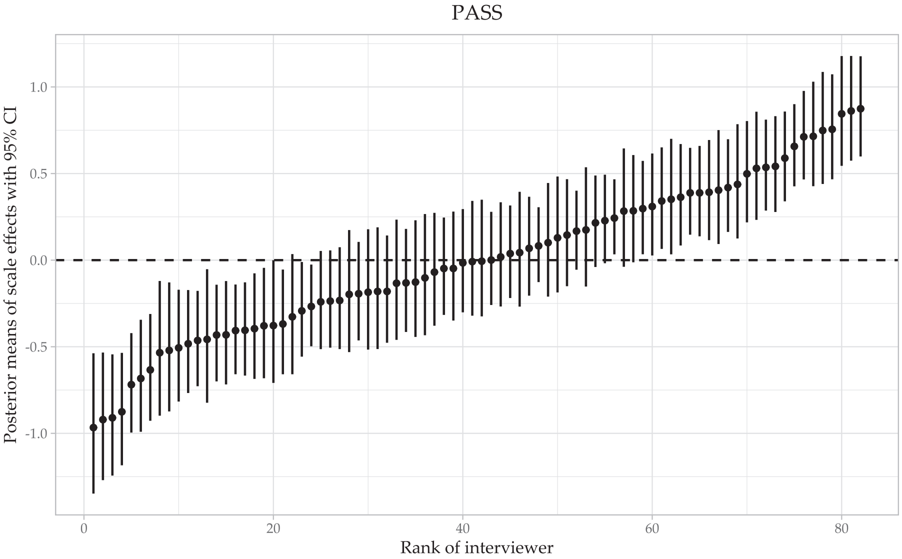


Fig. B4. Interviewer effects on scale, PASS.

References

AAPOR, 2003. Interviewer falsification in survey research: Current best methods for prevention, detection, and repair of its effects. 〈https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/falsification.pdf〉.

Batty, G.D., Shipley, M.J., Gunnell, D., Huxley, R., Kivimaki, M., Woodward, M., Lee, C.M.Y., Smith, G.D., 2009. Height, wealth, and health: An overview with new data from three longitudinal studies. Econ. Hum. Biol. 7 (2), 137–152.

Beste, J., Olbrich, L., & Schwanhäuser, S. (2021). Interviewer:innenkontrolle im Panel Arbeitsmarkt und soziale Sicherung (PASS). FDZ-Methodenbericht 4.

Boreham, R., Boldysevaite, D., Killpack, C., 2012. UKHLS: Wave 1 Technical Report.

Boström, G., Diderichsen, F., 1997. Socioeconomic differentials in misclassification of height, weight and body mass index based on questionnaire data. Int. J. Epidemiol. 26 (4), 860–866.

Bound, J., Brown, C., Mathiowetz, N., 2001. Measurement error in survey data. In: Heckman, J.J., Leamer, E.E. (Eds.), Handbook of Econometrics, fifth ed. Elsevier Science B.V, pp. 3705–3843.

Brunton-Smith, I., Sturgis, P., Leckie, G., 2017. Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model. J. R. Stat. Soc. Ser. A Stat. Soc. 180 (2), 551–568.

Buck, N., McFall, S., 2012. Understanding Society: design overview. Longitud. Life Course Stud. 3 (1), 5–17.

Burke, M.A., Carman, K.G., 2017. You can be too thin (but not too tall): Social desirability bias in self-reports of weight and height. Econ. Hum. Biol. 27, 198–222.

Bürkner, P.C., 2018. Advanced Bayesian multilevel modeling with the R package brms. R J. 10 (1), 395–411.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: A probabilistic programming language. J. Stat. Softw. 76 (1), 1–32.

Case, A., Paxson, C., 2008. Stature and status: height, ability, and labor market outcomes. J. Political Econ. 116 (3), 499–532.

Cawley, J., 2004. The impact of obesity on wages. J. Hum. Resour. 39 (2), 451–474.

Council of Europe, 2001. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press.

Crossley, T.F., Schmidt, T., Tzamourani, P., Winter, J.K., 2021. Interviewer effects and the measurement of financial literacy. J. R. Stat. Soc. Ser. A: Stat. Soc. 184 (1), 150–178.

Davillas, A., Jones, A.M., 2021. The implications of self-reported body weight and height for measurement error in BMI. Econ. Lett. 209, 110101.

De Haas, S., Winker, P., 2016. Detecting fraudulent interviewers by improved clustering methods – the case of falsifications of answers to parts of a questionnaire. J. Off. Stat. 32 (3), 643–660.

Deaton, A., Arora, R., 2009. Life at the top: the benefits of height. Econ. Hum. Biol. 7 (2), 133–136.

Eckman, S., Koch, A., 2019. Interviewer involvement in sample selection shapes the relationship between response rates and data quality. Public Opin. Q. 83 (2), 313–337.

Fernihough, A., McGovern, M.E., 2015. Physical stature decline and the health status of the elderly population in England. Econ. Hum. Biol. 16, 30–44.

Finn, A., Ranchhod, V., 2017. Genuine fakes: The prevalence and implications of data fabrication in a large South African survey. World Bank Econ. Rev. 31 (1), 129–157.

Fuller, W.A., 1987. Measurement Error Models. John Wiley & Sons, Inc.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis, third ed. CRC Press.

Gil, J., Mora, T., 2011. The determinants of misreporting weight and height: the role of social norms. Econ. Hum. Biol. 9 (1), 78–91.

Goebel, J., Grabka, M.M., Liebig, S., Kroh, M., 2019. The german socio-economic panel (SOEP). Jahrb. Natl. Stat. 239 (2), 345–360.

Gorber, S.C., Tremblay, M., Moher, D., Gorber, B., 2007. A comparison of direct vs. self-report measures for assessing height, weight and body mass index: a systematic review. Obes. Rev. 8 (4), 307–326.

Groves, R.M., 2004. Survey Errors and Survey Costs, 2nd ed..,. John Wiley & Sons, Inc,.

Hedeker, D., Mermelstein, R.J., Demirtas, H., 2008. An application of a mixed-effects location scale model for analysis of Ecological Momentary Assessment (EMA) data. Biometrics 64 (2), 627–634.

Heineck, G., 2006. Height and weight in Germany, evidence from the German Socio-Economic Panel, 2002. Econ. Hum. Biol. 4 (3), 359–382.

Josten, M., Trappmann, M., 2016. Interviewer effects on a network-size filter question. J. Off. Stat. 32 (2), 349–373.

Kosyakova, Y., Skopek, J., Eckman, S., 2015. Do interviewers manipulate responses to filter questions? Evidence from a multilevel approach. Int. J. Public Opin. Res. 27 (3), 417–431.

Kroh, M., 2005. Intervieweffekte bei der Erhebung des Körpergewichts in Bevölkerungsumfragen. Gesundheitswesen 67 (8–9), 646–655.

Larson, M.R., 2000. Social desirability and self-reported weight and height. Int. J. Obes. 24 (5), 663–665.

Lee, L.F., Sepanski, J.H., 1995. Estimation of linear and nonlinear errors-in-variables models using validation data. J. Am. Stat. Assoc. 90 (429), 130–140.

Meyer, B.D., Mittag, N., 2019. Misreporting of government transfers: how important are survey design and geography? South. Econ. J. 86 (1), 230–253.

Moore, C.J., 1988. Self/proxy response status and survey response quality. J. Off. Stat. 4 (2), 155–172.

National Center for Health Statistics, 1994. Plan and operation of the Third National Health and Nutrition Examination Survey, 1988-94. Vital.-. Health Stat. 1, 32.

O'Neill, D., Sweetman, O., 2013. The consequences of measurement error when estimating the impact of obesity on income. IZA J. Labor Econ. 2 (1), 1–20.

Persico, N., Postlewaite, A., Silverman, D., 2004. The effect of adolescent experience on labor market outcomes: the case of height. J. Political Econ. 112 (5), 1019–1053.

Pinkston, J.C., 2017. The dynamic effects of obesity on the wages of young workers. Econ. Hum. Biol. 27, 154–166.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 〈https://www.r-project.org/〉.

Schnell, R., Kreuter, F., 2005. Separating interviewer and sampling-point effects. J. Off. Stat. 21 (3), 389–410.

Schräpler, J.-P., Wagner, G.G., 2005. Identification, characteristics and impact of faked interviews in surveys: an analysis by means of genuine fakes in the raw data of SOEP. Allg. Stat. Arch. 89 (1), 7–20.

Schwanhäuser, S., Sakshaug, J.W., Kosyakova, Y., Kreuter, F., 2020. Statistical identification of fraudulent interviews in surveys - improving interviewer controls. In: Olson, K., Smyth, J.D., Dykema, J., Holbrook, A., Kreuter, F., West, B.T. (Eds.), Interviewer Effects from a Total Survey Error Perspective. CRC Press, pp. 91–106.

SOEP Group, 2020. SOEP-Core v35 – HEALTH. SOEP Survey Papers 830: Series D – Variable Descriptions and Coding.

Sturgis, P., Maslovskaya, O., Durrant, G., Brunton-Smith, I., 2021. The interviewer contribution to variability in response times in face-to-face interview surveys. J. Surv. Stat. Methodol. 9 (4), 701–721.

Tourangeau, R., Yan, T., 2007. Sensitive questions in surveys. Psychol. Bull. 133 (5), 859–883.

Trappmann, M., Bähr, S., Beste, J., Eberl, A., Frodermann, C., Gundert, S., Schwarz, S., Teichler, N., Unger, S., Wenzig, C., 2019. Data resource profile: panel study labour market and social security (PASS). Int. J. Epidemiol. 48 (5), 1411–1411 G.

Trappmann, M., Beste, J., Bethmann, A., Müller, G., 2013. The PASS panel survey after six waves. J. Labour Mark. Res. 46 (4), 275–281.

Uhrig, S.C.N., 2012. Understanding panel conditioning: an examination of social desirability bias in self-reported height and weight in panel surveys using experimental data. Longitud. Life Course Stud. 3 (1), 120–136.

University of Essex: Institute for Social and Economic Research, 2021. Understanding Society: Waves 1-11, 2009-2020 and Harmonised BHPS: Waves 1-18, 1991-2009, fourteenth ed. UK Data Service,. https://doi.org/10.5255/UKDA-SN-6614-15.

University of Essex: Institute for Social and Economic Research and National Centre for Social Research, 2014. Understanding Society: Waves 2 and 3 Nurse Health Assessment, 2010- 2012 [data collection], third e. UK Data Service,. https://doi.org/10.5255/UKDA-SN-7251-3.

Warren, J.R., Halpern-Manners, A., 2012. Panel conditioning in longitudinal social science surveys. Sociol. Methods Res. 41 (4), 491–534.

West, B.T., Blom, A.G., 2017. Explaining interviewer effects: a research synthesis. J. Surv. Stat. Methodol. 5 (2), 175–211.

West, B.T., Kreuter, F., Jaenichen, U., 2013. "Interviewer" effects in face-to-face surveys: a function of sampling, measurement error, or nonresponse? J. Off. Stat. 29 (2), 277–297.

6 Multivariate assessment of interviewer errors in a cross-national economic survey

Declaration of Contributions

Lukas Olbrich had the idea for the analysis approach, implemented the analyses, and wrote the paper.

Contributions of Co-authors

Elisabeth Beckmann provided information on the survey data and the data collection processes and managed the communication with the fieldwork institutes. She also wrote the "The OeNB Euro Survey" chapter, provided feedback on the analysis, and revised and proofread the paper. Joseph W. Sakshaug provided guidance and valuable input, and revised and proofread the paper.

Abstract

Interviewers have long been identified as a source of error in face-to-face surveys. However, previous studies have typically focused on a single source of interviewer error and single-country cross-sectional surveys. We extend this literature by investigating interviewer errors from multiple dimensions in the Oesterreichische Nationalbank (OeNB) Euro Survey, a cross-national survey conducted annually in ten Central, Eastern, and Southeastern European countries. Using data from ten rounds (i.e., 100 country-years), we apply several data quality indicators on various dimensions of interviewer error and investigate country-years with particularly exceptional patterns. To combine the indicators, we use a multivariate tree-based outlier detection method (isolation forest) that flags country-years and interviewers with outlying values and combine it with methods from the interpretable machine learning literature to identify the respective exceptional feature values. Lastly, we document the effects of interviewer errors on the bias and variance of survey estimates. In several instances, our results identify fieldwork institutes and supervisors rather than interviewers as the main source of error.

Keywords

interviewer effects, survey data quality, multilevel modeling, interviewer falsification, interviewer variance

Acknowledgements

This project was funded by a Klaus-Liebscher-Economic-Research-Scholarship (OeNB). The authors would like to thank Melanie Koch, Katharina Allinger, Julia Woerz, Helmut Stix, Thomas Scheiber, participants of the WAPOR 2023 conference, the LMU summer retreat 2024, internal IAB and OeNB seminars, and three anonymous reviewers for their helpful feedback and comments.

Note

An earlier version of the paper was published as an OeNB Working Paper: Olbrich, Lukas; Beckmann, Elisabeth; Sakshaug, Joseph W. (2024): Multivariate assessment of interviewer-related errors in a cross-national economic survey, Working Paper, No. 253, Oesterreichische Nationalbank (OeNB), Vienna.

6.1 Introduction

Interviewers are an important source of error in face-to-face surveys (e.g., Crespi, 1945; Kish, 1962). Throughout the survey lifecycle, interviewers can influence survey estimates in multiple ways, with the sampling (if interviewers are involved), recruitment, and measurement stages being particularly prone to interviewer errors (West & Blom, 2017). Deviations from sampling instructions induce sampling error, recruiting only specific respondent subgroups leads to nonresponse error, and influences on responses lead to measurement error. These errors may arise from unintentional or intentional errors. Unintentional errors are mainly driven by the respondent's impression of the interviewer (i.e., due to their observable characteristics) or the mere presence of interviewers (Kreuter et al., 2010). Intentional (or deliberate) interviewer errors arise from interviewers willingly deviating from their instructions and guidelines, for example, by skipping parts of the questionnaire text, skipping response options, fabricating (parts of) interviews, or intentionally deviating from random route instructions and are also labeled *interviewer falsification* (Groves, 2004). The main focus of this article are intentional errors.

Most research on the role of interviewers on various outcomes has focused on single-country cross-sectional surveys (e.g., Brunton-Smith et al., 2017; Olson & Peytchev, 2007; Schnell & Kreuter, 2005; West & Olson, 2010). In recent years, however, the availability of large-scale cross-national surveys has facilitated analyses on the role of interviewers by putting their effects into perspective, identifying countries that require more or less caution concerning interviewers, and identifying correlates of interviewer errors. These studies include research on interviewer variance in substantive questionnaire items and biomeasure collection (e.g., Beullens & Loosveldt, 2016; Waldmann et al., 2023; Zins & Burgard, 2020), interviewer variance in data quality indicators such as straightlining or interview duration (e.g., Loosveldt & Beullens, 2013, 2017; Vandenplas et al., 2018), interviewer influences on sample selection (e.g., Eckman & Koch, 2019; Kohler, 2007; Menold, 2014), and interviewers' role on data anomalies (e.g., Blasius & Thiessen, 2021). However, none of these studies have jointly analyzed multiple indicators of interviewer error.

This study investigates interviewer errors in the OeNB Euro Survey from 2012 to 2021, a cross-national survey conducted annually in ten Eastern European countries. The OeNB Euro Survey is particularly well-suited for investigating the role of interviewers as 1) respondents are sampled using random route sampling, which requires enhanced interviewer involvement, 2) the development of measures of interviewer error can be examined over time and compared between countries, and 3) the OeNB Euro Survey contains several questions on financial literacy which are particularly prone to interviewer effects (Crossley et al., 2021).

Our analysis proceeds in multiple steps. First, we develop a framework of interviewer tasks and interviewer errors and show how the indicators of interviewer errors we use (internal unit nonresponse bias indicator, daily interviews per interviewer, interviewer variance, item nonresponse, straightlining, and (near-)duplicate analysis) relate to interviewer tasks. Second, we describe how the indicators of interviewer error are calculated. Third, we discuss the results and investigate country-years with particularly exceptional outcomes in detail. Fourth, we use isolation forests to flag country-years and interviewers with outlying values based on the data quality measures. Lastly, we explore the extent to which interviewer errors can impact substantive analyses.

Our analysis contributes to the literature on interviewer errors in multiple ways. First, we provide evidence of interviewer errors for countries that are rarely investigated in-depth concerning the role of interviewers. Second, we assess the development of indicators of interviewer error over time, and as some countries changed fieldwork institutes over time, we can assess concurrent changes in the error indicators. Third, we provide a more holistic approach to interviewer errors than previous studies by examining different types of errors. Fourth, researchers may use the proposed analysis approach as a template to learn about potential sources of interviewer errors in other surveys.

6.2 Interviewer errors and cross-national surveys

In this study, we investigate interviewer errors from various perspectives. Figure 6.1 provides an overview of the indicators of interviewer error and which interviewer tasks are associated with the corresponding errors. Most of these indicators have also been used in previous literature on interviewer errors in cross-country surveys. We put "Data processing" in a dashed box as the fieldwork institute is also heavily involved in this step and thus may also contribute to errors. Below, we describe the indicators and their motivation in detail and summarize previous studies using cross-country survey data.

The first indicator is an *internal criterion of unit nonresponse bias* as suggested by Sodeur (1997) and implemented by Kohler (2007). Without appropriate external criteria for evaluating nonresponse bias, internal criteria are based on sub-group characteristics for which the true population value is known. The only criterion used in previous research is the proportion of female respondents in gender-heterogeneous couples living in the same two-person household. The expected proportion is 50 percent, and deviations beyond sampling variance intervals indicate nonresponse bias. In our framework, interviewers may contribute to deviations from 50 percent by deviating from random route instructions (i.e., during sampling) and during recruitment (i.e., recruiting respondents with higher availability, Menold, 2014). Alternative error sources such as measurement error or item nonresponse are unlikely to affect the proportion (Kohler, 2007). This criterion has been widely applied in cross-national studies to investigate cultural and methodological sources of nonresponse bias. Kohler (2007) used data from six cross-national surveys and found that correlates of higher bias (due to sampling method, backchecking procedures, substitutions) are strongly related to deflective interviewer behavior. Jabkowski and Cichocki (2019) used the internal bias measure to show that unit nonresponse bias in the ESS is higher for personal-register samples than for non-personal-register samples. Menold (2014) and Eckman and Koch (2019) used ESS data and compared nonresponse bias measures across sampling methods. Menold (2014) found larger biases when interviewers are more involved in sample selection (i.e., random route and listing-based samples) and have more leeway to deviate. Eckman and Koch (2019) used internal and external criteria (i.e., reference values from a benchmark survey) and found that more interviewer involvement correlates with higher bias and is a key mediator for the relationship between response rates and nonresponse bias. Hence, the OeNB Euro Survey is particularly prone to such deviations as previous literature has shown that random route samples tend to have higher deviations (Eckman & Koch, 2019; Kohler, 2007; Menold, 2014). None of the previous studies assessed changes in the unit nonresponse bias indicator over time.

The second indicator is the *daily interviewer workload*, which has been used for flagging suspicious interviewers in previous literature (Bushery et al., 1999). An interviewer's maximum number of successful interviews is restricted by the duration of interviews, unit nonresponse, the travel distance between respondents, and their daily working hours. Hence, exceptionally high values indicate that interviewers deviated from the prescribed random route instructions (i.e., interviewed

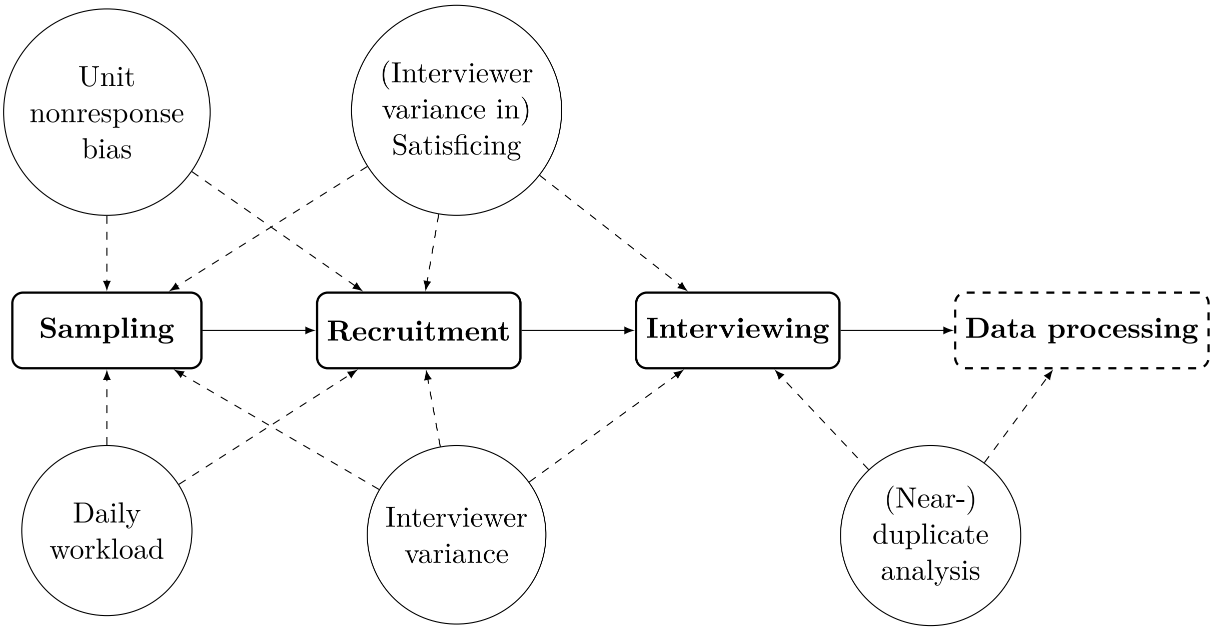


Figure 6.1: Interviewer tasks and indicators of interviewer error.

all available persons) or invested no effort into recruitment and proceeded to the next target person as fast as possible.

The third indicator is *interviewer variance*. Assuming random assignment of respondents to interviewers, respondents interviewed by the same interviewer might be more similar due to interviewers interviewing similar respondents or interviewers influencing the measurement (West & Olson, 2010; West et al., 2013, 2018). While these error types are difficult to disentangle, both lead to inflated variance estimates. In the OeNB Euro Survey, interviewers might also sample similar respondents if they deviate from random route instructions. Interviewer variance is typically denoted by the intraclass correlation coefficient (ICC) that measures the proportion of the residual variance explained by interviewers (see Section 6.4.3 for a detailed discussion). Beullens and Loosveldt (2016) analyze interviewer variance for ESS data. First, they show that academic articles using ESS data rarely take interviewer variance into account. Second, they estimate ICCs for 48 items in 36 countries and six rounds. The average ICCs within country-rounds vary between 1 and 28 percent. ICCs are also part of general data quality reports of the ESS (Ghirelli et al., 2022; Wuyts & Loosveldt, 2019) and show substantial heterogeneity across participating countries. Similarly, Zins and Burgard (2020) disentangled interviewer and design effects for round 6 of the ESS. Except for Bosnia and Herzegovina, all countries participating in the OeNB Euro Survey participated at least once in the ESS (Albania, North Macedonia, and Romania participated only once). Overall, they belong to the countries producing the largest interviewer variance estimates (see Ghirelli et al., 2022, p. 88), making a systematic analysis of these countries even more worthwhile. Moreover, the OeNB Euro Survey contains financial literacy questions that quiz respondents about economic concepts such as inflation, interest rates, or risk diversification (Lusardi & Mitchell, 2011). In contrast to other survey questions that ask for the respondent's opinions or attitudes, interviewers likely know the correct response to financial literacy questions and thus have additional capacity for intentionally influencing respondents, though the extent to which interviewers do so might differ. Using data from the German Panel on Household Finances, Crossley et al. (2021) document substantially higher interviewer variance for financial literacy questions than other questionnaire items. Interviewer effects on knowledge questions have also been documented for political knowledge (Johann & Mayer, 2021) and HIV transmission rates (Kerwin & Ordaz Reynoso, 2021). Thus, in line with Crossley et al. (2021), we expect higher interviewer variance for financial literacy questions than for other survey questions.

The fourth indicator is *satisficing*, which describes providing a satisfactory rather than optimal response to a survey item (Krosnick, 1991). Interviewers may engage in interviewing styles that induce satisficing. Interviewer variance in satisficing indicators may also result from interviewers sampling or recruiting similar types of respondents with regard to satisficing behavior (i.e., if interviewers differ in their ability to recruit reluctant respondents who at the same time show higher levels of satisficing). As the first measure of satisficing, we use straightlining, which states whether the responses to a set of same-scaled adjacent items are the same or vary across items (Yan, 2008). Straightlining is a widely used indicator of data quality and has been applied to assess both respondent satisficing and interviewer variance or falsification (e.g., Kim et al., 2019; Krosnick, 1991; Loosveldt & Beullens, 2017; Olbrich et al., 2024). Lower data quality is indicated by a lack of, or low, variance of responses across same-scaled items. Concerning substantive results, straightlining inflates the correlations between items in the same battery and thus conceals true differences between them (Yan, 2008). Loosveldt and Beullens (2017) investigated straightlining (or nondifferentiation) in the 6th round of the ESS and found that interviewers account for up to 20 percent of the variance in their straightlining indicators. However, they observe substantial heterogeneity across countries. Lastly, Vandenplas et al. (2018) jointly considered interviewer variance in straightlining and the duration of the module containing the respective item batteries. Using data from the 7th round of the ESS, they find that interviewers explain 8 to 38 percent of the variation in the module duration and 0 to 21 percent of the variation in straightlining. Blasius and Thiessen (2021) also used ESS data to identify suspicious interviewers using Categorical Principal Components Analysis (CatPCA) and straightlining. They identify several countries and interviewers with anomalous response patterns and find a strong correlation between their data quality measures and a corruption index. The second measure of satisficing is item nonresponse, which refers to the prevalence of "Don't know" and "No response" answers and is often used as an indicator for respondent satisficing (Krosnick, 1991). However, multiple studies provide evidence of the impact of interviewers on item nonresponse (e.g., Pickery & Loosveldt, 1998, 2001; Silber et al., 2021). Previous literature also used item nonresponse to identify fraudulent interviewers (Schäfer et al., 2005). Falsifiers might either avoid item nonresponse to avoid raising suspicion (Schäfer et al., 2005) or produce a lot of item nonresponse to reduce effort (Crespi, 1945). Hence, both extremely high and low shares of item nonresponse are suspicious. In cross-country surveys, interviewer effects on item nonresponse have received less attention than straightlining, although Bittmann (2020) reports sizeable ICCs for the ESS (note that he focuses on interviewer-respondent matching rather than interviewer errors). For the OeNB Euro Survey, item nonresponse is particularly relevant as the questionnaire contains multiple rather sensitive questions (i.e., on the respondent's debt or financial assets), which can induce item nonresponse (Grönemann, 2024).

The fifth indicator is a *(near-)duplicate analysis* (Kuriakose & Robbins, 2016). (Near-)duplicates may arise from interviewer behavior during the interview (e.g., repetitively fabricating parts of or entire interviews, strongly influencing respondents to respond in a particular way) or during data processing (e.g., ex-post filling in missing values). However, most studies investigating (near-)duplicates focused on the role of higher-level employees and fieldwork institutes and found evidence for manipulations in several cross-country surveys (e.g., Blasius & Thiessen, 2015; Koczela et al., 2015; Kuriakose & Robbins, 2016; Slomczynski et al., 2017). Accumulations of nearly identical interviews might indicate copy-pasting parts of, or entire, interviews by higher-level employees, though highly-similar interviews could occur for reasons unrelated to data quality (Simmons et al., 2016). Note that high similarities within an interviewer's workload indicate interviewer errors, whereas similarities across interviewers may occur due to collaboration between interviewers (Bergmann et al., 2019; Yamamoto & Lennon, 2018) or higher-level employees (Blasius & Thiessen, 2015). Sarracino and Mikucka (2017) showed that even small proportions of duplicates can lead to biased estimated regression coefficients and their standard errors.

These indicators investigate the role of interviewers from various perspectives, and the corresponding studies (with a particular focus on the ESS) document substantial heterogeneity across countries. In the present study, we combine and extend these approaches and use them as inputs to multivariate analysis methods. A priori, it is unknown which countries in the OeNB Euro Survey are more or less prone to interviewer errors. Concerning their development over time, errors might get worse as error-prone habits are acquired or errors are reduced as routines of preventing errors are established.

6.3 The OeNB Euro Survey

We use data from ten rounds (2012-2021) of the OeNB Euro Survey commissioned by the Austrian Central Bank. The survey covers Central, Eastern and Southeastern European countries that do not use the euro as a legal tender: six EU member states (Bulgaria, Croatia, Czech Republic, Hungary, Poland, and Romania[1]) and four candidate countries (Albania, Bosnia and Herzegovina, North Macedonia, and Serbia). The OeNB Euro Survey has been conducted regularly since 2007 as a face-to-face survey. The target population of the OeNB Euro Survey is defined as all persons aged 18 and over residing in the territory of the countries covered by the survey at the time of data collection.

In each country, an Austrian survey organization subcontracts fieldwork institutes to conduct the survey. In each country and each survey round, a sample of around 1,000 individuals is interviewed. National surveys are conducted by random route sampling of the adult population. For most countries and rounds, surveys are conducted using computer-assisted personal interviewing (CAPI). Especially in earlier rounds, a share of interviews are conducted via pen-and-paper-assisted personal interviewing (PAPI). Since 2015, only the Czech Republic, Hungary, and Poland have conducted some PAPI interviews. Hungary changed to 100 percent CAPI in 2018. Fieldwork is conducted in October and November and takes, on average, four weeks. The number of interviewers conducting the survey ranges from less than 30 to more than 100, with an average of 70 interviewers (see Figure 6.A1 in the online supplementary material).

Nonresponse varies across countries and survey rounds. AAPOR RR1 response rates (AAPOR, 2016) are reported in Table 6.B1 in the online supplementary material. Note that these response rates are based on reported disposition codes and are thus subject to reporting errors. The response rates vary between 10 and more than 80 percent. Within countries, we observe substantial changes in nonresponse often coinciding with changes in the fieldwork institute.

[1] Slovakia was included in the survey until 2008.

The survey uses a common questionnaire for all countries, which consists of core questions on euroization, trust, expectations, and related financial decisions that are repeated in each round, and flexible special topic modules. For each round, the final English questionnaire is translated into the national languages of the countries covered by the OeNB Euro Survey.

The OeNB Euro Survey was initially intended to run for three years only and has since evolved into a long-term survey project. Therefore, some methodological changes and data quality controls were introduced and developed over time. For example, information on the duration of interviews and interviewer IDs have only been collected since 2012, and further information on interviewers' survey experience has been collected since 2017.

6.4 Methods

In the forthcoming analysis, we evaluate data quality at the country-year (i.e., at the survey level) and interviewer level. The latter analysis is conducted year-by-year as the interviewer staff changed over time. The subsequent sections describe the indicators used, and their respective aggregation to the country-year and interviewer level.

6.4.1 Internal unit nonresponse bias indicator

We calculate the proportion of female respondents in gender-heterogeneous couples living in the same two-person household. A key challenge for this indicator is the identification of gender-heterogeneous couples living in the same household, which is particularly problematic for repeated cross-sectional surveys if questions on relevant items change over time. In our case, we cannot differentiate between gender-heterogeneous and gender-homogeneous couples. However, Rybak (2023) showed that this problem has a minor influence on the final measure. Furthermore, before 2018, we do not know whether couples live in the same household. For the data from 2018 onward, we know that at most three percent of respondents per round are married or have a partner but live in separate households. Thus, including couples who live in separate households should have negligible consequences. To ensure that sampling error does not influence the results, we follow Eckman and Koch (2019) and divide the difference between the proportion and 50 percent by the standard error ($\sqrt{50 \times 50/n}$). Note that the described internal measure is only a proxy to unit nonresponse bias in the respective sample as nonresponse bias might differ across variables and other sample subsets (Kohler, 2007).

Country-year level. The unit nonresponse bias measure is only available on the country-year level and we use the measure suggested by Eckman and Koch (2019) as the indicator.

6.4.2 Daily interviews per interviewer

For each country-year, we calculate the daily number of interviews per interviewers.

Country-year level. We use the maximum number of daily interviews per interviewer within the respective country-year as an indicator.

Interviewer level. At the interviewer level, we rely on the maximum number of daily interviews per interviewer.

6.4.3 Interviewer variance

The key variables for our interviewer variance analysis are financial literacy questions (Lusardi & Mitchell, 2008; Reiter & Beckmann, 2020). The OeNB Euro Survey contains four financial literacy questions[2] on inflation, interest rates, risk diversification, and exchange rates (see Table 6.1); the question on risk diversification was not included in 2017 and 2020. We also generate a financial literacy score that is the sum of correct responses (without the risk diversification question to ensure comparability across years). Following previous literature (Crossley et al., 2021), we code wrong answers, "Don't know", and "No response" answers as zero and correct answers as one. To put the interviewer variance estimates for the financial literacy questions into perspective, we estimate the interviewer variance for further questionnaire items, though we exclude several variables (socio-demographic variables, variables that could be over-filtered, extremely unbalanced binary variables, i.e., at least 80 percent have the same value, and variables with more than 15 percent item nonresponse). On average, we estimate the interviewer variance for 50 variables in each country-year.

The established approach to estimate interviewer variance is multilevel modeling (e.g., Brunton-Smith et al., 2017; Davis & Scott, 1995; Hox, 1994; Hox et al., 1991; O'Muircheartaigh & Campanelli, 1998; Schnell & Kreuter, 2005; Sturgis et al., 2021). The model for a continuous survey measure y is

$$y_{ijk} = \beta_0 + \sum_m \beta_m x_m + \theta_j + \mu_k + \varepsilon_{ijk} \tag{6.1}$$

where y_{ijk} is observed for each respondent i nested in interviewer j and PSU k, β_0 is a constant, β_m are the coefficients for control variables x_m, θ_j is each interviewer's effect on y and assumed to follow a normal distribution with mean zero and variance σ_θ^2. Similarly, the PSU effects μ_k are assumed to follow a normal distribution with mean zero and variance σ_μ^2. The residuals ε_{ijk} follow a normal distribution with mean zero and variance σ_ε^2. The ICC is calculated as $\sigma_\theta^2(\sigma_\theta^2 + \sigma_\mu^2 + \sigma_\varepsilon^2)^{-1}$ and denotes the proportion of variance explained by interviewers. The ICC can be further used to estimate the variance inflation caused by interviewers using the *deff* $(= 1 + ICC(b-1))$ or *deft* $(= \sqrt{deff})$ where b denotes the (average) interviewer workload (Kish, 1962; Schnell & Kreuter, 2005). In our data, respondents were not randomly assigned to interviewers, and thus, variance in the outcome across interviewers may reflect differences in their sample assignments (Elliott et al., 2022). We add multiple control variables (age, gender, education, employment, household size, town size, nightlight activity, dwelling characteristics, and household income quintiles) that should adjust for respondent composition differences across interviewers (Elliott et al., 2022; Hox, 1994). The online supplementary material provides a detailed description of the control variables in Table 6.C1. Interviewer characteristics (age, gender, experience) are not available for all years. Thus, we will only briefly discuss their significance in the online supplementary materials 6.D. Furthermore, the partial interpenetration of PSUs and interviewers is not sufficient to disentangle their effects in most country-years. Due to the random route sampling approach, interviews in most PSUs are conducted by a single interviewer (average number of interviewers per PSUs across all country-years: 1.08) and interviewers often only work in very few PSUs (average number of PSUs per interviewer across all country years: 1.98). Thus, we use the GPS coordinates of the

[2] A fifth question on legal obligations as a guarantor was included in 2018 and 2019. As this does not allow for assessing developments over time, we refrain from detailed discussions of this question.

PSUs and iteratively merge PSUs where only one interviewer worked to the closest PSUs until at least two interviewers worked in each (aggregate) PSU (see online supplementary material 6.A for more information).

For binary outcome variables we fit multilevel logistic regression models denoted as:

$$log\left\{\frac{P(y_{ijk}=1)}{P(y_{ijk}=0)}\right\} = \beta_0 + \sum_m \beta_m x_m + \theta_j + \mu_k \tag{6.2}$$

As before, y_{ijk} is the outcome variable, β_0 is a constant, β_m are the coefficients for control variables x_m, θ_j are the random interviewer intercepts with mean zero and variance σ^2_θ, and μ_k denote the PSU effects with mean zero and variance σ^2_μ. Assuming an underlying logistic distribution for the residuals, the ICC is calculated as $\sigma^2_\theta(\sigma^2_\theta + \sigma^2\mu + \frac{\pi^2}{3})^{-1}$. We fit the multilevel linear models using the R package `lme4` (Bates et al., 2015) and restricted maximum likelihood estimation and the multilevel logistic models using the R package `glmmTMB` (Brooks et al., 2017), which implements Laplace approximation. For ordinal outcomes, we fit multilevel ordinal logistic regressions using the `ordinal` package (Christensen, 2022). Note, however, that we use multilevel linear models for the financial literacy score to obtain results comparable to previous literature (Crossley et al., 2021). We refrain from estimating a single model with countries and years as higher levels since the control variables might affect the outcomes differently across countries and are not fully comparable across countries (e.g., the education level).

Table 6.1: Financial literacy questions in the OeNB Euro Survey.

Topic	Question
Inflation	Suppose that the interest rate on your savings account was 4% per year and inflation was 5% per year. Again disregarding any bank fees – after 1 year, would you be able to buy more than, exactly the same as, or **less than today** with the money in this account?
Interest rate	Next, we would like to ask some general questions concerning saving and borrowing. Suppose you had 100 [LOCAL CURRENCY] in a savings account and the interest rate was 2% per year. Disregarding any bank fees, how much do you think you would have in the account after 5 years if you left the money to grow: **more than 102**, exactly 102, less than 102 [LOCAL CURRENCY]?
Risk diversification	When an investor spreads his money among different assets, does the risk of losing money ...- increase - **decrease** - stay the same?
Exchange rates	Suppose that you have taken a loan in EURO. Then the exchange rate of the [LOCAL CURRENCY] depreciates against the EURO. How does this change the amount of local currency you need to make your loan installments? The amount of local currency ...- **increases** - stays exactly the same - decreases

Note: Correct responses in bold.

Country-year level. For the country-year analysis, we derive two indicators from the interviewer variance analysis. First, we calculate the average ICC for financial literacy questions. Second, we calculate the average ICC for all other items. To ensure comparability, we only include the items for which all countries fulfill our criteria concerning extreme imbalance and item nonresponse in the respective year.

Interviewer level. To identify anomalous interviewers, we predict interviewer effects from the estimated models. Then, we standardize these predictions for each variable and country-year and calculate the mean squared values separately for the financial literacy questions and all other questions for which interviewer variance was estimated. Again, we only use questions included in all countries in the respective year.

6.4.4 Satisficing

Item nonresponse

We measure item nonresponse by calculating the proportion of "Don't know" and "No response" answers for each interview.

Country-year level. To ensure that questionnaire characteristics do not influence differences across years, we pool the data from all countries and years and fit a linear regression with the item nonresponse as a dependent variable and the survey year as the sole explanatory variable. The country-year level indicator is the country-year average of the residuals. Furthermore, we dichotomize the proportion of item nonresponse ($D = 1\, if\, item\, nonresponse > 0.05, 0\, otherwise$) and fit separate multilevel logistic regressions as denoted in Equation 6.2. We use the estimated ICCs as a separate indicator based on item nonresponse.

Interviewer level. Based on the model described above, we use the standardized predicted interviewer effects as the interviewer-level indicators.

Straightlining

We use an item battery on trust in institutions and restrict the straightlining indicator to the items on trust in the government, the police, domestic banks, foreign banks, and the EU. Interviews with item nonresponse to at least two of these items are excluded to ensure that item nonresponse does not influence the results. Note that item nonresponse itself is analyzed with the indicator described in the previous section.

Country-year level. We follow the same procedure for straightlining as for item nonresponse. Here, the dependent variable is whether there is any variation within the trust item battery or not and we calculate the country-year level average of the residuals. We also fit a separate multilevel logistic regression for each country-year with the binary straightlining variable as the dependent variable and use the estimated ICCs as the indicator.

Interviewer level. On the interviewer level, we extract the interviewer-level predictions based on the models described above.

6.4.5 (Near-)duplicate analysis

To identify (nearly) identical survey records, we follow Kuriakose and Robbins (2016). This approach requires calculating the proportion of identical responses between each observation and every other observation in the dataset and obtaining the maximum similarity for each observation. A key challenge for the identification of (near-)duplicates is an appropriate threshold of similarity between two interviews that is unlikely to occur in the absence of data manipulation. Kuriakose and Robbins (2016) conducted multiple simulation analyses and suggested using 85 percent as a threshold since maximum similarities in their simulations never exceeded this value. Simmons et al. (2016) critically evaluated the 85 percent threshold and found that more observations, fewer variables, and more response options increase the probability of obtaining high similarities.

We refrain from using a fixed threshold for each country-year and use a mixture modeling approach to identify interviews with high similarities. For each country-year, we calculate the maximum similarities and fit mixture models with up to three clusters to the maximum similarities distribution. Based on BIC comparisons, the mixture model with the best fit is selected. If the one-cluster model has the best fit, no interviews are flagged for high similarities. If the two- or three-cluster solution is selected, we flag interviews who belong to the cluster with the highest average maximum similarity with more than 90 percent posterior probability to ensure that interviews are flagged with sufficient certainty. As a further condition, the proportion of flagged interviews must not exceed 50 percent of the sample size to avoid "normal" interviews being flagged in cluster solutions with a cluster of low maximum similarities and average maximum similarities. For calculating the maximum similarities, we follow Kuriakose and Robbins (2016) and exclude variables with more than 10 percent missings to ensure that filtering patterns do not drive our results.

Country-year level. For each country-year, we calculate the proportion of interviews flagged by the mixture modeling approach.

Interviewer level. Similarly, we calculate the proportion of interviews flagged by the mixture modeling approach. While the likelihood of higher similarities increases with the interviewer's workload, flagged interviews indicate suspicious behavior irrespective of the number of interviews.

6.4.6 Isolation forests for outlier detection

While all of the indicators can work as standalone measures to flag suspicious interviewers or country-years, we seek to combine all the measures to identify the most exceptional cases. Previous methods used to aggregate indicators and identify falsifying interviewers include summing up z-scores of indicators (Schwanhäuser et al., 2022) or cluster analysis (De Haas & Winker, 2014, 2016). For the former approach, a case with an exceptional value for only one indicator might not be flagged because of inconspicuous values for all other indicators. For cluster analysis, a disadvantage is that we do not know which of the resulting clusters is suspicious nor do we know why cases were assigned to specific clusters.

We rely on a tree-based outlier detection method called Isolation Forests (Liu et al., 2008). To build a single isolation tree, the algorithm randomly selects an indicator x, samples a value from $unif[min(x), max(x)]$, and splits the sample by this value. These steps are repeated until every observation ends up in a singular tree branch. Outliers likely require fewer splits until they are *isolated*, whereas more common observations will require more splits as they are closer to other observations. The isolation depth (i.e., the number of splits until an observation is isolated) describes to which extent an observation is an outlier. Based on the isolation depth, Liu et al. (2008) derived a standardized outlier score that simplifies interpretation. Scores above 0.5 indicate that the observation is an outlier. Combining many isolation trees results in an isolation forest that allows for calculating average outlier scores and ensures that results are not determined by a single set of draws. For a more in-depth description of isolation forests, see Liu et al. (2008). To fit the isolation forests, we use the R package `isotree` (Cortes, 2023) and use 1,000 trees for each model.

While the identification of outliers itself can provide valuable insights on data quality problems, information on *why* a particular observation is flagged can guide further in-depth investigations. Therefore, we calculate Shapley values for each observation (Štrumbelj & Kononenko, 2014). Shapley values are based on a concept from game theory (Shapley, 1953) and provide each feature's contribution to the difference between the observed prediction (or score) and the average prediction (or score) for the entire sample. This informs us about the indicators that cause the respective case's outlier score. An in-depth discussion on Shapley values is provided in Molnar (2022). We use the R package `fastshap` (Greenwell, 2021) to calculate the Shapley values.

We fit isolation forests on the country-year level to determine the most exceptional OeNB Euro Survey samples and on the interviewer level for each year to identify the most suspicious interviewers. For each analysis, we also calculate Shapley values to enhance our understanding of exceptional cases.

6.5 Results

6.5.1 Internal unit nonresponse bias indicator

Replicating Kohler (2007), the proportion of female respondents for each country in each year is shown in Figure 6.2. In total, 32 country-years statistically significantly deviate from 50 percent. The deviations are most frequent for Albania (6) and Croatia (8), whereas countries such as the Czech Republic, Poland, Bulgaria, or Serbia are almost always close to 50 percent. In Albania, the proportion is around 55 percent from 2012 to 2015, but in 2016 the proportion drops to 40 to 45 percent. Notably, this change was concurrent with a change in the fieldwork institute. In Croatia, the proportions increase up to 68 percent in 2021, indicating an increase in bias. Bosnia and Herzegovina and North Macedonia show extreme deviations for a subset of subsequent years.

The estimated proportions are broadly in line with previous estimates for other cross-national surveys, such as the ESS. Nonetheless, the patterns observed for Albania, Bosnia and Herzegovina, North Macedonia, and Croatia suggest problems during sampling and/or recruitment. Given that interviewers play a critical role in random route surveys (interviewers are responsible for sampling, recruitment, and measurement), they are likely the main drivers behind these deviations from 50 percent (Kohler, 2007). Furthermore, the estimates point to the importance of the fieldwork institutes as switching fieldwork institutes can lead to substantial changes in sample composition.

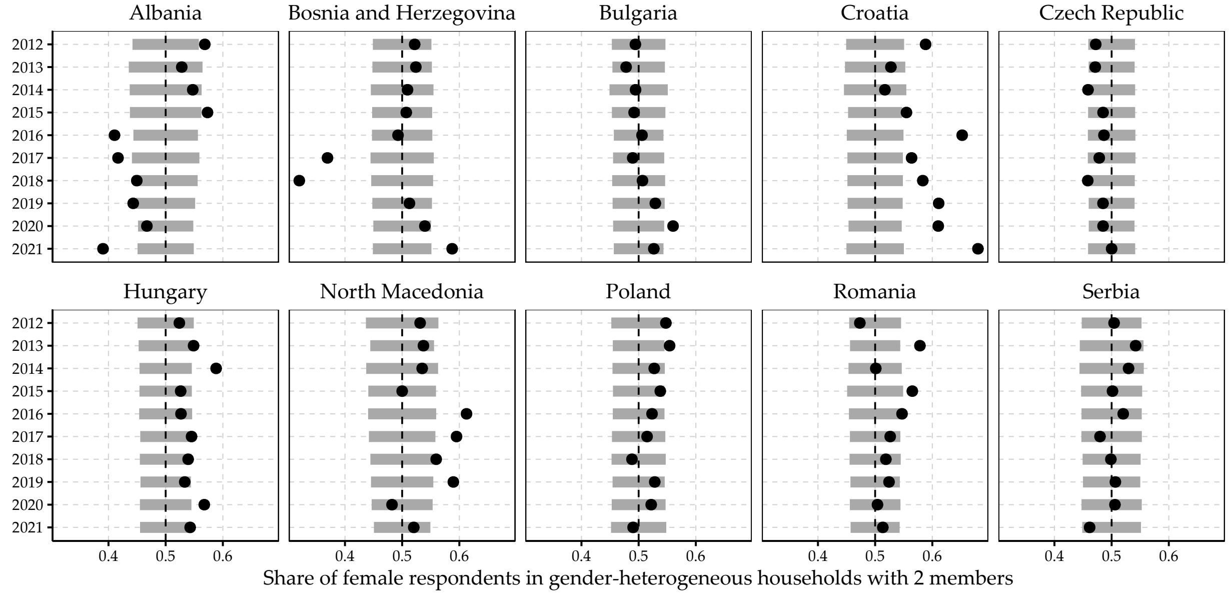


Figure 6.2: Share of female respondents in gender-heterogeneous households with two members across countries and years with 95 percent confidence interval.

6.5.2 Interviewer workload

The distribution of the daily number of successful interviews per interviewer varies substantially over time and countries (see Figure 6.E1 in the online supplementary materials). For example, the average number of daily successful interviews was 9.4 in Bulgaria in 2013 and decreased to 3.9 in 2021. On the contrary, the average never exceeds 4.5 in Croatia. Generally, the average number of daily interviews per interviewer decreased over time. In Bulgaria in 2013, some interviewers had implausibly high workloads of more than 40 completed interviews per day. Such deviations might arise from technical problems, interviewers sharing an interviewer ID, or fraudulent behavior. As the most extreme cases occurred more than ten years ago, closer investigations of these cases are unfortunately not possible. Regardless of the reason, these extreme workloads indicate deviations from survey protocols.

6.5.3 Interviewer variance

Figure 6.3 depicts the estimated ICCs for the financial literacy score (with 90 percent confidence intervals based on 200 bootstrap replications).[3] The estimated ICCs for all financial literacy questions are provided in Tables 6.F1 to 6.F5 in the supplementary materials. The grey dots depict the ICCs for other questionnaire items. Only items for which ICCs were estimated every year for the respective country are included to ensure adequate comparisons over time (ranging from 11 to 18 items).[4] We observe highly heterogeneous ICCs across countries, and years. In general, the

[3] For several country-years, multiple bootstrap replications fail due to the absence of aggregated PSU variance. We show the interviewer variance results for all replications (including the failed replications). Excluding failed replications changes the confidence intervals only marginally.

[4] Figure 6.G1 in the online supplementary material depicts boxplots of the ICC estimate for each country-year to ease the observation of trends and outliers.

ICCs – in particular for financial literacy – are high. Using results from the ESS as comparison (Ghirelli et al., 2022; Wuyts & Loosveldt, 2019), the ICCs are similar to the ESS countries with the highest ICCs. The consequences of such high ICCs are discussed in Section 6.6. Poland is the only country in which we observe a steady decrease in ICCs over time. In Croatia in 2013, the ICCs are substantially lower than in any other country and year for all variables.[5]

Compared to other questionnaire items, the financial literacy items are most prone to interviewer variability. Across the 100 country-years, the interest rate question has the highest ICC in 46 cases, the exchange rate question in 16 cases, the inflation question in 13 cases, and the risk question in 1 case (summing up to 76 total cases). In line with Crossley et al. (2021), we find that ICCs are highest for the inflation rate question and lowest for the risk diversification question for several countries. Crossley et al. (2021) estimated ICCs of 0.290 for the inflation question, 0.386 for the interest rate question, and 0.183 for the risk diversification question (multilevel logistic models with controls). For the literacy score, they estimated an ICC of 0.170. Compared to Crossley et al. (2021), the ICCs for the inflation question are statistically significantly higher for 31 country-years, 28 country-years for the interest rate question, and 20 out of 80 country-years for the risk diversification question (see Tables 6.F1 to 6.F5); note, however, that the bootstrapped confidence intervals are relatively wide due to the small number of observations in each country-year.[6]

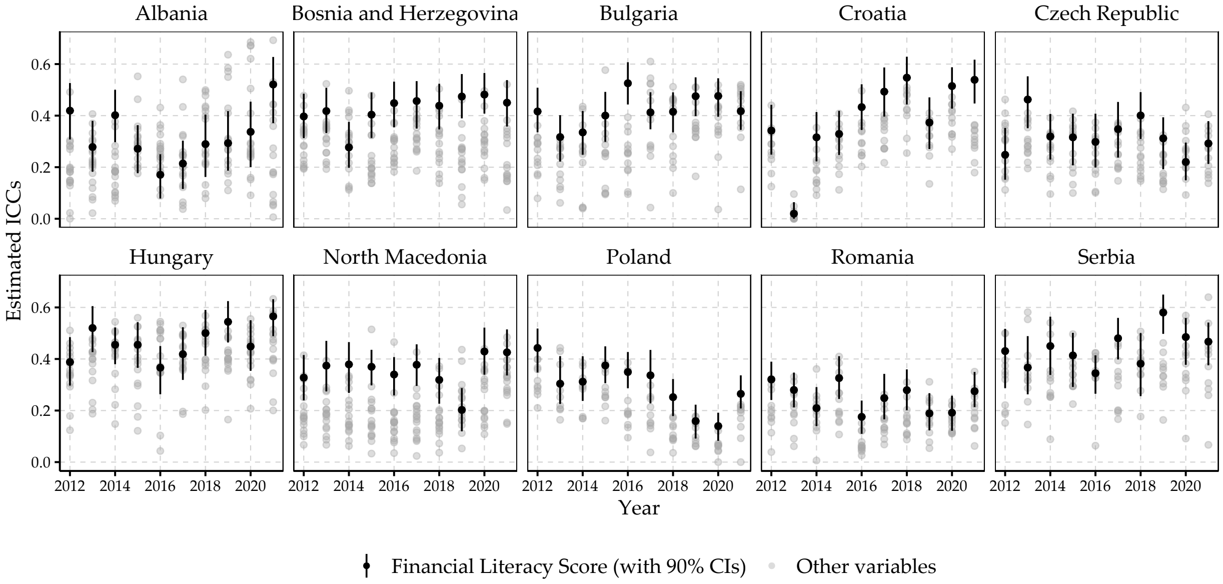


Figure 6.3: Estimated ICCs for financial literacy and other variables across countries and years.

[5] In the fall of 2013, the institute conducting the survey in Croatia changed the interviewer team from full-time employees to part-time contractual workers, most of whom were very young and inexperienced. The data collected by this team was of very poor quality with regard to all indicators analyzed at that time by the OeNB. The subcontract with the institute was subsequently terminated and in spring of 2014, a new fieldwork institute repeated the survey that had been conducted in Croatia in the fall of 2013. The low ICCs for Croatia in 2013 are from the 2014 spring survey conducted by the new institute. Again, these results point toward the importance of the fieldwork institute when investigating interviewer errors.

[6] As we coded "Don't know" and "No response" values as wrong answers, we also tested whether excluding these observations leads to different results. While the overall developments remain unchanged, the ICCs are on average slightly higher across all financial literacy questions (inflation: 5.7 percentage points higher; interest rate: 6.2 percentage points higher; exchange rate: 5.3 percentage points higher; risk: 2.1 percentage points higher).

Concerning the role of regional homogeneity, Figure 6.H1 (in the online supplementary materials) shows that interviewer variance plays a more important role than the aggregated PSU variance for the vast majority of items. Similarly, fitting models with or without the aggregate PSU random effects (see Figure 6.H2) does not substantially change the interviewer variance estimates and conclusions drawn from the results (in 17.9 percent of all estimations, the ICC changes by more than five percentage points). In line with these findings, the results of an alternative analysis approach to disentangle PSU from interviewer variance that exploits the longitudinal structure of the data and the prevalence of changes of the fieldwork institute in several countries (described in detail in the supplementary materials 6.I) show that interviewer variance is substantially larger than PSU variance.

6.5.4 Satisficing

Item nonresponse

The proportion of "Don't know" and "No response" responses varies substantially both within and across country-years. The averages vary between 3 to 8 percent in most countries (see 6.J1 in the online supplementary material). In some countries and years, the proportions exceed 10 percent, but these values are mostly driven by a few outlying cases with extreme item nonresponse values. A particularly noteworthy change occurred in Albania, where the average proportion of item nonresponse dropped from 5.8 percent in 2020 to less than 0.9 percent in 2021. Over time, item nonresponse decreased in the OeNB Euro Survey.

The ICC estimates for item nonresponse are reported in Figure 6.J3 of the online supplementary materials. With some exceptions, the estimated ICCs vary between 20 and 60 percent. These consistently high values indicate substantial interviewer variance and thus show that interviewers strongly influence the prevalence of more than 5 percent item nonresponse. Within countries, the estimated ICCs are rather stable with slight decreases in some countries, while countries such as Albania, Croatia, Romania, and Serbia have higher variation in the estimated ICCs. The ICCs above 50 percent are driven by a very uneven distribution of item nonresponse across interviewers where most interviewers either rarely or almost always have item nonresponse shares above 5 percent. In Croatia, we observe the same pattern as for the ICCs estimated for the questionnaire items (see section 6.5.3).

Straightlining

In many countries, the proportion of respondents who selected the same response to the trust items varies around 20 percent over time (see Figure 6.J2 in the online supplementary material). In Croatia and Hungary, the share increased since 2017, whereas it slightly decreased in Bosnia and Herzegovina and Serbia. As for other indicators, Albania is a clear outlier, both with regard to the variation over time and the level of straightlining. From 2018 to 2019 straightlining decreased from 48 to 20 percent, increased again to 48 percent in 2020, and decreased to 24 percent in 2021. These results provide further evidence of potential irregularities during data collection in Albania during the most recent years of the survey.

In most countries, the ICC estimates for straightlining vary between 20 and 60 percent and none of the countries consistently has ICCs below 20 percent (see Figure 6.J3), though Poland steadily improves over time. In Croatia, we observe the same pattern as for the interviewer variance in financial literacy and other items and item nonresponse. In general, the high ICCs indicate that straightlining is to large parts driven by interviewers in many country-years.

6.5.5 (Near-)duplicate analysis

For the (near-)duplicate analysis, Table 6.2 reports the proportion of interviews flagged by the mixture modeling approach for each country-year. In total, eight country-years have more than 10 percent of flagged respondents, Albania accounts for six of these. The most exceptional country-year is Albania in 2020 with 46.8 percent of flagged observations, which is also an outlier when the 85%-threshold is considered.

Table 6.2: Proportion of observations flagged by mixture modeling approach (proportion above 0.85 in parentheses).

Country	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Albania	3.56	6.58	2.77	0.00	20.50	24.20	36.60	18.20	46.76	25.87
	(5.29)	(8.02)	(2.19)	(0.18)	(6.80)	(1.60)	(14.30)	(4.20)	(44.17)	(19.58)
Bosnia and Herzegovina	6.30	4.36	1.09	2.45	0.00	0.00	0.00	2.80	0.00	0.00
	(0.39)	(0.00)	(0.20)	(0.59)	(0.50)	(0.78)	(0.88)	(4.60)	(4.90)	(1.70)
Bulgaria	8.31	1.36	8.96	0.00	0.00	0.00	10.90	0.00	0.00	1.20
	(4.30)	(1.36)	(0.00)	(0.29)	(0.39)	(0.79)	(0.60)	(1.10)	(4.78)	(3.99)
Croatia	14.00	0.00	0.00	0.00	0.00	3.17	0.00	0.00	0.00	2.27
	(11.70)	(0.20)	(0.00)	(1.50)	(4.09)	(0.59)	(0.00)	(0.48)	(2.36)	(1.38)
Czech Republic	4.46	6.00	2.45	2.74	1.00	2.70	8.30	2.40	0.00	5.30
	(3.32)	(1.24)	(0.28)	(3.30)	(2.20)	(0.00)	(1.70)	(2.10)	(9.70)	(6.30)
Hungary	2.70	6.50	1.89	9.47	0.00	2.60	5.90	0.00	0.00	3.50
	(2.40)	(0.60)	(1.79)	(0.00)	(2.90)	(1.50)	(5.40)	(2.90)	(7.20)	(4.10)
North Macedonia	3.24	3.39	1.08	0.00	0.00	0.70	0.78	0.00	0.00	1.19
	(0.00)	(0.00)	(0.39)	(0.20)	(0.00)	(0.20)	(0.20)	(1.19)	(3.56)	(0.60)
Poland	4.90	6.30	3.29	3.07	0.99	2.79	2.46	0.00	0.80	0.00
	(0.90)	(5.10)	(0.50)	(3.07)	(3.16)	(3.99)	(1.57)	(2.26)	(5.86)	(0.20)
Romania	4.78	1.31	1.71	1.15	3.19	0.57	4.35	4.52	0.58	4.91
	(0.83)	(0.53)	(0.85)	(1.15)	(2.59)	(0.57)	(3.56)	(4.52)	(4.87)	(5.68)
Serbia	6.34	3.07	8.83	0.00	3.99	0.00	0.00	0.00	0.00	5.95
	(6.72)	(3.07)	(8.83)	(1.11)	(1.60)	(0.30)	(0.00)	(2.48)	(4.17)	(2.48)

Due to the exceptional values observed for Albania in 2020, we investigate these results more closely. First, we analyze whether high matches occur within or between interviewers. Using all pairs of interviews with maximum similarities equal to or larger than the minimum value flagged by the mixture modeling threshold, we find that only 1,293 of the 8,176 pairs (15.8 percent) with a maximum match flagged by the mixture modeling approach share the same interviewer. This indicates that interviewers are unlikely to be the main source of the high similarities.

Second, we evaluate whether interviews flagged by the mixture modeling approach are connected and build an adjacency matrix of all flagged interviews. We fill the $n \times n$ matrix with values x_{ij} where x_{ij} is 1 if interviews i and j have a similarity equal to or larger than the mixture modeling threshold and 0 otherwise. Figure 6.4 illustrates the connections between interviews for Albania in 2020 using Fruchterman-Reingold layouts (Fruchterman & Reingold, 1991). Serbia in 2014 and Croatia in 2012 are shown for reference. For the Serbian sample (90 interviews), the maximum component size is four with no larger components emerging (network density = 0.015). For the Croatian sample (140 interviews), we observe a large connected component accounting for 62.9 percent of all interviews, the remaining sample consists of small components (network density = 0.034). 11 interviewers conducted the interviews in the large component. For Albania in 2020

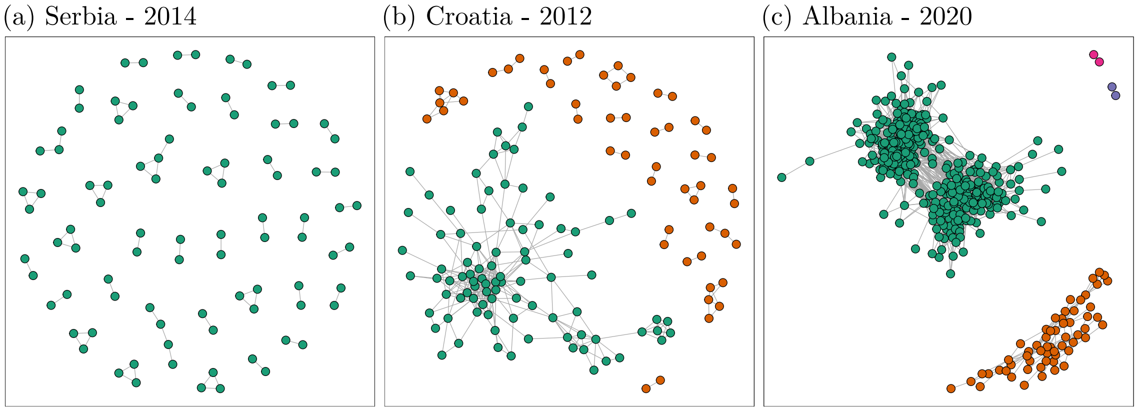


Figure 6.4: Fruchterman-Reingold layout plots. Each point corresponds to an interview that has at least one match exceeding the mixture modeling-based threshold. Grey lines connect interviews that are a match.

(469 interviews), the network (network density = 0.074) consists of four components of varying size (407, 58, 2, and 2 interviews). Two interviewers working in the same region account for all interviews in the second-largest component (orange dots in Figure 6.4c). 14 interviewers were involved in the largest component (green dots in Figure 6.4c). These interviews were conducted in northern regions of Albania with two supervisors being responsible for all except three interviews. Our results show that the high similarities did not occur by chance between pairs of interviews but represent networks of similar interviews. While regional homogeneities could lead to highly similar interviews, our results (see Table 6.2) show that high similarities only occurred from 2016 onwards, coinciding with the change to a new fieldwork institute. Regional homogeneities should be observed in preceding years as well and should not abruptly end at the supervisor's region of responsibility.

Lastly, we investigate the interview start date and time and the interview duration. In Albania in 2020, 24.3 percent of observations share the same interviewer start date and time and interview duration with at least one other observation. Given that these data should be captured automatically, such duplicates are highly unlikely and point toward manipulation of the data. Indeed, for other countries in 2020, the maximum proportion is just 3.3 percent. In earlier rounds (i.e., 2012-2014), however, non-unique timestamps were more common (for example, 38.8 percent in Albania in 2014). As interviews were partly conducted in PAPI mode in these years, such cases are likely driven by ex-post filling-in the respective data or a lack of guidelines for reporting the start time and duration. The non-unique timestamps in 2020 are exclusively between interviewers. In the most extreme cases, two interviewers working in different PSUs conducted up to nine interviews starting at the same time and taking the same time within one day. Combining the paradata analysis with the duplicate analysis, we find that in Albania in 2020 the proportion of interviews with a maximum match flagged by the mixture modeling analysis for observations with a non-unique interview date, time, and duration is 95.1 percent, while it is only 31.2 percent for interviews with unique paradata. For the earlier data, we do not observe such differences, which indicates that lack of reporting guidelines and ex-post filling-in could be the main reason for the non-unique data. In summary, the follow-up analyses indicate that the large proportions of nearly duplicated data in Albania are driven by data manipulations. It is more likely that these manipulations were carried out by supervisors rather than interviewers.

6.5.6 Isolation forests for outlier detection

The previous sections provide a detailed description of each data quality indicator. Here, we evaluate whether isolation forests can enhance the efficiency of detecting exceptional country-years and interviewers and may guide researchers during quality controls.

Country-year level analysis

The country-year level analysis is based on nine indicators. In total, nine country-years have scores above the outlier threshold of 0.5 (see Figure 6.K1 in the online supplementary material). The highest value belongs to Albania in 2020, followed by Croatia in 2013 and Albania in 2021. Albania accounts for five of the outliers, none of the other countries are outliers more than twice. With regard to the survey year, no pattern emerges as at most two outliers are from the same year.

While the isolation forest scores highlight country-years that require closer investigation, Shapley values can indicate which indicators are particularly noteworthy in the respective country-year. The Shapley values for the country-years flagged as outliers are shown in Figure 6.5. Higher values indicate that the respective indicator contributes more to the difference between the observed and the average score. Note that low values depict that the indicator values are either not anomalous or that other indicators are more exceptional. The outlying country-years differ widely with regard to the indicators with the largest contribution. In Albania, several indicators seem to play important roles, although the (near-)duplicate indicator is the most important in all years. In Bulgaria in 2013, the extreme maximum daily workload drives the outlier score. Croatia in 2013 is flagged due to the exceptionally low ICCs, which do not imply low data quality but call for closer investigation due to their extreme deviation from other countries and years. For Croatia in 2021 and Romania in 2016, several indicators contribute to the outlying score.

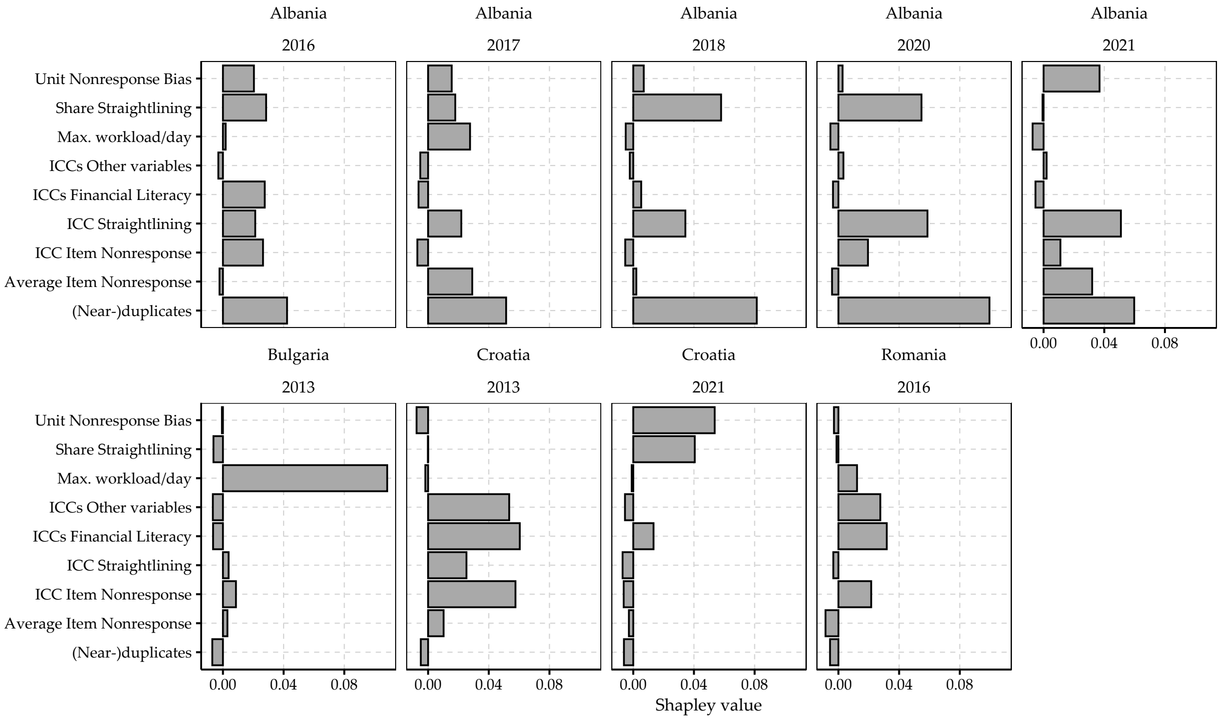


Figure 6.5: Shapley values for countries with scores above 0.5.

Interviewer-level analysis

For the interviewer-level analysis, we use six indicators and estimate an isolation forest for each year separately (see Figure 6.K2 in the online supplementary material for the distribution of the scores in each year). As reported in Table 6.3, the extent to which interviewers from specific countries have values above 0.5 varies substantially. For example, more than half of the interviewers working in Albania in 2020 are flagged. In the other countries, the proportion never exceeds 20 percent.

Table 6.3: Proportion of interviewers with score above 0.5 from isolation forest analysis.

Country	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Albania	5.26	11.90	9.30	2.78	33.33	43.33	45.16	45.16	53.57	27.59
Bosnia and Herzegovina	8.33	6.58	7.81	5.63	2.78	2.63	7.14	4.41	1.49	3.33
Bulgaria	11.48	13.79	17.46	2.70	2.50	4.76	8.65	2.91	2.00	5.94
Croatia	15.07	8.93	1.69	3.17	4.69	3.57	7.69	8.45	2.74	9.21
Czech Republic	6.35	5.77	6.12	10.71	5.56	5.45	9.80	12.00	2.04	14.29
North Macedonia	7.27	4.11	4.35	2.33	1.18	4.44	2.50	1.45	3.70	3.51
Hungary	4.17	7.14	4.81	13.13	4.35	7.77	5.05	5.26	2.00	7.29
Poland	8.86	11.54	6.32	7.22	3.16	11.96	10.64	8.51	9.47	7.78
Romania	8.62	2.73	6.42	9.09	10.67	9.43	10.71	10.96	7.79	8.60
Serbia	14.89	7.41	12.28	9.26	8.45	1.33	5.13	4.00	3.95	13.04

As before, the Shapley values provide more detailed insights into the indicators' contribution to the respective scores. As an example, Figure 6.6 shows the Shapley values for the four interviewers with the highest scores in 2020. Similar to the country-year level analysis, the most important indicators vary across interviewers. For some interviewers (i.e., 2, 3, 4) only single indicators drive the outlier scores, for other interviewers (i.e., 3) several indicators contribute to the outlier score.

In both cases, practitioners can use these insights to investigate the respective indicators and their sources more closely.

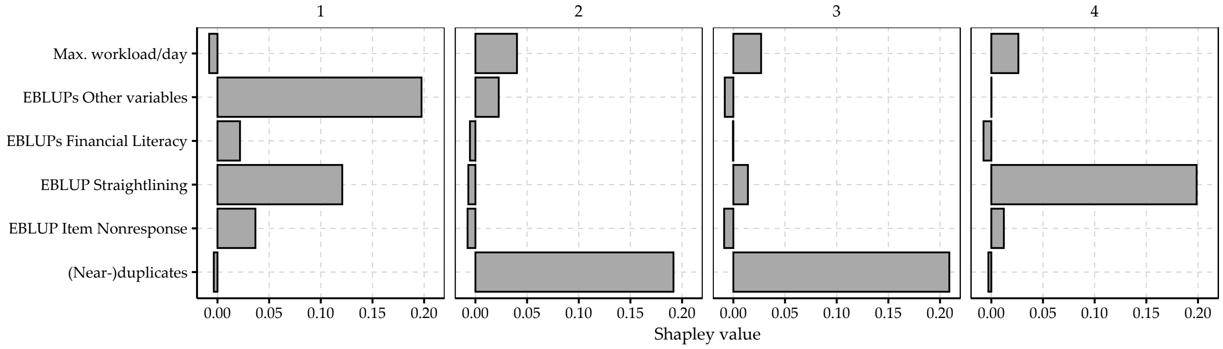


Figure 6.6: Shapley values for four interviewers with highest scores in 2020.

6.6 Impacts on substantive analyses

To explore the extent to which the described data quality issues can impact substantive analyses, we take two approaches. First, we focus on the (near-)duplicate analysis and Albania in 2020 to evaluate the extent to which financial literacy values differ between flagged and non-flagged observations. Second, we discuss how interviewer variance affects inference for the most recent OeNB Euro Survey round (2021).

Table 6.4 reports the proportion of correct responses for the three financial literacy questions and the financial literacy score for the observations flagged and not flagged by the mixture modeling threshold. For the interest rate question, the differences between both groups are negligible. In contrast, only 3.0 percent of the flagged observations provided a correct response to the inflation question. For the observations below the threshold, this share is 48.9 percent. For the exchange rate question, the difference between both groups is 9.5 percentage points. The financial literacy score differs by 0.35 points which is driven by the large difference for the inflation question. These results show that (presumably) manipulated data can severely bias substantive analyses.

Table 6.4: Differences in financial literacy questions for observations below or above simulated threshold, Albania 2020.

	Interest rate	Inflation	Exchange rate	Score	N
Not flagged	0.270 (0.019)	0.489 (0.022)	0.238 (0.018)	0.996 (0.042)	534
Flagged	0.281 (0.021)	0.030 (0.008)	0.333 (0.022)	0.644 (0.037)	469
Full	0.275 (0.014)	0.274 (0.014)	0.282 (0.014)	0.832 (0.029)	1003

Notes: Standard errors in parentheses.

Using the *deff* formula presented in section 6.4.3 and the data and results from 2021 for the financial literacy score, Table 6.5 reports the average interviewer workloads, the ICC, the *deff*, and the effective sample sizes. The sizeable variations in interviewer workloads and ICCs across countries lead to large differences in the *deff* across countries. Poland has the lowest *deff* of 3.68

which implies that the variance is 3.68 times higher than the variance from a simple random sample. Albania has high values for the average interviewer workload and the ICC which leads to an extreme design effect of 18.44 and an effective sample size of around 54. Hence, the interviewer variance leads to substantial variance inflation, complicating comparisons across countries or over time.

Table 6.5: Interviewer effects on financial literacy score in 2021.

Country	N	Int. workload	ICC	deff	Eff. sample size
Albania	1000	34.483	0.521	18.439	54.233
Bosnia and Herzegovina	979	16.317	0.450	7.896	123.985
Bulgaria	995	9.851	0.417	4.695	211.906
Croatia	1008	13.263	0.539	7.610	132.463
Czech Republic	1000	20.408	0.292	6.667	149.982
Hungary	996	10.375	0.565	6.300	158.105
North Macedonia	974	17.088	0.428	7.887	123.498
Poland	1000	11.111	0.265	3.676	272.020
Romania	1027	10.926	0.275	3.729	275.393
Serbia	1007	14.594	0.467	7.351	136.983

Notes: Interviewer workloads based on sample sizes for multilevel models.

6.7 Discussion

Interviewers play a key role in face-to-face surveys as their tasks include contacting respondents, convincing them to participate, and conducting the interviews. However, all these tasks are prone to errors. In this study, we investigated interviewer errors from various perspectives in ten rounds of a cross-national survey. We implemented several data quality indicators related to interviewer error (internal unit nonresponse bias measure, interviewer workloads, interviewer variance, satisficing, near-duplicates) and identified multiple country-years with suspicious values. Concerning interviewer variance, we found that financial literacy questions are particularly prone to interviewer variability (in line with Crossley et al., 2021). To facilitate the efficient identification of outlying cases, we combined the data quality indicators in an isolation forest analysis both on the country-year and interviewer level. Using Shapley values, we also illustrated an approach that can guide applied researchers to potential sources of data quality issues. Lastly, we showed that the described data quality issues can severely affect substantive analyses.

While multiple country-years have exceptional patterns for single indicators, Albania stands out across most analyses. This finding is emphasized in the isolation forest analysis where Albania is flagged in five out of ten years for the country-year level analysis and large shares of interviewers working in Albania are flagged in the interviewer-level analysis. These findings point to problems during data collection in Albania. As a consequence, subsets of the Albanian data have since been excluded from the OeNB Euro Survey. Further details on how OeNB addressed the data quality issues in Albania can be found at https://www.oenb.at/en/Monetary-Policy/Surveys/OeNB-Euro-Survey.html. Our follow-up analyses indicate that this might be a rather local issue related to supervisors. For other country-years, no sample is as suspicious as Albania, though several interviewers have suspicious values on one or multiple indicators. In sum, our analysis shows that interviewer errors can severely harm data quality and induce biased and imprecise survey estimates.

Using ten rounds of data collection allows for observing changes within countries over time. In repeated surveys, data quality might either increase due to a learning process or decrease due to acquiring habits harming quality. In our case, only Poland improved over time, while we observed no clear changes in other countries. Our results also highlight the importance of the contracted fieldwork institutes (Blasius & Sausen, 2023; Blasius & Thiessen, 2015, 2021). In various analyses, we observed substantial changes when the contracted fieldwork institute in the respective country changed. Some changes led to improved quality, while others resulted in a decline. Future research on data quality on cross-country surveys may investigate the role of fieldwork institutes (and interviewer supervisors) in more depth. Interviewer errors can only be reduced if interviewers are made aware of standardized guidelines, are thoroughly trained, are monitored and receive feedback on their work, and receive recommendations for improving their work (Fowler & Mangione, 1990; Groves et al., 2004). If these factors are absent or of low quality, interviewer errors are unavoidable and will lower data quality. Of course, reducing interviewer errors requires adequate behavior by supervisors and higher-level field work institute employees.

Our results are subject to several limitations. First, the interviewers are not randomly assigned to regions and respondents. However, we account for (aggregated) PSU effects and multiple control variables in the multilevel model analysis. Our results suggest that regional differences or respondent compositions play a minor role. Second, in most cases, data were collected years ago which prohibits follow-up investigations for suspicious cases. Third, in special cases such as Albania, we cannot conclusively identify the source of suspicious data, i.e., whether the interviewer, supervisors, or the fieldwork institute is primarily responsible.

The results emphasize the necessity of implementing thorough data quality controls for interviewer-administered surveys. In particular, when data are collected in multiple countries and researchers cannot observe the contracted fieldwork institute's work, quality controls should be conducted during or shortly after the field period has ended to ensure that potential problems can be corrected (such procedures have now been implemented for the OeNB Euro Survey).

Literature

AAPOR. (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys*. AAPOR.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bergmann, M., Schuller, K., & Malter, F. (2019). Preventing interview falsifications during fieldwork in the Survey of Health, Ageing and Retirement in Europe (SHARE). *Longitudinal and Life Course Studies*, *10*(4), 513–530.

Beullens, K., & Loosveldt, G. (2016). Interviewer effects in the European Social Survey. *Survey Research Methods*, *10*(2), 103–118.

Bittmann, F. (2020). The more similar, the better? *Survey Research Methods*, *14*(3), 301–323.

Blasius, J., & Sausen, L. (2023). Perceived corruption, trust, and interviewer behavior in 26 european countries. *Survey Research Methods*, *17*(2), 131–145.

Blasius, J., & Thiessen, V. (2015). Should we trust survey data? Assessing response simplification and data fabrication. *Social Science Research*, *52*, 479–493.

Blasius, J., & Thiessen, V. (2021). Perceived corruption, trust, and interviewer behavior in 26 European countries. *Sociological Methods & Research*, *50*(2), 740–777.

Brooks, M., E., Kristensen, K., Benthem, K. van, Magnusson, A., Berg, C., W., Nielsen, A., Skaug, H., J., Mächler, M., & Bolker, B., M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378.

Brunton-Smith, I., Sturgis, P., & Leckie, G. (2017). Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(2), 551–568.

Bushery, J. M., Reichert, J. W., Albright, K. A., & Rossiter, J. C. (1999). Using date and time stamps to detect interviewer falsification. *Proceedings of the Survey Research Method Section, American Statistical Association*, 316–320.

Christensen, R. H. B. (2022). *Ordinal-Regression models for ordinal data*. https://CRAN.R-project.org/package=ordinal. https://CRAN.R-project.org/package=ordinal

Cortes, D. (2023). *Isotree: Isolation-based outlier detection*. https://CRAN.R-project.org/package=isotree

Crespi, L. P. (1945). The cheater problem in polling. *Public Opinion Quarterly*, *9*(4), 431–445.

Crossley, T. F., Schmidt, T., Tzamourani, P., & Winter, J. K. (2021). Interviewer effects and the measurement of financial literacy. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(1), 150–178.

Davis, P., & Scott, A. (1995). The effect of interviewer variance on domain comparisons. *Survey Methodology*, *21*(2), 99–106.

De Haas, S., & Winker, P. (2014). Identification of partial falsifications in survey data. *Statistical Journal of the IAOS*, *30*(3), 271–281.

De Haas, S., & Winker, P. (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), 643–660.

Eckman, S., & Koch, A. (2019). Interviewer involvement in sample selection shapes the relationship between response rates and data quality. *Public Opinion Quarterly*, *83*(2), 313–337.

Elliott, M. R., West, B. T., Zhang, X., & Coffey, S. (2022). The anchoring method: Estimation of interviewer effects in the absence of interpenetrated sample assignment. *Survey Methodology*, *48*(1), 25–48.

Fowler, F., & Mangione, T. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. SAGE Publications, Inc.

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129–1164.

Ghirelli, N., Lynn, P., Dorer, B., Schwarz, H., Kappelhof, J., van de Maat, J., Kessler, G., Briceno-Rosas, R., & Rød, L.-M. (2022). *Quality report for the European Social Survey, Round 9*. GESIS.

Greenwell, B. (2021). *Fastshap: Fast approximate shapley values*. https://CRAN.R-project.org/package=fastshap

Grönemann, M. (2024). How to reduce item nonresponse in face- to-face surveys? A review and evidence from the European Social Survey. *methods, data, analyses*, 1–20. https://doi.org/10.12758/MDA.2024.02

Groves, R. 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, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. John Wiley & Sons, Inc.

Hox, J. J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods & Research*, *22*(3), 300–318.

Hox, J. J., de Leeuw, E. D., & Kreft, I. G. G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: A multilevel model. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 439–461). John Wiley & Sons, Inc.

Jabkowski, P., & Cichocki, P. (2019). Within-household selection of target-respondents impairs demographic representativeness of probabilistic samples: Evidence from seven rounds of the European Social Survey. *Survey Research Methods*, *13*(2), 167–180.

Johann, D., & Mayer, S. J. (2021). Do interviewers affect measures of factual political knowledge? Evidence from Austria and Germany. *International Journal of Public Opinion Research*, *33*(4), 998–1011.

Kerwin, J. T., & Ordaz Reynoso, N. (2021). You know what I know: Interviewer knowledge effects in subjective expectation elicitation. *Demography*, *58*(1), 1–29.

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail – web mixed-mode surveys. *Social Science Computer Review*, *37*(2), 214–233.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, *57*(297), 92–115.

Koczela, S., Furlong, C., McCarthy, J., & Mushtaq, A. (2015). Curbstoning and beyond: Confronting data fabrication in survey research. *Statistical Journal of the IAOS*, *31*(3), 413–422.

Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods*, *1*(2), 55–67.

Kreuter, F., Couper, M., & Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. *Proceedings of the Survey Research Method Section, American Statistical Association*, 282–296.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Kuriakose, N., & Robbins, M. (2016). Don't get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS*, *32*(3), 283–291.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Loosveldt, G., & Beullens, K. (2013). 'How long will it take?' An analysis of interview length in the fifth round of the European Social Survey. *Survey Research Methods*, *7*(2), 69–78.

Loosveldt, G., & Beullens, K. (2017). Interviewer effects on non-differentiation and straightlining in the European Social Survey. *Journal of Official Statistics*, *33*(2), 409–426.

Lusardi, A., & Mitchell, O. S. (2008). Planning and financial literacy: How do women fare? *American Economic Review*, *98*(2), 413–417.

Lusardi, A., & Mitchell, O. S. (2011). Financial literacy around the world: An overview. *Journal of Pension Economics and Finance*, *10*(4), 497–508.

Menold, N. (2014). The influence of sampling method and interviewers on sample realization in the European Social Survey. *Survey Methodology*, *40*(1), 105–123.

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable.* Lulu.com.

Olbrich, L., Kosyakova, Y., Sakshaug, J. W., & Schwanhäuser, S. (2024). Detecting interviewer fraud using multilevel models. *Journal of Survey Statistics and Methodology*, *12*(1), 14–35.

Olson, K., & Peytchev, A. (2007). Effect of interviewer experience on interview pace and interviewer attitudes. *Public Opinion Quarterly*, *71*(2), 273–286.

O'Muircheartaigh, C., & Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *161*(1), 63–77.

Pickery, J., & Loosveldt, G. (1998). The impact of respondent and interviewer characteristics on the number of "No Opinion" answers. *Quality & Quantity*, *32*, 31–45.

Pickery, J., & Loosveldt, G. (2001). An exploration of question characteristics that mediate interviewer effects on item nonresponse. *Journal of Official Statistics*, *17*(3), 337–350.

Reiter, S., & Beckmann, E. (2020). How financially literate is CESEE? Insights from the OeNB Euro Survey. *Focus on European Economic Integration Q3/2020. OeNB*, 36–59.

Rybak, A. (2023). Survey mode and nonresponse bias: A meta-analysis based on the data from the international social survey programme waves 1996–2018 and the European social survey rounds 1 to 9. *PLOS ONE*, *18*(3), e0283092.

Sarracino, F., & Mikucka, M. (2017). Bias and effiency loss in regression estimates due to duplicated observations: A Monte Carlo simulation. *Survey Research Methods*, *11*(1), 17–44.

Schäfer, C., Schräpler, J.-P., Müller, K.-R., & Wagner, G. G. (2005). Automatic identification of faked and fraudulent interviews in the German SOEP. *Schmollers Jahrbuch*, *125*(1), 183–193.

Schnell, R., & Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, *21*(3), 389–410.

Schwanhäuser, S., Sakshaug, J. W., & Kosyakova, Y. (2022). How to catch a falsifier: Comparison of statistical detection methods for interviewer falsification. *Public Opinion Quarterly*, *81*(1), 1–31.

Shapley, L. S. (1953). A value for n-person games. In *Contributions to the theory of games* (pp. 307–3017).

Silber, H., Roßmann, J., Gummer, T., Zins, S., & Weyandt, K. W. (2021). The effects of question, respondent and interviewer characteristics on two types of item nonresponse. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(3), 1052–1069.

Simmons, K., Mercer, A., Schwarzer, S., & Kennedy, C. (2016). Evaluating a new proposal for detecting data falsification in surveys: The underlying causes of "high matches" between survey respondents. *Statistical Journal of the IAOS*, *32*(3), 327–338.

Slomczynski, K. M., Powalko, P., & Krauze, T. (2017). Non-unique records in international survey projects: The need for extending data quality control. *Survey Research Methods*, *11*(1), 1–16.

Sodeur, W. (1997). Interne Kriterien zur Beurteilung von Wahrscheinlichkeitsauswahlen. *ZA-Information*, *41*, 58–82.

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Sturgis, P., Maslovskaya, O., Durrant, G., & Brunton-Smith, I. (2021). The interviewer contribution to variability in response times in face-to-face interview surveys. *Journal of Survey Statistics and Methodology*, *9*(4), 701–721.

Vandenplas, C., Loosveldt, G., Beullens, K., & Denies, K. (2018). Are interviewer effects on interview speed related to interviewer effects on straight-lining tendency in the European Social Survey? An interviewer-related analysis. *Journal of Survey Statistics and Methodology*, *6*, 516–538.

Waldmann, S., Sakshaug, J. W., & Cernat, A. (2023). Interviewer effects on the measurement of physical performance in a cross-national biosocial survey. *Journal of Survey Statistics and Methodology*, smad031. https://doi.org/10.1093/jssam/smad031

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, *5*(2), 175–211.

West, B. T., Conrad, F. G., Kreuter, F., & Mittereder, F. (2018). Nonresponse and measurement error variance among interviewers in standardized and conversational interviewing. *Journal of Survey Statistics and Methodology*, *6*(3), 335–359.

West, B. T., Kreuter, F., & Jaenichen, U. (2013). “Interviewer” effects in face-to-face surveys: A function of sampling, measurement error, or nonresponse? *Journal of Official Statistics*, *29*(2), 277–297.

West, B. T., & Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, *74*(5), 1004–1026.

Wuyts, C., & Loosveldt, G. (2019). *Quality matrix for the European Social Survey, Round 8.* Centre for Sociological Research, KU Leuven.

Yamamoto, K., & Lennon, M. L. (2018). Understanding and detecting data fabrication in large-scale assessments. *Quality Assurance in Education*, *26*(2), 196–212.

Yan, T. (2008). Nondifferentiation. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 520–521). SAGE Publications, Inc.

Zins, S., & Burgard, J. P. (2020). Considering interviewer and design effects when planning sample sizes. *Survey Methodology*, *46*(1), 93–119.

Online supplementary material

6.A Algorithm for merging PSUs

Algorithm 1 Merging PSUs based on GPS proximity

```
while only one interviewer in at least one PSU do
    calculate distance matrix of all PSUs
    subset rows to PSUs where only one interviewer worked
    among these PSUs, find PSU $i$ with shortest distance to any other PSU $j$
    merge $i$ and $j$
    use midpoint between $i$ and $j$ as updated GPS coordinate
end while
```

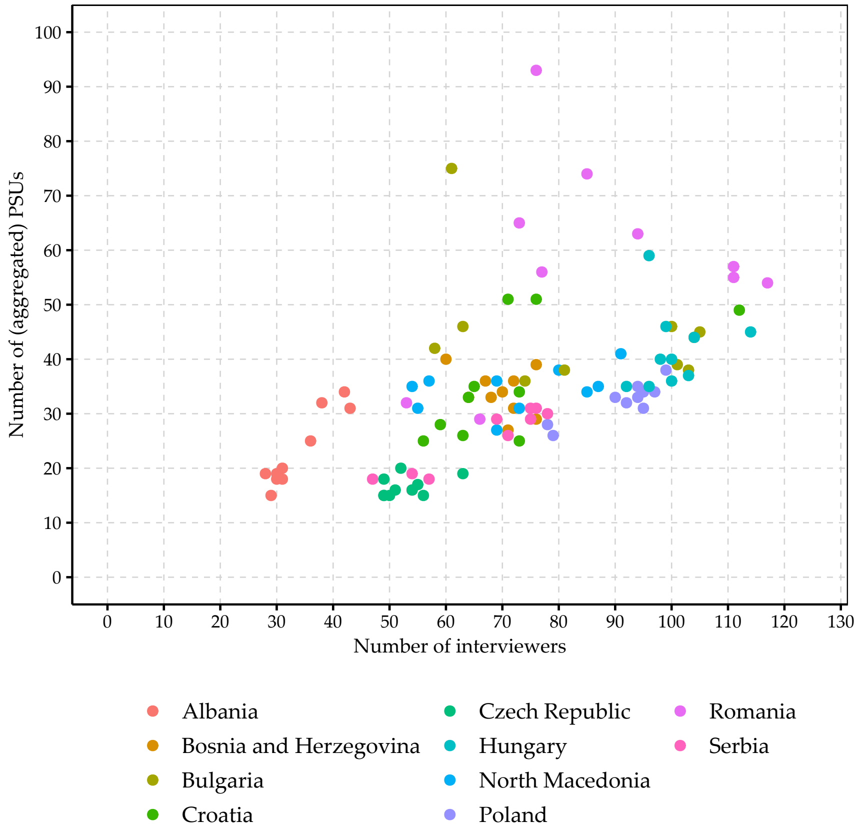


Figure 6.A1: Number of interviewers and (aggregated) PSUs in each country-year.

6.B Response rates

Table 6.B1: Response rates (AAPOR RR1).

Country	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Albania	0.670	0.671	0.699	0.667	0.803	0.534	0.668	0.660	0.635	0.686
Bosnia and Herzegovina	0.626	0.219	0.724	0.745		0.783	0.754	0.717	0.592	0.625
Bulgaria	0.332	0.340	0.364	0.361	0.095	0.401	0.450	0.397	0.429	0.447
Croatia	0.308	0.280	0.299	0.342	0.348	0.351	0.344	0.349	0.326	0.316
Czech Republic	0.523	0.563	0.600	0.588	0.572	0.555	0.559	0.529	0.455	0.468
North Macedonia	0.594	0.576	0.606	0.571	0.525	0.505	0.422	0.455	0.740	0.757
Hungary	0.422	0.408	0.392	0.418	0.429	0.419	0.400	0.383	0.343	0.312
Poland	0.572	0.498	0.417	0.418	0.396	0.359	0.318	0.295	0.271	0.270
Romania	0.807	0.806	0.770	0.800	0.683	0.607	0.747	0.785		0.601
Serbia	0.752	0.729	0.732	0.778		0.685	0.686	0.640		0.597

Notes: For missing country-years, the gross sample size is not available.

6.C Control variables

Table 6.C1: Description of control variables.

Variable	Description
Respondent characteristics	
Age	Missing for 15 observations across all country-years, which are excluded from the multilevel analysis. Scaled to mean zero and standard deviation one for multilevel models.
Gender	0 = male, 1 = female
Education	Three categories: No formal/primary education, (Post-)secondary education, tertiary education. In some country-years, categories 1 and 2 are combined. If the number of missings < 50, these observations are excluded from the multilevel models. If the number of missings ≥ 50, a separate "Missing" category is added. In several country-years, the "No formal/primary education" contains only a few observations which lead to (quasi-)complete separation in the multilevel logistic regressions. In this case, these observations are added to the "(Post-)secondary education". The corresponding country-years are Albania in 2015 and 2021, Bulgaria in all years, Romania in 2017 and 2021, the Czech Republic in 2019 and 2021, Hungary, Croatia, and Poland in 2021.
Employment	Four categories: employed, self-employed, retired/student/maternity leave, unemployed. If the number of missings < 50, these observations are excluded from the multilevel models. If the number of missings exceeds ≥ 50, a separate "Missing" category is added. In Croatia in 2013, 2016, and 2021, Serbia in 2019, and Hungary in 2015 the self-employed were combined with the employed to avoid quasi-complete separation.
Household characteristics	
Household size	4 categories: 1, 2, 3, ≥ 4. In Albania in 2013, 2017, 2018, 2020, and 2021, categories 1 and 2 are merged due to few cases with a single household member.
Dwelling condition	2 categories: excellent/good condition, poor condition. Based on the interviewer's assessment.
Household income quintiles	Quintiles based on household income. Separate category for missings. In Hungary in 2015, Bulgaria in 2020, and Croatia in 2015 categories were combined due to (quasi-)complete separation.
Regional characteristics	
Town size	4 categories: 0 - 9,999, 10,000 - 49,999, 50,000 - 99,999, $\geq$ 100,000
Nightlight activity	VIIRS nightlight (annual VNL V2) within a radius of 5km around the random route starting point, source: Earth Observation Group

6.D Interviewer characteristics

We also investigate the role of interviewer characteristics (age, gender, experience) on the financial literacy score for the years 2017 to 2021. Figure 6.D1 shows the discrepancy between ICCs from models with and without interviewer characteristics. In sum, the differences are minor which suggests that observable characteristics cannot explain the interviewer variance. Figure 6.D2 shows the coefficients for the three covariates for all country-years. Interviewer age is positively correlated with financial literacy in several country-years, for the interviewer gender the coefficients are more mixed, and interviewer experience does not seem to correlate with financial literacy. These results are in line with Crossley et al. (2021) who also found that interviewer age is a significant predictor of financial literacy. The mechanism behind this relationship is, however, unclear. We tested whether the match between the interviewer and respondent age (maximum 5 years difference) influences financial literacy scores, but found no evidence.

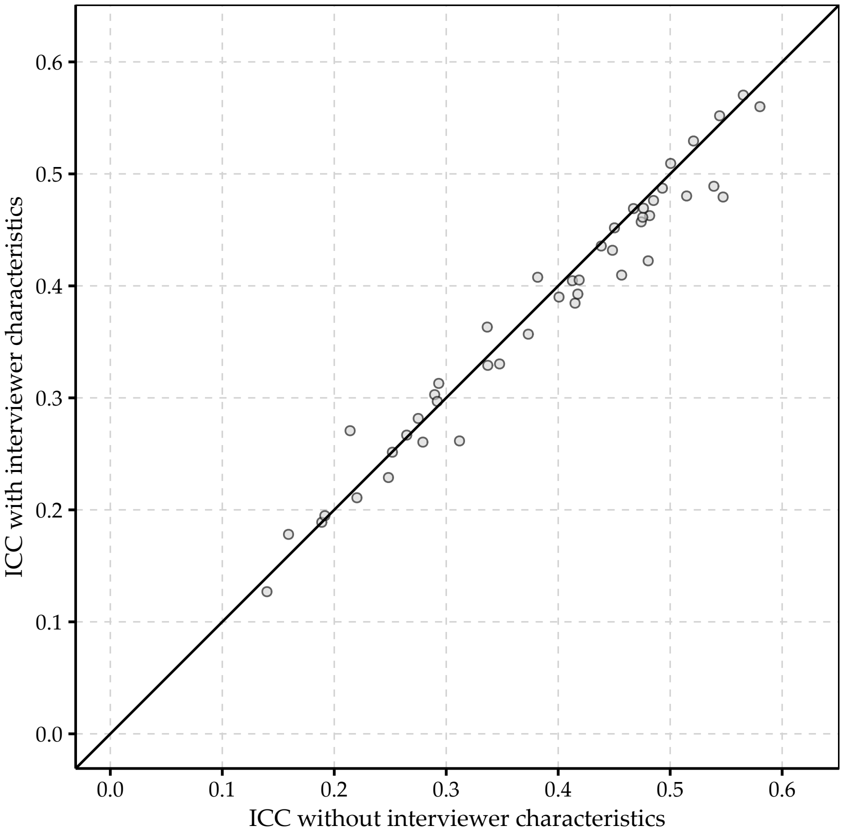


Figure 6.D1: Estimated ICCs for financial literacy score with and without interviewer characteristics.

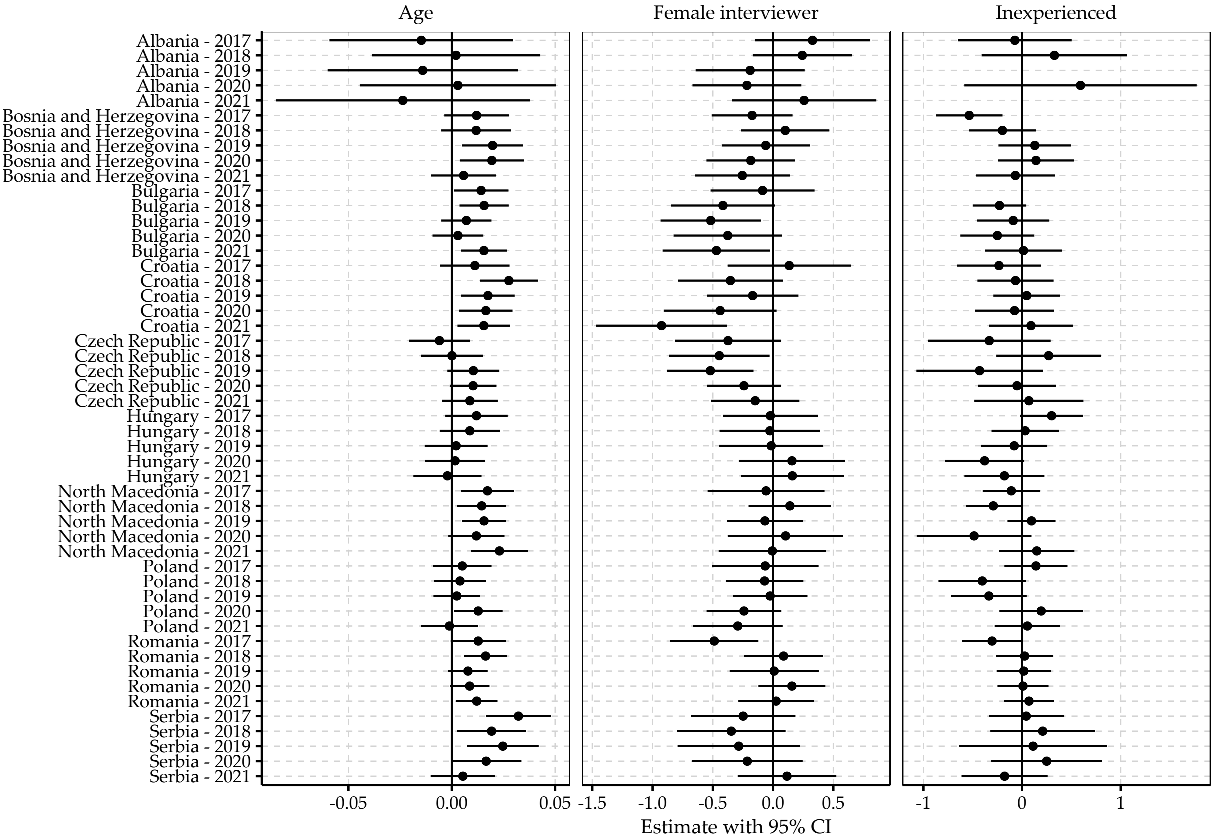


Figure 6.D2: Estimated coefficients for interviewer characteristics in multilevel models with financial literacy score as the dependent variable.

6.E Daily interviewer workload

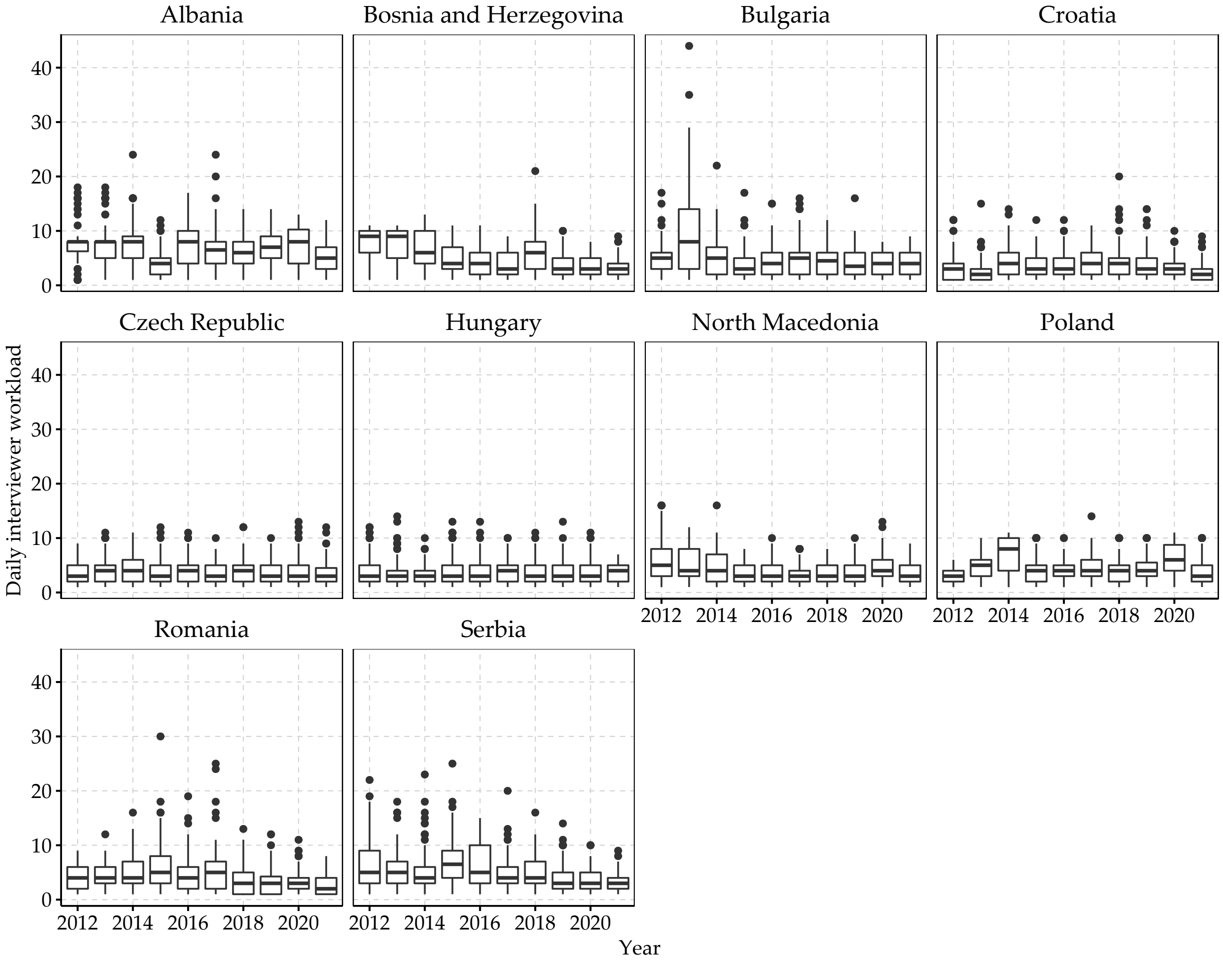


Figure 6.E1: Daily interviewer workloads across countries and years.

6.F ICC estimates with confidence intervals

Table 6.F1: ICCs for financial literacy - interest rate question.

Country	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Albania	0.596	0.438	0.628	0.447	0.283	0.379	0.439	0.508	0.520	0.497
	[0.398,0.731]	[0.278,0.579]	[0.477,0.753]	[0.271,0.586]	[0.112,0.392]	[0.168,0.539]	[0.228,0.604]	[0.264,0.704]	[0.307,0.675]	[0.167,0.712]
Bosnia and Herzegovina	0.536	0.583	0.449	0.499	0.454	0.510	0.568	0.473	0.711	0.540
	[0.368,0.664]	[0.441,0.673]	[0.284,0.587]	[0.347,0.589]	[0.290,0.575]	[0.367,0.630]	[0.436,0.708]	[0.342,0.595]	[0.570,0.818]	[0.350,0.688]
Bulgaria	0.560	0.529	0.441	0.683	0.718	0.484	0.516	0.569	0.662	0.623
	[0.389,0.666]	[0.397,0.650]	[0.299,0.601]	[0.511,0.797]	[0.607,0.803]	[0.355,0.588]	[0.375,0.633]	[0.437,0.658]	[0.535,0.731]	[0.490,0.705]
Croatia	0.403	0.042	0.344	0.458	0.510	0.514	0.670	0.572	0.561	0.617
	[0.196,0.600]	[0.000,0.112]	[0.183,0.490]	[0.305,0.566]	[0.359,0.618]	[0.336,0.629]	[0.499,0.780]	[0.410,0.708]	[0.393,0.664]	[0.459,0.716]
Czech Republic	0.362	0.601	0.504	0.535	0.442	0.506	0.641	0.386	0.401	0.365
	[0.176,0.485]	[0.411,0.720]	[0.321,0.607]	[0.336,0.619]	[0.219,0.553]	[0.314,0.676]	[0.439,0.732]	[0.186,0.552]	[0.239,0.524]	[0.189,0.491]
Hungary	0.542	0.715	0.588	0.634	0.549	0.471	0.733	0.711	0.667	0.705
	[0.391,0.655]	[0.556,0.780]	[0.422,0.673]	[0.483,0.709]	[0.353,0.696]	[0.301,0.637]	[0.557,0.844]	[0.516,0.803]	[0.474,0.776]	[0.576,0.790]
North Macedonia	0.212	0.289	0.403	0.362	0.304	0.427	0.349	0.172	0.418	0.571
	[0.045,0.352]	[0.139,0.418]	[0.268,0.522]	[0.217,0.472]	[0.161,0.403]	[0.277,0.540]	[0.189,0.479]	[0.030,0.311]	[0.260,0.582]	[0.427,0.685]
Poland	0.533	0.380	0.311	0.446	0.387	0.337	0.271	0.284	0.204	0.247
	[0.379,0.640]	[0.227,0.524]	[0.149,0.432]	[0.280,0.539]	[0.239,0.491]	[0.204,0.439]	[0.136,0.361]	[0.135,0.404]	[0.086,0.299]	[0.120,0.350]
Romania	0.396	0.348	0.448	0.337	0.135	0.385	0.219	0.211	0.236	0.357
	[0.256,0.493]	[0.222,0.488]	[0.275,0.576]	[0.191,0.463]	[0.035,0.245]	[0.209,0.532]	[0.118,0.329]	[0.093,0.329]	[0.118,0.347]	[0.243,0.469]
Serbia	0.565	0.522	0.492	0.363	0.396	0.516	0.528	0.588	0.522	0.396
	[0.343,0.711]	[0.326,0.670]	[0.276,0.677]	[0.214,0.515]	[0.233,0.542]	[0.354,0.612]	[0.314,0.721]	[0.404,0.699]	[0.354,0.634]	[0.226,0.544]

Notes: 95 percent confidence intervals based on 200 bootstrap replications in brackets.

Table 6.F2: ICCs for financial literacy - exchange rate question.

Country	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Albania	0.347	0.427	0.371	0.221	0.234	0.560	0.212	0.216	0.554	0.227
	[0.201,0.475]	[0.249,0.551]	[0.216,0.505]	[0.086,0.340]	[0.075,0.361]	[0.303,0.679]	[0.061,0.385]	[0.069,0.394]	[0.372,0.715]	[0.089,0.333]
Bosnia and Herzegovina	0.545	0.433	0.318	0.382	0.598	0.541	0.358	0.585	0.500	0.439
	[0.367,0.650]	[0.296,0.599]	[0.160,0.472]	[0.240,0.473]	[0.438,0.695]	[0.408,0.666]	[0.199,0.511]	[0.428,0.709]	[0.357,0.626]	[0.273,0.601]
Bulgaria	0.449	0.402	0.417	0.505	0.568	0.549	0.506	0.512	0.511	0.518
	[0.297,0.618]	[0.265,0.529]	[0.257,0.545]	[0.330,0.652]	[0.453,0.676]	[0.400,0.651]	[0.361,0.596]	[0.378,0.603]	[0.379,0.594]	[0.347,0.637]
Croatia	0.392	0.033	0.491	0.390	0.571	0.477	0.610	0.414	0.606	0.643
	[0.212,0.524]	[0.000,0.082]	[0.309,0.627]	[0.259,0.496]	[0.431,0.677]	[0.311,0.565]	[0.457,0.715]	[0.228,0.568]	[0.447,0.712]	[0.490,0.746]
Czech Republic	0.270	0.420	0.286	0.428	0.342	0.476	0.442	0.405	0.406	0.169
	[0.120,0.418]	[0.217,0.537]	[0.129,0.374]	[0.238,0.509]	[0.164,0.508]	[0.297,0.632]	[0.279,0.556]	[0.245,0.517]	[0.234,0.514]	[0.051,0.253]
Hungary	0.425	0.489	0.430	0.477	0.422	0.399	0.546	0.571	0.536	0.558
	[0.263,0.584]	[0.282,0.643]	[0.270,0.535]	[0.311,0.564]	[0.254,0.515]	[0.258,0.550]	[0.393,0.691]	[0.431,0.685]	[0.343,0.684]	[0.420,0.679]
North Macedonia	0.215	0.170	0.368	0.315	0.294	0.299	0.277	0.288	0.521	0.509
	[0.104,0.339]	[0.009,0.290]	[0.200,0.494]	[0.174,0.400]	[0.163,0.391]	[0.096,0.465]	[0.159,0.367]	[0.151,0.427]	[0.372,0.655]	[0.369,0.621]
Poland	0.475	0.260	0.329	0.293	0.321	0.383	0.210	0.185	0.060	0.174
	[0.329,0.592]	[0.142,0.384]	[0.156,0.460]	[0.156,0.366]	[0.188,0.412]	[0.223,0.468]	[0.075,0.276]	[0.081,0.293]	[0.000,0.112]	[0.075,0.260]
Romania	0.394	0.277	0.250	0.342	0.142	0.211	0.267	0.288	0.377	0.337
	[0.259,0.502]	[0.158,0.382]	[0.119,0.359]	[0.207,0.449]	[0.059,0.225]	[0.092,0.333]	[0.147,0.379]	[0.147,0.398]	[0.240,0.494]	[0.197,0.465]
Serbia	0.531	0.380	0.545	0.341	0.409	0.487	0.437	0.526	0.543	0.416
	[0.257,0.649]	[0.211,0.531]	[0.374,0.663]	[0.188,0.429]	[0.281,0.522]	[0.339,0.605]	[0.284,0.574]	[0.327,0.625]	[0.364,0.652]	[0.271,0.550]

Notes: 95 percent confidence intervals based on 200 bootstrap replications in brackets.

Table 6.F3: ICCs for financial literacy - inflation question.

Country	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Albania	0.326	0.183	0.191	0.373	0.091	0.294	0.173	0.129	0.504	0.211
	[0.179,0.478]	[0.063,0.305]	[0.074,0.316]	[0.214,0.514]	[0.009,0.189]	[0.106,0.432]	[0.067,0.274]	[0.022,0.269]	[0.300,0.651]	[0.076,0.303]
Bosnia and Herzegovina	0.371	0.409	0.361	0.391	0.327	0.389	0.411	0.487	0.424	0.293
	[0.223,0.494]	[0.274,0.549]	[0.215,0.508]	[0.243,0.501]	[0.179,0.450]	[0.271,0.512]	[0.291,0.537]	[0.336,0.615]	[0.284,0.554]	[0.165,0.404]
Bulgaria	0.471	0.356	0.447	0.410	0.492	0.484	0.460	0.429	0.338	0.376
	[0.290,0.601]	[0.200,0.512]	[0.310,0.589]	[0.268,0.532]	[0.347,0.617]	[0.306,0.593]	[0.320,0.537]	[0.270,0.549]	[0.178,0.461]	[0.230,0.508]
Croatia	0.479	0.064	0.287	0.323	0.342	0.459	0.498	0.492	0.517	0.455
	[0.288,0.638]	[0.000,0.133]	[0.161,0.418]	[0.173,0.450]	[0.189,0.476]	[0.273,0.570]	[0.341,0.614]	[0.343,0.616]	[0.355,0.625]	[0.288,0.576]
Czech Republic	0.217	0.433	0.278	0.264	0.300	0.409	0.414	0.299	0.316	0.341
	[0.082,0.299]	[0.252,0.544]	[0.119,0.381]	[0.112,0.343]	[0.128,0.406]	[0.228,0.554]	[0.236,0.547]	[0.169,0.403]	[0.141,0.417]	[0.167,0.441]
Hungary	0.572	0.564	0.414	0.435	0.549	0.453	0.643	0.528	0.475	0.617
	[0.436,0.692]	[0.396,0.676]	[0.278,0.495]	[0.264,0.561]	[0.386,0.646]	[0.268,0.589]	[0.493,0.748]	[0.350,0.691]	[0.313,0.577]	[0.497,0.726]
North Macedonia	0.521	0.592	0.450	0.438	0.364	0.403	0.320	0.248	0.491	0.522
	[0.340,0.659]	[0.470,0.700]	[0.270,0.575]	[0.303,0.544]	[0.234,0.459]	[0.248,0.509]	[0.200,0.417]	[0.108,0.387]	[0.344,0.624]	[0.372,0.635]
Poland	0.443	0.423	0.519	0.454	0.335	0.306	0.288	0.102	0.264	0.334
	[0.314,0.566]	[0.272,0.572]	[0.309,0.639]	[0.270,0.544]	[0.193,0.407]	[0.167,0.427]	[0.150,0.387]	[0.021,0.182]	[0.128,0.356]	[0.161,0.444]
Romania	0.332	0.374	0.282	0.309	0.195	0.190	0.323	0.120	0.115	0.198
	[0.184,0.454]	[0.249,0.495]	[0.139,0.407]	[0.182,0.416]	[0.094,0.274]	[0.066,0.300]	[0.216,0.434]	[0.031,0.211]	[0.026,0.201]	[0.106,0.309]
Serbia	0.552	0.426	0.565	0.440	0.449	0.513	0.286	0.506	0.450	0.575
	[0.318,0.660]	[0.263,0.628]	[0.353,0.662]	[0.260,0.552]	[0.318,0.559]	[0.369,0.629]	[0.144,0.470]	[0.319,0.610]	[0.290,0.547]	[0.414,0.680]

Notes: 95 percent confidence intervals based on 200 bootstrap replications in brackets.

Table 6.F4: ICCs for financial literacy - risk diversification question.

Country	2012	2013	2014	2015	2016	2018	2019	2021
Albania	0.235	0.179	0.241	0.426	0.197	0.248	0.117	0.025
	[0.112,0.346]	[0.079,0.282]	[0.132,0.351]	[0.244,0.547]	[0.057,0.309]	[0.102,0.377]	[0.025,0.217]	[0.000,0.054]
Bosnia and	0.249	0.258	0.352	0.310	0.274	0.272	0.411	0.244
Herzegovina	[0.118,0.371]	[0.132,0.375]	[0.191,0.500]	[0.171,0.418]	[0.144,0.377]	[0.167,0.369]	[0.272,0.533]	[0.126,0.353]
Bulgaria	0.283	0.232	0.382	0.449	0.416	0.511	0.445	0.389
	[0.149,0.388]	[0.127,0.346]	[0.227,0.526]	[0.306,0.551]	[0.297,0.518]	[0.341,0.645]	[0.304,0.539]	[0.251,0.487]
Croatia	0.300	0.029	0.345	0.325	0.311	0.353	0.320	0.288
	[0.139,0.437]	[0.000,0.077]	[0.209,0.450]	[0.174,0.464]	[0.163,0.430]	[0.199,0.493]	[0.190,0.445]	[0.149,0.412]
Czech Republic	0.213	0.329	0.299	0.254	0.289	0.098	0.161	0.154
	[0.094,0.288]	[0.152,0.488]	[0.124,0.397]	[0.120,0.343]	[0.130,0.381]	[0.018,0.175]	[0.043,0.266]	[0.053,0.241]
Hungary	0.447	0.382	0.269	0.428	0.462	0.441	0.297	0.372
	[0.328,0.567]	[0.243,0.468]	[0.141,0.385]	[0.278,0.519]	[0.275,0.553]	[0.283,0.583]	[0.164,0.431]	[0.256,0.457]
North Macedonia	0.191	0.292	0.319	0.219	0.107	0.229	0.257	0.330
	[0.047,0.321]	[0.175,0.381]	[0.177,0.446]	[0.107,0.314]	[0.000,0.253]	[0.074,0.358]	[0.123,0.385]	[0.181,0.458]
Poland	0.447	0.252	0.238	0.251	0.303	0.222	0.162	0.207
	[0.303,0.554]	[0.128,0.341]	[0.094,0.382]	[0.130,0.337]	[0.176,0.382]	[0.115,0.307]	[0.064,0.264]	[0.090,0.287]
Romania	0.286	0.241	0.260	0.227	0.110	0.229	0.109	0.284
	[0.133,0.395]	[0.115,0.380]	[0.117,0.372]	[0.103,0.344]	[0.018,0.201]	[0.123,0.372]	[0.000,0.209]	[0.160,0.392]
Serbia	0.365	0.261	0.592	0.208	0.270	0.307	0.282	0.311
	[0.165,0.499]	[0.092,0.436]	[0.327,0.724]	[0.090,0.318]	[0.136,0.415]	[0.173,0.423]	[0.158,0.365]	[0.170,0.404]

Notes: 95 percent confidence intervals based on 200 bootstrap replications in brackets.

Table 6.F5: ICCs for financial literacy score.

Country	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Albania	0.419	0.278	0.402	0.272	0.171	0.214	0.290	0.293	0.337	0.521
	[0.283,0.538]	[0.168,0.386]	[0.280,0.527]	[0.161,0.375]	[0.064,0.267]	[0.095,0.320]	[0.149,0.423]	[0.169,0.450]	[0.184,0.472]	[0.339,0.642]
Bosnia and	0.397	0.418	0.277	0.404	0.449	0.457	0.438	0.474	0.482	0.450
Herzegovina	[0.291,0.499]	[0.312,0.527]	[0.176,0.405]	[0.288,0.516]	[0.343,0.543]	[0.326,0.546]	[0.330,0.543]	[0.366,0.575]	[0.396,0.572]	[0.347,0.556]
Bulgaria	0.416	0.317	0.335	0.400	0.526	0.413	0.415	0.475	0.476	0.417
	[0.307,0.519]	[0.197,0.428]	[0.224,0.436]	[0.287,0.512]	[0.436,0.613]	[0.327,0.512]	[0.326,0.516]	[0.382,0.558]	[0.376,0.562]	[0.334,0.510]
Croatia	0.342	0.020	0.316	0.329	0.433	0.493	0.547	0.373	0.515	0.539
	[0.221,0.454]	[0.000,0.076]	[0.200,0.435]	[0.229,0.427]	[0.335,0.542]	[0.373,0.599]	[0.431,0.637]	[0.257,0.476]	[0.403,0.605]	[0.429,0.629]
Czech Republic	0.248	0.462	0.319	0.316	0.298	0.348	0.401	0.312	0.220	0.292
	[0.139,0.371]	[0.334,0.571]	[0.225,0.418]	[0.173,0.420]	[0.182,0.422]	[0.222,0.470]	[0.292,0.508]	[0.167,0.409]	[0.124,0.310]	[0.184,0.386]
Hungary	0.388	0.520	0.455	0.455	0.366	0.419	0.500	0.544	0.448	0.565
	[0.282,0.509]	[0.396,0.624]	[0.365,0.537]	[0.352,0.549]	[0.246,0.478]	[0.296,0.539]	[0.394,0.606]	[0.432,0.638]	[0.340,0.575]	[0.481,0.646]
North Macedonia	0.328	0.374	0.379	0.370	0.340	0.378	0.319	0.202	0.429	0.426
	[0.231,0.434]	[0.266,0.480]	[0.243,0.483]	[0.272,0.443]	[0.246,0.428]	[0.276,0.464]	[0.220,0.422]	[0.103,0.297]	[0.322,0.547]	[0.303,0.521]
Poland	0.443	0.304	0.312	0.375	0.349	0.337	0.252	0.159	0.140	0.265
	[0.319,0.534]	[0.212,0.423]	[0.217,0.422]	[0.292,0.465]	[0.264,0.446]	[0.219,0.444]	[0.166,0.329]	[0.081,0.234]	[0.074,0.213]	[0.180,0.350]
Romania	0.321	0.280	0.209	0.326	0.176	0.248	0.279	0.189	0.192	0.275
	[0.226,0.412]	[0.206,0.365]	[0.126,0.295]	[0.220,0.416]	[0.099,0.253]	[0.151,0.358]	[0.183,0.389]	[0.103,0.276]	[0.107,0.268]	[0.200,0.373]
Serbia	0.431	0.367	0.450	0.414	0.345	0.480	0.382	0.580	0.485	0.467
	[0.256,0.533]	[0.244,0.508]	[0.304,0.590]	[0.279,0.515]	[0.252,0.432]	[0.364,0.564]	[0.235,0.532]	[0.465,0.655]	[0.356,0.577]	[0.356,0.560]

Notes: 95 percent confidence intervals based on 200 bootstrap replications in brackets.

6.G Boxplot of ICCs

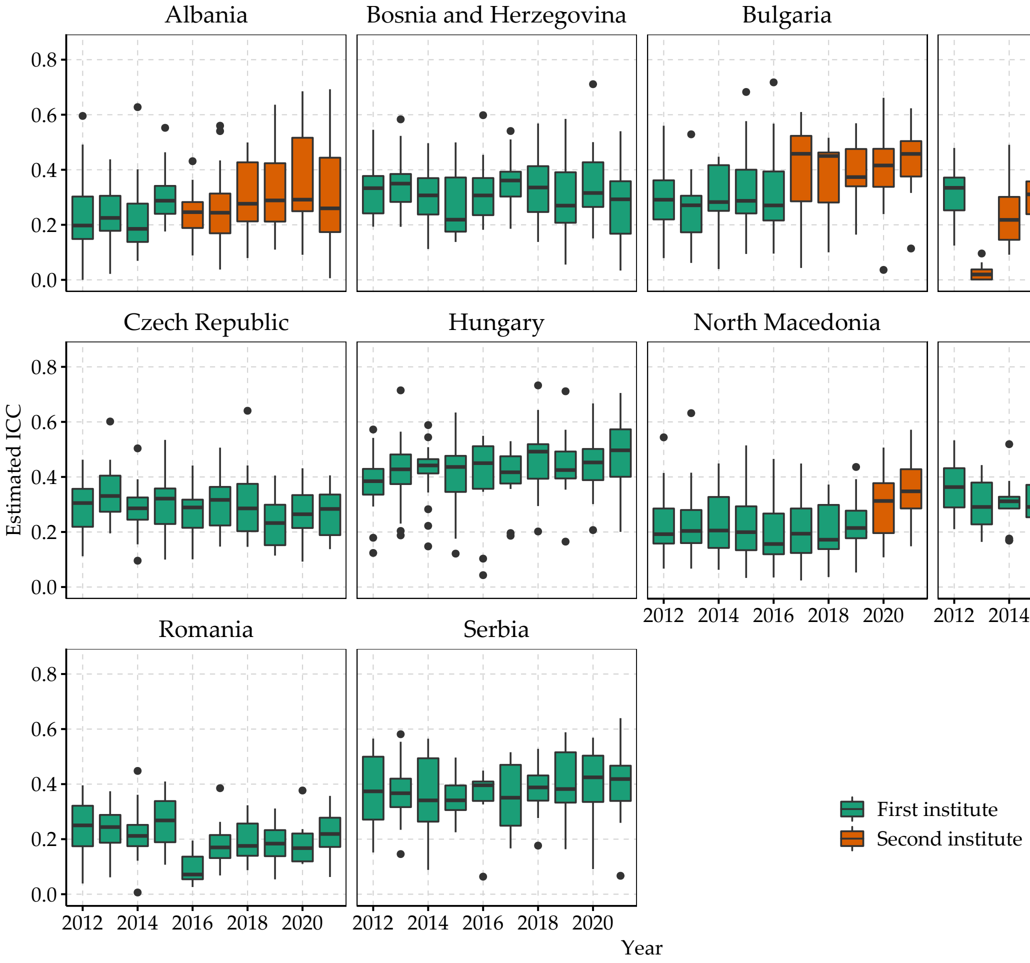


Figure 6.G1: Boxplots of ICCs, restricted to variables for which ICCs were estimated in all years in the respective country.

6.H Interviewer and PSU effects

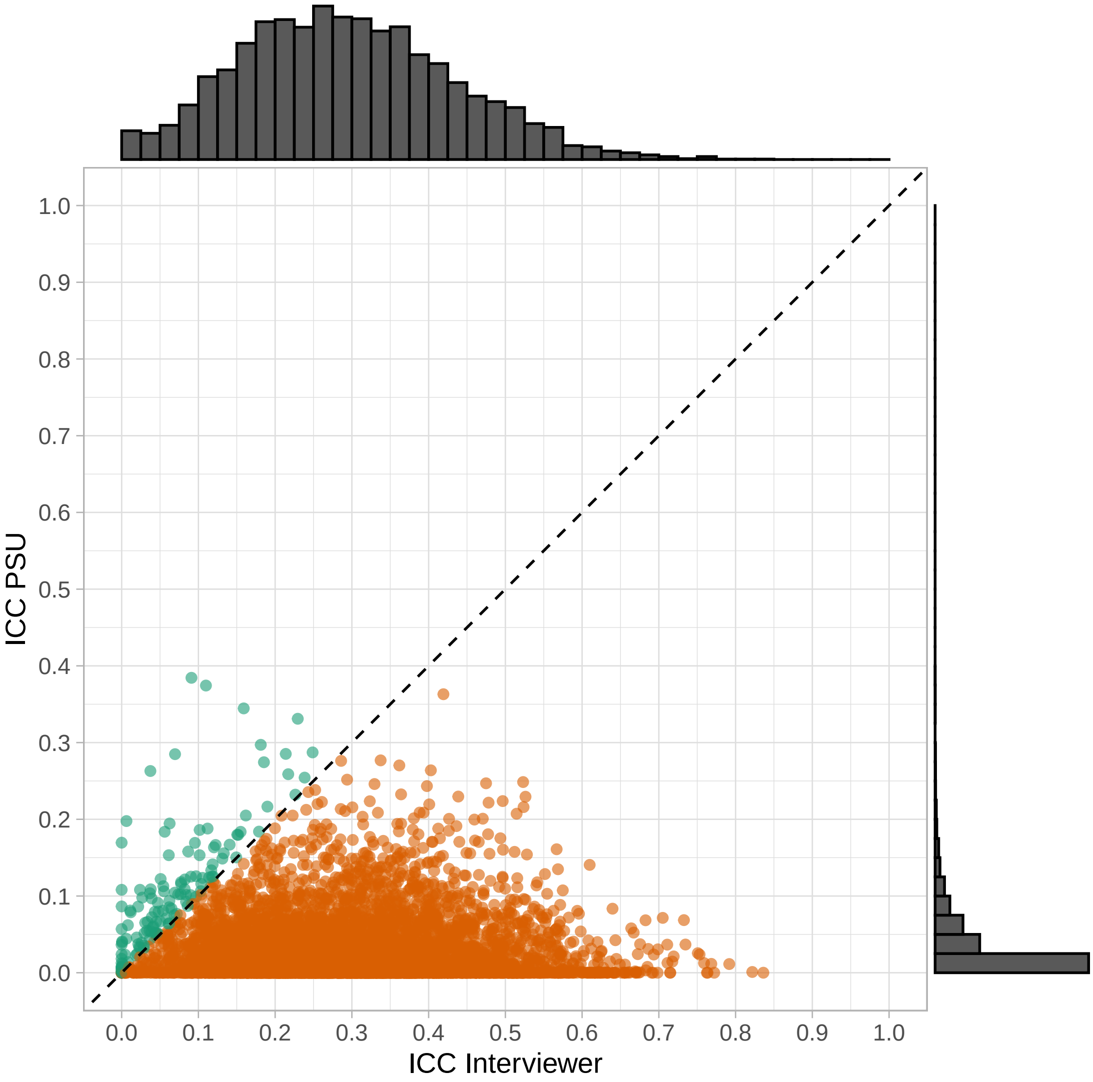


Figure 6.H1: Interviewer and PSU ICCs for all estimated models across all country-years.

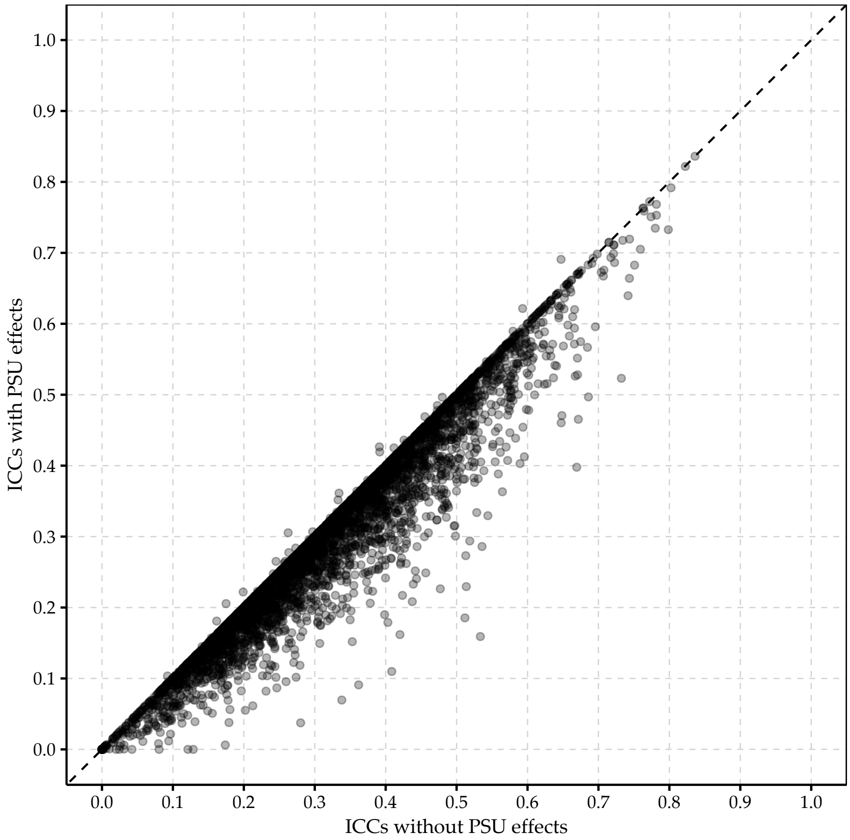


Figure 6.H2: ICCs in models with and without PSU random effects for all estimated models across all country-years.

6.I Change of fieldwork institute

In a further attempt to assess the extent to which regional differences inflate the variance estimates, we exploit the fact that four countries switched fieldwork institutes over time (Albania 2015/2016, Bulgaria 2016/2017, Croatia 2012/2013, North Macedonia 2019/2020). Hence, the set of interviewers was likely exchanged from one round to another. At the same time, many regions were sampled before and after the institute change. We restrict the data collected in the year before and after the change to 10km-radius-regions sampled in both years. To estimate the influence of the 10km-radius-regions, we fit multilevel models with several control variables (age, gender, education, employment, household size, town size, nightlight activity, dwelling characteristics, and household income quintiles), an indicator for the survey year, and random effects for the 10km-radius-regions and the interviewers. We calculate the proportions of explained variance for all questions that were part of both questionnaires and apply the same restrictions as listed above concerning item nonresponse, filtering, and extreme response distributions. If the 10km-radius-regions accounted for the majority of variation in the outcome variables, then the interviewer variance should be close to zero, while the 10km-radius-regions variance should exceed the interviewer variance. The main assumptions for this approach are that 1) regional effects are stable over time, and 2) interviewers were indeed exchanged between rounds and not re-hired by the new fieldwork institute. We cannot test the former assumption but given that we only consider a one-year difference, substantial changes are unlikely. To test the latter assumption, we use data on the interviewer characteristics and merge interviewers from the pre-change year with interviewers from the post-change year working in the same region to evaluate whether they share the same characteristics. We restrict this analysis to Bulgaria and North Macedonia as interviewer characteristics are not available for earlier years and focus specifically on the interviewers' age and gender. In Bulgaria, 9 out of 170 interviewer matches (5.3 percent) who worked in the same regions share the same gender and age (i.e., increased by one between year). In North Macedonia, 9 out of 285 interviewer matches (3.2 percent) share the same gender and age. As some of these changes may also occur due to random chance, re-hiring seems to play only a minor role.

6.I1 provides a summary of the results for the four countries. For Albania, interviewers explain on average 31.0 percent of the variation, while the regions explain 1.6 percent. For Bulgaria, interviewers explain on average 36.0 percent, while the regions explain 2.7 percent. For Croatia, the interviewers explain on average around 14.7 percent, while regions explain around 0.9 percent of the variance. In North Macedonia, interviewers explain on average 25.7 percent, regions only 2.0 percent. Thus, the results show that interviewers play a more important role than the region the interviewers are working in. However, one variable is subject to substantial regional variation, which is the question on the time it takes to reach the next bank branch. Since the proximity to the next bank branch is expected to vary across regions, this is no surprise and validates the estimation approach. In Albania, the regional effects explain 6.8 percent of the variation, in Bulgaria 25.4 percent, and in North Macedonia 27.0 percent, while regional effects are irrelevant for Croatia. Removing this variable before calculating averages results in the values denoted in parentheses in Table 6.I1. In particular, in North Macedonia this leads to a substantial decrease in the average explained variance for the regions. In summary, these results do not suggest that regional homogeneities lead to substantial inflation of estimated interviewer variance.

Table 6.I1: Results for switching fieldwork institutes.

Country	N pre	N post	N regions	N vars.	$\overline{ICC}_{int}$	$\overline{ICC}_{region}$
Albania	738	726	28	43	0.310 (0.303)	0.016 (0.015)
Bulgaria	647	537	37	31	0.360 (0.346)	0.027 (0.018)
Croatia	408	525	17	42	0.147 (0.141)	0.009 (0.009)
North Macedonia	866	836	42	32	0.257 (0.247)	0.020 (0.012)

Notes: Averages without ICCs for time to bank branch in parentheses.

6.J Item nonresponse and straightlining

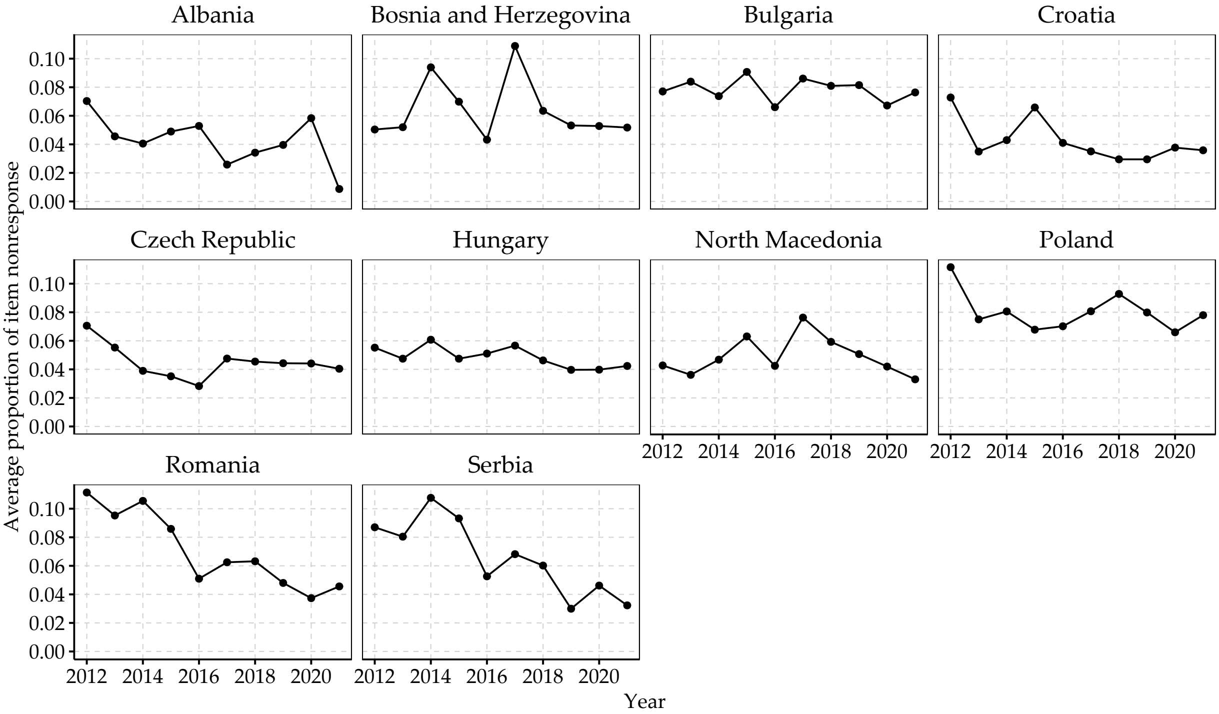


Figure 6.J1: Average item nonresponse across countries and years.

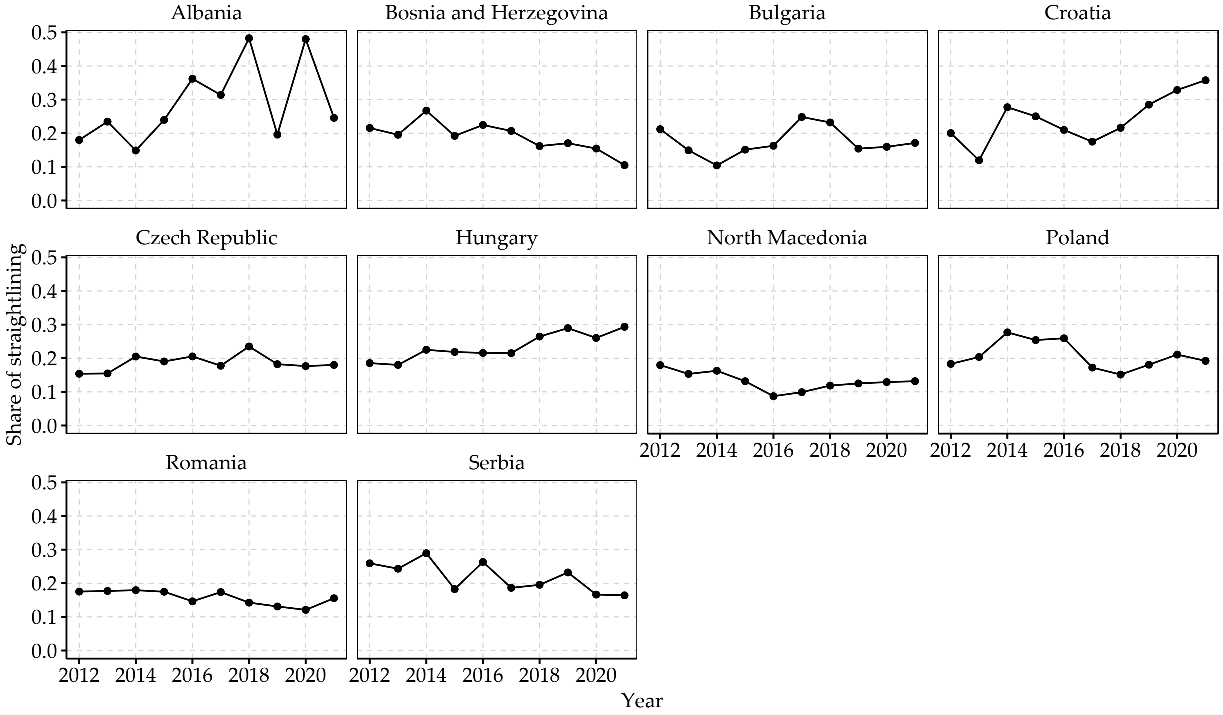


Figure 6.J2: Average share of straightlining in the trust item battery across countries and years.

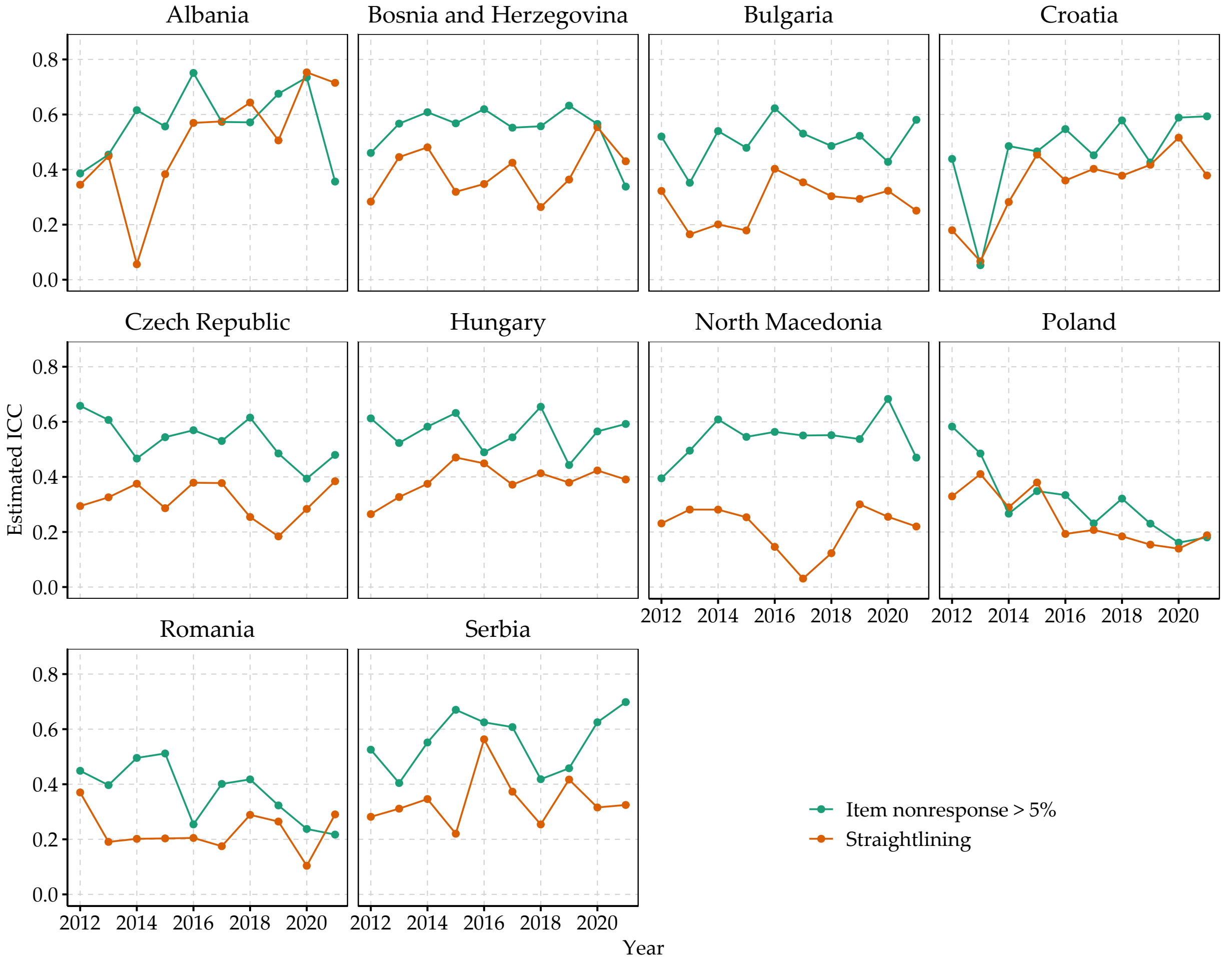


Figure 6.J3: Estimated ICCs for item nonresponse and straightlining across countries and years.

6.K Distribution of isolation forest outlier scores

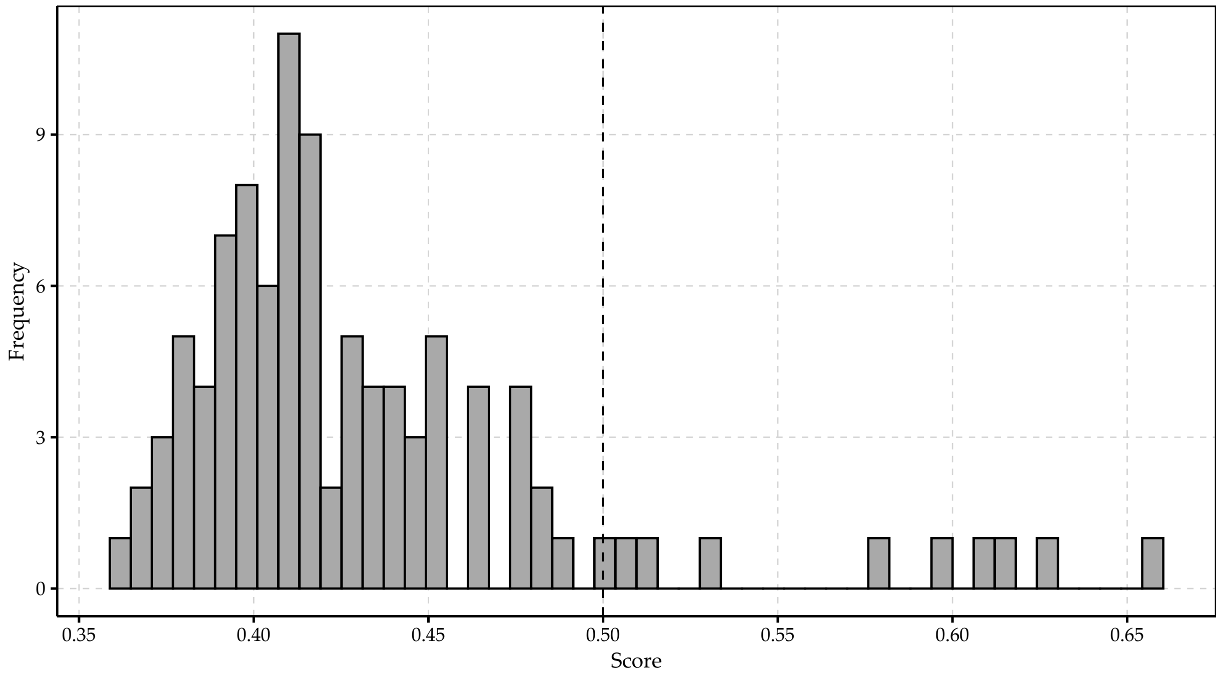


Figure 6.K1: Distribution of isolation forest outlier scores, country-year analysis.

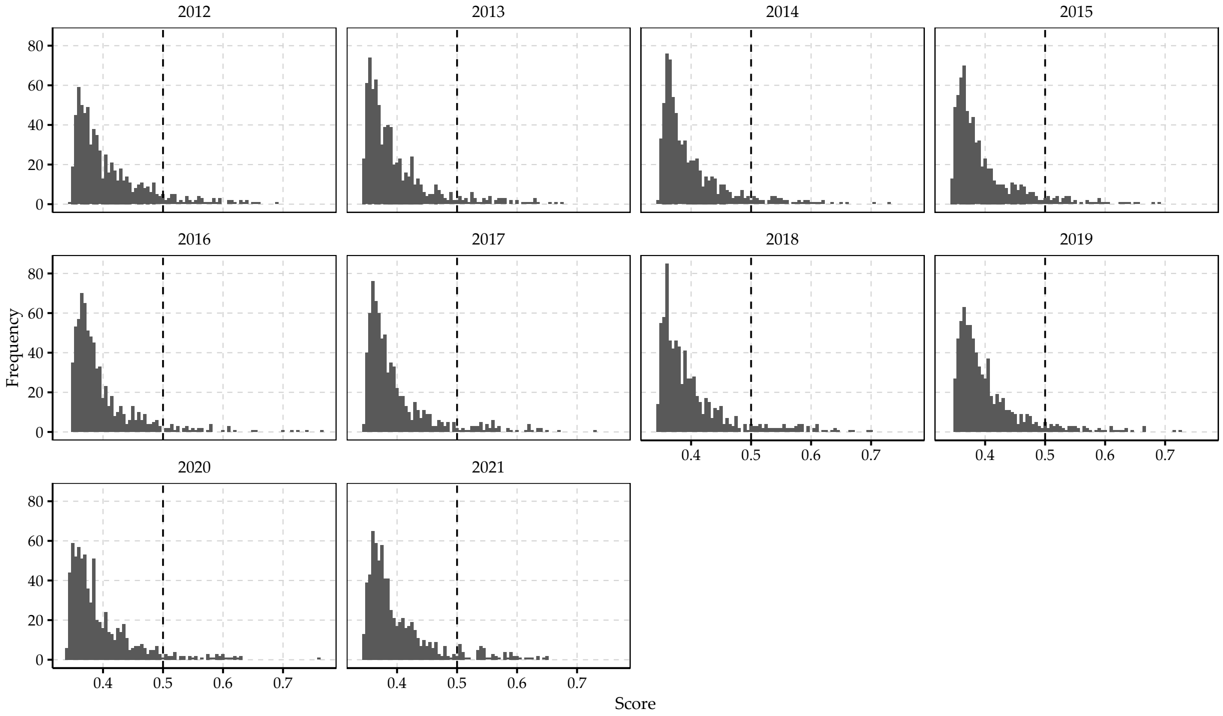


Figure 6.K2: Distribution of isolation forest outlier scores, interviewer-level analysis.

7 Evaluating methods to prevent and detect inattentive respondents in web surveys

Declaration of Contributions

Lukas Olbrich had the idea of clustering the timestamp data and jointly with Joseph W. Sakshaug developed the experimental setup. Lukas Olbrich conducted all analyses and wrote the paper.

Contributions of Co-authors

Joseph W. Sakshaug contributed to the development of the experimental setup, provided guidance and valuable input, and revised and proofread the paper. Eric Lewandowski was responsible for communicating the experimental setup to the survey panel provider and managing the data delivery.

Abstract

Inattentive respondents pose a substantial threat to data quality in web surveys. To minimize this threat, we evaluate methods for preventing and detecting inattentive responding and investigate its impacts on substantive research. First, we test the effect of asking respondents to commit to providing high-quality responses at the beginning of the survey on various data quality measures. Second, we compare the proportion of flagged respondents for two versions of an attention check item instructing them to select a specific response vs. leaving the item blank. Third, we propose a timestamp-based cluster analysis approach that identifies clusters of respondents who exhibit different speeding behaviors. Lastly, we investigate the impact of inattentive respondents on univariate, regression, and experimental analyses. Our findings show that the commitment pledge had no effect on the data quality measures. Instructing respondents to leave the item blank instead of providing a specific response significantly increased the rate of flagged respondents (by 16.8 percentage points). The timestamp-based clustering approach efficiently identified clusters of likely inattentive respondents and outperformed a related method, while providing additional insights on speeding behavior throughout the questionnaire. Lastly, we show that inattentive respondents can have substantial impacts on substantive analyses.

Keywords

cluster analysis, inattentive responding, data quality, commitment pledge, instructed response item, paradata, speeders

Acknowledgements

Lukas Olbrich gratefully acknowledges financial support from the Graduate Programme of the Institute for Employment Research (IAB) and the University of Erlangen-Nuremberg (GradAB). The authors would like to thank participants of the GOR 2024 conference and internal IAB seminars for their helpful feedback and comments.

7.1 Introduction

Researchers increasingly rely on online modes to collect survey data (Baker et al., 2010). The main reasons driving this trend are the relatively low costs of web surveys and the speed at which data can be collected (Couper, 2017). In particular, non-probability online panels provide an easy-to-use and relatively inexpensive infrastructure to collect data quickly and have gained increasing popularity in the social sciences, despite concerns about their selection bias (e.g., Mercer et al., 2017). However, the shifting popularity towards self-administered web surveys may lead to a change in error sources relative to traditional interviewer-administered modes. While the absence of interviewers avoids interviewer effects (West & Blom, 2017) and lessens the risk of social desirability bias (Kreuter et al., 2008), negative consequences include the decreased ability to keep respondents engaged and focused on the response task, while ensuring that they comprehend the questions and provide thoughtful answers (Chang & Krosnick, 2010). A particularly problematic consequence is inattentive respondents (also called careless respondents, speeders, or insufficient effort respondents), which have long been identified as a significant and growing threat to data quality in web surveys (Meade & Craig, 2012).

The increasing proportion of inattentive respondents (see Ternovski et al., 2022, for a recent example) and its potential for introducing measurement errors have led to the development of a variety of prevention and detection methods. While research on prevention methods, such as commitment pledges (Conrad et al., 2017; Hibben et al., 2022), is still scarce, detection methods are abundant and range from rather obvious attention checks that instruct respondents to perform specific tasks (Oppenheimer et al., 2009) to sophisticated data analysis methods (e.g., Read et al., 2022; Ulitzsch et al., 2022). However, none of these approaches come without limitations. For example, attention checks can take various forms and previous studies have shown how different wordings (e.g., Silber et al., 2022) and task difficulty (e.g., Anduiza & Galais, 2017; Leiner, 2019; Shamon & Berning, 2020) can lead to differing proportions of respondents being flagged as inattentive. Some respondents might even fail attention checks on purpose (Liu & Wronski, 2018; Silber et al., 2022), or pass them by chance. Furthermore, while sophisticated data analysis approaches provide valuable insights into response behaviors that aid in identifying inattentive respondents, these methods can be challenging to implement and their results difficult to interpret.

We contribute to the literature on preventing and detecting inattentive responding in four ways. First, we evaluate the effectiveness of introducing a commitment pledge at the beginning of the survey with regard to respondents' performance on several indicators of attentiveness and data quality, and test whether pledge effects fade out over the course of the survey. Second, we compare two attention checks instructing respondents to either select a specific response option or select none of the response options, and use the results to test whether respondents likely pass the former attention check by chance. Third, we propose an easy-to-use cluster analysis method that uses screen-level durations to identify inattentive respondents with graphical representation of results. We show that the proposed method outperforms a related method with respect to differentiating attentive and inattentive respondents. Lastly, we explore the impacts of inattentive respondents on substantive results with regard to univariate analyses, regression analyses, and survey experiments.

The study results will inform researchers designing web surveys and analyzing data collected from them. In particular, the commitment pledge and attention check experiments will inform on the effectiveness of inserting additional items to the questionnaire for preventing and detecting inattentive responding. The proposed cluster analysis offers an additional quality control tool for researchers that provides insights into the prevalence of inattentive respondents in the data. As we also document the potential impacts of inattentive responding on substantive findings, the results will reinforce the importance of anticipating data quality issues during web survey design and analysis.

7.2 Inattentive responding

7.2.1 Prevention methods

The literature documents several approaches to preventing inattentive responding, including explicitly warning respondents against this behavior (e.g., Berinsky et al., 2016; Huang et al., 2012) or providing immediate feedback to the respondent when this behavior is observed (e.g., Conrad et al., 2017). However, (repeatedly) warning respondents might be perceived as intrusive for attentive respondents, and providing immediate feedback is technically challenging and requires thresholds to trigger feedback. An alternative approach is to use commitment pledges as tested by Conrad et al. (2017) and evaluated in-depth by Hibben et al. (2022) and Cibelli (2017). Building on previous findings from interviewer-administered surveys (e.g., Cannell et al., 1981), commitment pledges ask respondents whether they will commit to providing accurate data, with the question usually posed at the beginning of the survey. Respondents can either state that they will commit or not commit to doing so, though the proportion of respondents who do not commit is usually negligible (Cibelli, 2017; Conrad et al., 2017; Hibben et al., 2022), and often these respondents are filtered out of the survey immediately.

Conducting an experiment in a non-probability survey, Conrad et al. (2017) find that respondents who received a text message asking for their commitment were less likely to speed and highly educated respondents gave more accurate responses than those who received a neutral message containing no pledge. Similarly, Hibben et al. (2022) show that receiving a commitment pledge reduced item nonresponse, increased income reporting accuracy, and increased the total interview duration in a probability-based survey. However, they find no effect on straightlining (a lack of response differentiation in a set of same-scaled items, Yan, 2008) and a higher break-off rate (the proportion of respondents who start the survey but end the survey before completion) for respondents receiving the commitment pledge, which was driven by non-committing respondents. Lastly, Cibelli (2017, Chapter 2) investigates the provision of a commitment pledge in a probability-based survey. She finds that respondents who received a commitment pledge provided more accurate responses for some questions and had longer interview durations and more acquiescent responses. In contrast, she finds no difference for overall item nonresponse, straightlining, and social desirability. However, the commitment pledge increased item nonresponse for rather difficult questions and increased the break-off rate, presumably because committed respondents would rather not answer a question or break-off than provide inaccurate information. Relatedly, Clifford and Jerit (2015) showed that asking for the respondents' commitment at the beginning of the questionnaire reduces attention check failure but can induce socially desirable responding for some respondent groups (i.e., highly-educated respondents).

Building on these findings, we first evaluate whether commitment pledges have positive impacts on multiple measures of inattention and data quality (attention checks, straightlining, screen durations, item nonresponse, break-off rates) in a large-scale non-probability survey with a target population of young people prone to inattentive responding. Second, we extend previous research by investigating to which extent a commitment pledge at the beginning of the survey leads to higher data quality throughout the entire questionnaire. As the distance to the commitment increases, we hypothesize that any benefits from the commitment pledge will fade out as respondents progress further in the questionnaire.

7.2.2 Detection methods

Ex-ante survey design

Approaches to identify inattentive respondents in web surveys can be classified into ex-ante and ex-post methods (Meade & Craig, 2012). Ex-ante methods include measures researchers can take before the data are collected and involve adding attention check items to the questionnaire that flag inattentive respondents. Such direct attention checks include so-called bogus items (Meade & Craig, 2012) and instructed manipulation checks (IMCs) (Oppenheimer et al., 2009). Bogus items are illogical statements, such as “I am paid biweekly by leprechauns” (Meade & Craig, 2012, p. 5), with affirmative responses flagged as inattentive. IMCs can take various forms. Oppenheimer et al. (2009) use a long text on a separate screen that asks respondents to click on a specific field (e.g., the title of the screen), while other versions ask respondents to write specific words in an open-text field. Instructed response items (IRIs) are a special case of IMCs that instruct respondents to select a specific response option for an item embedded within a larger set of items (e.g., an item battery) on the same screen (Gummer et al., 2021; Meade & Craig, 2012).

Previous studies have shown that respondents flagged by IMCs provide lower-quality data and more measurement error than non-flagged respondents (e.g., Gummer et al., 2021; Huang et al., 2015; Meade & Craig, 2012; Oppenheimer et al., 2009; Silber et al., 2019). While these methods were designed to identify inattentive responding, they may also induce attentive responding and change response behavior to subsequent items (Oppenheimer et al., 2009). Hauser and Schwarz (2015) found spillover effects for complex tasks in an Amazon Mechanical Turk sample, however, these effects have not been replicated for standard survey questionnaires (Berinsky et al., 2014; Gummer et al., 2021; Hauser et al., 2016).

A significant disadvantage of these methods is that they provide only a snapshot of attentiveness, which is suboptimal given that inattentive behavior might change over the questionnaire. In addition, there is no guarantee that they will work as intended and identify the majority of inattentive respondents. Some attention checks might be too obvious to fail and previous research indicates that respondents might fail the checks on purpose due to confusion or the additional response burden (Liu & Wronski, 2018; Silber et al., 2022). Such checks often increase the response burden because the questionnaire is lengthened and attentive respondents might be annoyed by the additional items testing their attention (Silber et al., 2022).

Another concern for IRIs is that respondents may pass the checks by chance, for example, when providing a random response or straightlining the instructed response option. In the present study, we attempt to quantify the extent of this occurrence by running an experiment with two versions of the same IRI at the same position within an item battery, where one IRI asks respondents to provide a specific response and the second IRI asks respondents to provide no response. Some respondents in the first group might pass by chance if they randomly select one of the response options. In the second group, this is not possible as respondents should not select any response. Hence, we expect a higher proportion of respondents to pass the IRI for the first than for the second version. In addition to addressing these differences, we further contribute to the literature by investigating IRI spillover effects on subsequent items.

Ex-post data analysis

Ex-post detection methods are less intrusive but require researchers to analyze the collected data to identify likely inattentive cases. Such analysis approaches include the analysis of response data and paradata (Meade & Craig, 2012). Response data-based approaches often involve the analysis of same-scaled item batteries using neural networks (e.g., Melipillan, 2019; Welz & Alfons, 2024), straightlining/LongString indicators (e.g., Johnson, 2005; Meade & Craig, 2012), person-fit statistics (e.g., Emons, 2008), or the Mahalanobis distance (e.g., Meade & Craig, 2012). Most of these methods require setting a threshold above which respondents are deemed inattentive.

Paradata are data about the data collection process that are mainly generated during computer-assisted interviewing (Couper, 1998; Kreuter et al., 2010). Paradata-based detection methods often rely on timestamp data collected at the item-level, screen-level, or interview-level and researchers have used these data to develop multiple approaches and thresholds to identify presumably inattentive respondents (Matjašič et al., 2018). In the commercial survey industry, a widely employed threshold is one-third of the median total interview duration (McPhee et al., 2022). While easy to calculate, this rather arbitrary threshold cannot account for respondents taking breaks during the survey and differences in questionnaire length due to filter questions. As the threshold itself depends on the prevalence of very fast respondents, a risk of false positives in samples with few inattentive respondents and a risk of false negatives in samples with many inattentive respondents is introduced.

Previously proposed analysis approaches most relevant to the present study were developed by Read et al. (2022) and Ulitzsch and coauthors (Ulitzsch, Pohl, et al., 2024; Ulitzsch, Shin, & Lüdtke, 2024; Ulitzsch et al., 2022), who use mixture modeling approaches to disentangle attentive from inattentive respondents. Read et al. (2022) take the natural logarithm of the screen-level durations, apply principal component analysis for dimensionality reduction, and use mixture modeling to identify classes of "attentive", "slow inattentive", and "fast inattentive" respondents. The first two steps are required to limit the influence of extreme durations on cluster solutions. They find that the clusters of inattentive respondents pass fewer attention checks, have less consistent response patterns in item batteries, and attenuated treatment effects in a survey experiment. Ulitzsch et al. (2022) develop a latent mixture model that incorporates screen-level timestamp data and response data to disentangle attentive from inattentive respondents. However, their approach is computationally intensive and requires advanced statistical knowledge which might hinder its implementation in practice. Ulitzsch, Pohl, et al. (2024) extend this approach by accounting for response styles expected of attentive respondents. Lastly, Ulitzsch, Shin, and Lüdtke (2024) suggest using separate Gaussian mixture models to identify inattentive respondents at the screen level where the cluster with the shortest average duration is deemed as the inattentive cluster.

In the present study, we propose a further cluster-based approach to identify likely inattentive respondents. Specifically, we develop a non-parametric distance-based clustering approach that is robust to outliers, requires no researcher decisions on preprocessing, can increase the number of clusters efficiently, and provides an easy-to-interpret visualization of clustering results. The results for the distance-based approach are compared to multivariate mixture modeling, which is similar to the previously proposed approaches described above.

7.2.3 Impacts of inattentive responding on substantive analyses

Previous research on the influence of inattentive respondents on descriptive analyses suggests only minor effects on the estimates. For example, Anduiza and Galais (2017) show that excluding respondents based on IMCs can increase bias in univariate analyses while substantially reducing standard errors, suggesting a reduction of noise. For regression analyses, they find little evidence of biases due to inattentive respondents. Similarly, Greszki et al. (2015) find no strong differences in univariate or regression analyses when respondents flagged by timestamp-based speeder indices are excluded. Gummer et al. (2021) come to the same conclusions for excluding respondents who fail IRIs. However, inattentive respondents can have a larger influence on psychometric scores derived from item batteries (e.g., Huang et al., 2015; Maniaci & Rogge, 2014).

In (factorial) survey experiments, inattentive respondents will likely ignore the treatment or different experimental conditions and thus inattentive respondents assigned to different experimental groups will likely provide similar (random) responses. Hence, the estimated effect will be the intention-to-treat (ITT) effect rather than the average treatment effect. In this case, inattentive responding can attenuate treatment effect estimates (e.g., Kane, 2024; Read et al., 2022). However, excluding likely inattentive respondents flagged by post-treatment attention checks or paradata might induce post-treatment biases (e.g., Aronow et al., 2019; Montgomery et al., 2018). To counter such biases, Kane et al. (2023) suggest implementing a mock vignette before the actual vignette to obtain a measure of attentiveness, which can be used to calculate conditional average treatment effects (CATE) that depict the influence of respondents' attention on the experiment's results.

In the forthcoming study, we investigate the influence of likely inattentive respondents on univariate, regression, and experimental analyses. Given that the error induced by inattentive respondents depends on their response strategy (e.g., random responding, middle responding, acquiescence), their influence on univariate and regression analyses is a priori unknown, while treatment effects estimated for survey experiments should be attenuated.

7.3 Data sources

7.3.1 Study 1

The first study is a non-probability web survey, which was fielded in the U.S. in March 2022 by a contracted survey vendor. The target population was 16 to 25-year-olds. The main questionnaire topics were climate change and anxiety due to climate change (e.g., feelings about climate change, satisfaction with the government's work on climate change) and general demographics. We are aware of the general pitfalls of using non-probability-based samples for population-based inference (see Cornesse et al., 2020, for an overview), but refrain from engaging in this discussion as our focus is on response behavior in web-based samples. The sample consists of 15,990 respondents. The time spent on each screen was recorded for all respondents and measured in milliseconds. Each screen contained either informational text, a single item, or an item battery. A brief assessment of the validity of the screen-level timestamp data is described in Appendix 7.A.

We restrict the sample to smartphone respondents only (N = 13,758, 86 percent) as item batteries were depicted differently across devices (grid format for desktop respondents, scrolling for smartphone respondents), which prohibits a meaningful comparison with regard to timestamps. The results for the analysis of desktop respondents are provided in Appendix 7.B. Furthermore, we exclude four screens that depended on a preceding filter question. We keep durations for the welcome page and six screens containing only informational text. While inattention on these screens is not directly linked to the response process, skipping screens without reading their content indicates a tendency for inattentive behavior. In total, the sample consists of 742,932 durations for 54 screens and 13,758 respondents. The median total interview duration is 9.5 minutes, with a minimum of 3.1 minutes and a maximum of 1,499 minutes. In this sample, only 11 respondents are below the industry standard of one-third of the median duration.

7.3.2 Study 2

The second study was conducted in July and August 2023 with the same target population and topics as Study 1, and fielded by the same non-probability web survey vendor. Several questionnaire components were adopted from Study 1, though the order and number of questionnaire items changed. The sample consists of 6,002 respondents, though, again, we solely focus on smartphone respondents (N = 5,520, 92 percent). Commitment pledge and IRI experiments were implemented and are described in more detail in the next section. The median total interview duration was 7.3 minutes (minimum: 1.0 minutes; maximum: 1,257 minutes). Only 50 respondents are below the industry threshold of one-third of the median duration.

7.4 Methods and experimental designs

7.4.1 Clustering timestamp data

Several timestamp data characteristics often complicate the application of methods to identify inattentive respondents. First, duration distributions are often heavily skewed and have extreme outliers, for instance, due to respondents taking breaks and doing something else during the survey (e.g., Höhne et al., 2020; Sendelbah et al., 2016). Second, screens differ in length (amount of questionnaire text, number of items per screen, and number of response options). Third, for some screens (e.g. socio-demographic questions) both attentive and inattentive respondents are expected to be very fast. Clustering approaches that require dimensionality reductions (e.g., Read et al., 2022) might mask such differences across screens. Below, we describe a distance-based clustering approach that overcomes these issues.

Distance-based clustering approach

The proposed distance-based approach consists of two steps. First, we handle the outlier problem by scaling the duration data. For each screen s, the durations are sorted and the relative rank of each respondent i ($rank(dur_{is})/N$) is assigned. Hence, fast respondents receive relative ranks close to zero, whereas slow respondents receive ranks close to one. As a consequence, extreme outliers are less influential. While this step comes with a loss of information, we still know whether respondents spent more or less time on the respective screen compared to other respondents.

Second, we use the relative ranks as inputs to a hierarchical agglomerative cluster analysis with Ward's linkage and Euclidean distance (Kaufman & Rousseeuw, 1990). Both internal and external criteria are used to determine the most adequate number of clusters. We use 22 cluster validity indices as internal criteria and determine the recommended number of clusters based on the modal number of clusters (Charrad et al., 2014). The respective indices are listed in Appendix 7.C. As external criteria, we rely on a straightlining indicator which is a frequently used measure of inattention and strong satisficing (Kim et al., 2019; Krosnick, 1991; Meade & Craig, 2012).

We illustrate the clustering approach using the Study 1 data and calculate the straightlining indicator for four separate item batteries (see Appendix 7.D for the question wordings). The first item battery (Q5) contains 16 items on how climate change makes respondents feel (e.g. sad, anxious, powerless) with response options *not at all - a little - moderately - very much - extremely.* The second battery (Q7) asks what climate change makes respondents think (e.g. the future is frightening, I'm hesitant to have children) with the same response options as Q5. The third item battery (Q18) contains 9 items about their beliefs regarding the US Government's actions on climate change (e.g. acting in line with climate science, failing young Americans) with Yes-No response options. Lastly, the fourth item battery (Q19) contains 14 items on how they feel about the US government's response to climate change (e.g. hopeful, reassured, angry) with the same response options as for Q5 and Q7. All four item batteries contain items that contradict each other if the same response option is selected (e.g., Q18 contains an item on whether the respondent believes that the US Government is "trustworthy", another item asks whether the respondent believes that the US Government is "lying about the effectiveness of the actions they're taking"). Thus, any straightlining would produce rather inconsistent answers indicative of potential inattentive responding. As the cluster solutions with increasing numbers of clusters are nested models, we assess the external criteria by fitting logistic regressions for all cluster solutions and straightlining indicators separately and use AIC and BIC to determine the number of clusters that best separates straightlining from non-straightlining respondents. In total, we consider up to 15 clusters to ensure the detection of a variety of screen duration patterns over the interview.[1]

Model-based clustering approach

For comparison with our proposed approach, we implement a model-based clustering approach. Specifically, we follow previous research and use multivariate Gaussian mixture modeling (Read et al., 2022; Ulitzsch, Pohl, et al., 2024; Ulitzsch, Shin, & Lüdtke, 2024; Ulitzsch et al., 2022). To account for the skewness of the duration data, we take the natural logarithm of the durations (Ulitzsch, Shin, & Lüdtke, 2024). However, we refrain from preprocessing steps such as dimensionality reduction (Read et al., 2022) or defining cut-off values for extreme durations (Ulitzsch et al., 2022). We do so to limit the potential influence of researcher degrees of freedom. Nonetheless, we provide results for the model with durations top-coded at the 99th percentile for each screen in Appendix 7.E. Since Gaussian mixture model solutions depend on the provided starting values (e.g., obtained via k-means clustering or random draws), we initialize the model using a Hierarchical agglomerative clustering model. With regard to the number of clusters, internal and external criteria are used. As internal criteria, the AIC and BIC are used. As external criteria, we also rely on the straightlining indicators described above to determine the number of clusters that best separate straightlining from non-straightlining. As before, up to 15 clusters are considered.

[1] The information criteria do not indicate meaningful additional cluster solutions when more than 15 clusters are considered.

7.4.2 Commitment pledge

The commitment pledge experiment was implemented in Study 2. After the welcome page and two questions on the respondent's age and state of residency, respondents were randomized to receive a screen that showed the commitment pledge text (N = 2,747), whereas the other half of the sample (N = 2,773) did not receive the pledge and continued with the questionnaire. The exact pledge text followed Hibben et al. (2022, p. 17) and Geisen (2022): "We care about the quality of our survey data. For us to get the most accurate measures of your opinions, it is important that you provide thoughtful answers to each question in this survey. To do this, it is important to think carefully about each question, search your memory, and take time to answer. Are you willing to do this?". Compared to previous approaches (e.g., Conrad et al., 2017; Hibben et al., 2022), respondents were only provided a checkbox to commit and only allowed to continue the questionnaire after checking the box. Hence, the experiment explicitly focuses on having agreed to provide thoughtful responses. Whether respondents who received the pledge provide higher-quality data is first tested by comparing the proportions of respondents failing an attention check and the proportion of item nonresponse across groups. Extending on previous literature, we also investigate whether commitment effects fade out over the course of the questionnaire. Using 16 indicators of straightlining distributed over the questionnaire, we assess whether potential data quality differences change as respondents progress through the questionnaire. To estimate the effects of the commitment pledge, we follow Gomila (2021) and fit the following linear regression model:

$$y_{it} = \beta_0 + \lambda_t + \sum_{t=1}^{T} \beta_t d_i + \varepsilon_{it} \tag{7.1}$$

where y_{it} denotes whether respondent i failed (= 1) or passed (= 0) straightlining indicator t. β_0 is the constant, λ_t is the estimate for the control group for each indicator, and β_t denote the treatment effects for each indicator. d_i denotes whether the respondent receives the pledge or not. As we observe multiple indicators for each respondent, standard errors are clustered at the respondent level.

Lastly, we assess the impact of the commitment pledge on the time spent on each screen. Similar to model 7.1, we fit a linear regression model with the natural logarithm of the duration observed for each screen t and respondent i as the dependent variable. The estimated coefficients β_t denote the differences between the experimental groups for each screen. As before, standard errors are clustered at the respondent level.

$$ln(duration_{it}) = \beta_0 + \lambda_t + \sum_{t=1}^{T} \beta_t d_i + \varepsilon_{it} \tag{7.2}$$

7.4.3 Instructed response item (IRI)

Study 2 contained two IRIs for most respondents. The first IRI was embedded within the item battery on what climate change makes the respondent think (Q7). Each respondent received the IRI stating "Please select 'Very much' to show us that you are paying attention". Its position was fixed, while the order of the other items on the battery was randomized. The second attention check was implemented as part of an experimental manipulation. It was embedded within an item battery measuring the extent to which different factors contribute to the respondent's feelings about climate change split across two screens (see Appendix 7.F). One-third of respondents received no IRI, one-third received an IRI which instructed respondents "To show you have read this sentence please mark 'very much'" (Response-IRI), and one-third received an IRI stating "To show you have read this sentence please leave the question blank" (Blank-IRI). In addition to comparing the proportions of respondents who failed these attention checks, we follow previous literature (Berinsky et al., 2014; Gummer et al., 2021; Hauser & Schwarz, 2015; Hauser et al., 2016) and evaluate whether the IRIs affect response behavior to subsequent items (i.e., spillover effects). To estimate the spillover effects, we fit equation 7.1 with the IRI treatment status defined by d_i and the straightlining indicators as dependent variables.

7.5 Results

7.5.1 Clustering timestamp data

In this section, the results of the clustering approaches for the Study 1 data are described. The results using the Study 2 data are provided in Appendix 7.G.

Distance-based approach

For the distance-based clustering approach based on the relative ranks, the internal criteria suggest a 2-cluster solution (11 indices suggest a 2-cluster solution, 7 indices suggest a 3-cluster solution). The external criteria suggest a 7-cluster solution on which the AIC and BIC agree for all four item batteries (see Appendix 7.H). Figures 7.1 and 7.2 depict the results for both cluster solutions. The x-axis denotes the order of the screens in the questionnaire and the y-axis denotes the proportion of respondents within each cluster that belong to the respective decile. The relative ranks are aggregated into deciles to simplify interpretation.

The 2-cluster solution assigns respondents either to the rather fast Cluster 2 (on average 43.7 percent of respondents lie within the two fastest deciles and only 6.5 percent lie within the two slowest deciles) or a rather slow Cluster 1 (on average only 7.6 percent of respondents lie within the two fastest deciles and 27.0 percent lie within the two slowest deciles). While the cluster compositions are relatively stable throughout the questionnaire, we observe a slight change from screen 37 onward, where Cluster 1 gets slightly faster and Cluster 2 gets slightly slower. The socio-demographic items start at this point in the questionnaire for which even attentive respondents are expected to provide rather fast responses.

Figure 7.2 shows the decile compositions for the 7-cluster solution, which provides a more fine-grained depiction of screen time patterns in the data. Cluster 3 accounts for 15 percent of respondents and is characterized by very short durations. 36.4 percent of all durations are in the first decile. Except for a slight decrease in relative durations at the beginning and the shift for the socio-demographic questions, this pattern is stable over the entire questionnaire. All 11 respondents flagged by the industry standard of one-third of the median completion time lie within this cluster. Cluster 5 shows similar signs of very short durations, though to a lesser extent. Note that Clusters 3 and 5 are subsets of Cluster 2 in the 2-cluster solution. Clusters 4 and 6 show a

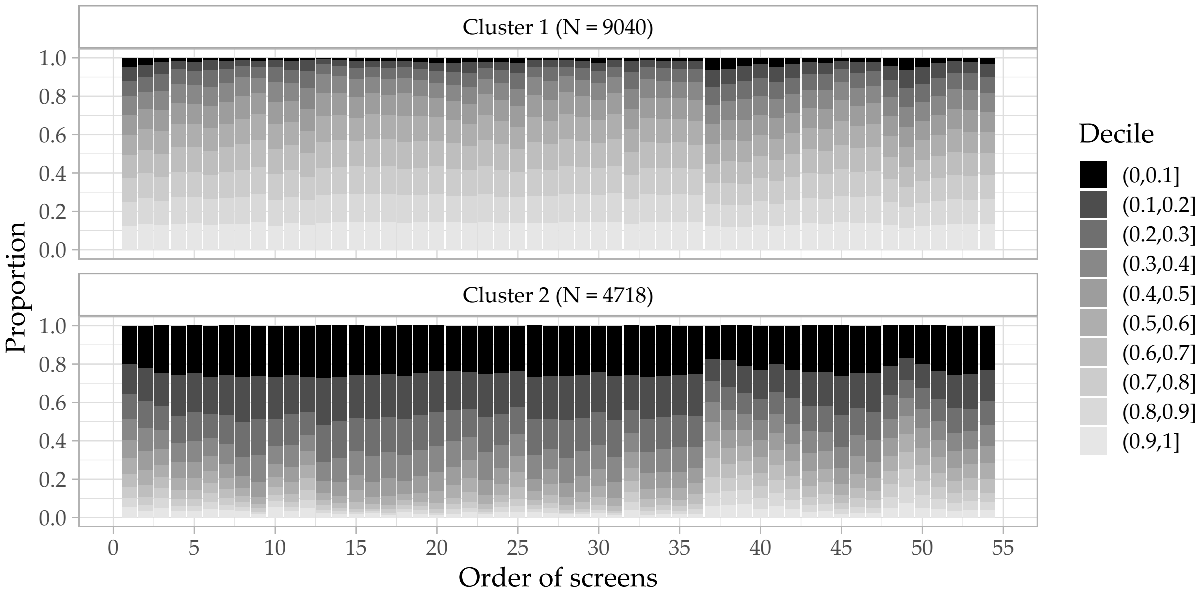


Figure 7.1: Duration decile composition of 2-cluster solution.

lower prevalence of very short durations but seem to speed up in the middle of the questionnaire. On the contrary, Clusters 1 and 2 seem to speed up for the socio-demographic questions. Cluster 7 is a slow cluster, with more than 20 percent in the slowest decile throughout the questionnaire. In summary, Clusters 3 and 5 contain the respondents with the shortest relative durations, whereas Clusters 4 and 6 speed up for the middle part of the questionnaire and Clusters 1 and 2 for the socio-demographic section at the end.

Next, we compare the cluster solutions with regard to the prevalence of straightlining. Table 7.1 reports the proportion of straightlining by item battery. As the 7-cluster solution was determined based on separating straightlining from non-straightlining, we expect larger discrepancies for the 7-cluster solution. Since straightlining might also arise from a lack of cognitive skills (e.g., Kim et al., 2019), a perfect classification of straightlining is unlikely. For the 2-cluster solution, the prevalence of straightlining is at least twice as high in Cluster 2 than in Cluster 1 for all item batteries except Q19. Hence, the straightlining indicator supports the notion that Cluster 2 is more prone to inattentive behavior. For the 7-cluster solution, several notable differences arise. First, Cluster 3 – the cluster with the shortest relative durations – never has the largest proportion of straightlining across all clusters, although its values are generally high. Second, Cluster 7 has the smallest proportions of straightlining across all item batteries. Third, Clusters 1 and 2 which are rather fast for the sociodemographic items have minor proportions of straightlining. Fourth, Clusters 4 and 6 which speed up during the questionnaire have proportions of straightlining similar to Clusters 3 and 5. This is particularly noteworthy as they are assigned to the slower cluster in the 2-cluster solution. In addition, Cluster 4 has more than 10 percent of durations in the slowest decile throughout most of the questionnaire, which is in line with *slow inattentive responding* proposed by (Read et al., 2022). In this case, increasing the number of clusters and consulting external criteria results in the detection of variation throughout the questionnaire which uncovers special clusters of likely inattentive respondents.

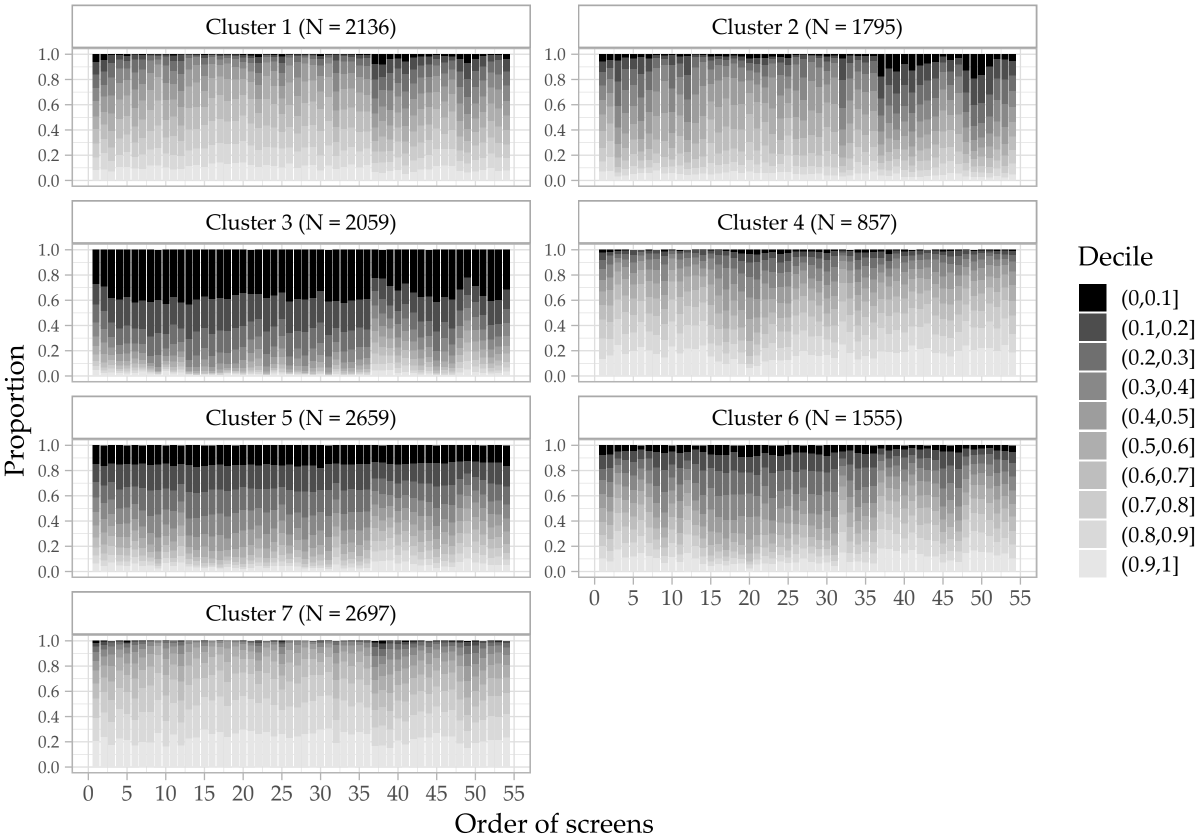


Figure 7.2: Duration decile composition of 7-cluster solution.

Table 7.1: Proportion of straightlining across clusters, distance-based approach.

Solution	Cluster	Q5	Q7	Q18	Q19
2	Cluster 1 (N = 9040)	0.059	0.071	0.106	0.103
2		[0.054,0.064]	[0.066,0.077]	[0.100,0.113]	[0.097,0.110]
2	Cluster 2 (N = 4718)	0.122	0.146	0.252	0.147
2		[0.113,0.132]	[0.136,0.156]	[0.240,0.265]	[0.137,0.157]
7	Cluster 1 (N = 2136)	0.032	0.033	0.056	0.061
7		[0.025,0.040]	[0.026,0.041]	[0.046,0.066]	[0.051,0.072]
7	Cluster 2 (N = 1795)	0.036	0.046	0.070	0.071
7		[0.028,0.046]	[0.036,0.056]	[0.059,0.083]	[0.060,0.084]
7	Cluster 3 (N = 2059)	0.119	0.145	0.271	0.130
7		[0.105,0.133]	[0.130,0.160]	[0.252,0.291]	[0.116,0.146]
7	Cluster 4 (N = 857)	0.109	0.133	0.207	0.221
7		[0.088,0.131]	[0.111,0.158]	[0.180,0.235]	[0.193,0.250]
7	Cluster 5 (N = 2659)	0.125	0.147	0.237	0.159
7		[0.113,0.138]	[0.134,0.161]	[0.221,0.254]	[0.145,0.174]
7	Cluster 6 (N = 1555)	0.144	0.214	0.276	0.237
7		[0.127,0.162]	[0.193,0.235]	[0.254,0.299]	[0.216,0.259]
7	Cluster 7 (N = 2697)	0.029	0.016	0.041	0.044
7		[0.023,0.036]	[0.012,0.022]	[0.034,0.049]	[0.036,0.052]

Notes: 95 percent confidence intervals in brackets.

Model-based approach

The model-based clustering approach suggests a 5-cluster solution for the internal criteria and a 4-cluster solution for the external criteria. For the 5-cluster solution, posterior probabilities for cluster membership exceed 90 percent for 85.9 percent of respondents. For the 4-cluster solution, posterior probabilities for cluster membership exceed 90 percent for 89.3 percent of respondents. Respondents are assigned to clusters based on the maximum posterior probability.

Table 7.2 reports the proportions of straightlining across clusters for the 4- and 5-cluster solution. For the 5-cluster solution, Cluster 3 has the highest proportion across all item batteries, whereas Cluster 1 has the lowest proportion across all item batteries. The remaining clusters are between these extremes, with proportions above 7 percent for Q5, above 9 percent for Q7, above 15 percent for Q18, and above 10 percent for Q19. For the 4-cluster solution, the results are similar. Cluster 3 has the highest proportions, Cluster 1 has the lowest, and Clusters 2 and 4 are in-between. In sum, the model-based approach identifies several clusters which fail to separate straightlining from non-straightlining.

Table 7.2: Proportion of straightlining across clusters, model-based approach.

Solution	Cluster	Q5	Q7	Q18	Q19
5	Cluster 1 (N = 3078)	0.017	0.007	0.025	0.031
5		[0.013,0.022]	[0.005,0.011]	[0.020,0.031]	[0.025,0.038]
5	Cluster 2 (N = 2300)	0.091	0.111	0.182	0.140
5		[0.080,0.104]	[0.098,0.124]	[0.166,0.198]	[0.126,0.155]
5	Cluster 3 (N = 2750)	0.136	0.182	0.285	0.184
5		[0.123,0.149]	[0.168,0.197]	[0.268,0.302]	[0.170,0.199]
5	Cluster 4 (N = 1917)	0.077	0.096	0.161	0.108
5		[0.066,0.090]	[0.083,0.110]	[0.144,0.178]	[0.094,0.123]
5	Cluster 5 (N = 3713)	0.087	0.099	0.152	0.133
5		[0.078,0.097]	[0.090,0.110]	[0.140,0.164]	[0.122,0.144]
4	Cluster 1 (N = 3677)	0.020	0.012	0.029	0.035
4		[0.016,0.025]	[0.008,0.016]	[0.024,0.035]	[0.029,0.041]
4	Cluster 2 (N = 2566)	0.086	0.106	0.176	0.137
4		[0.076,0.098]	[0.094,0.118]	[0.161,0.191]	[0.124,0.151]
4	Cluster 3 (N = 3288)	0.134	0.179	0.283	0.183
4		[0.122,0.146]	[0.166,0.192]	[0.268,0.299]	[0.170,0.196]
4	Cluster 4 (N = 4227)	0.088	0.101	0.156	0.129
4		[0.080,0.097]	[0.092,0.111]	[0.146,0.168]	[0.119,0.139]

Notes: 95 percent confidence intervals in brackets.

Comparison of clustering approaches

To compare the two clustering approaches, we first assess to which extent the distance-based and the model-based approaches assign respondents to the same clusters. We consider the 7-cluster solution for the distance-based approach and the 5-cluster approach for the model-based approach. The Adjusted Rand Index is 0.096, which signifies low agreement between the two approaches. Considering only the most suspicious clusters (for both approaches, Cluster 3), we find that 95.9 percent of respondents assigned to Cluster 3 in the model-based approach are assigned to the suspicious Clusters 3, 4, 5, and 6 in the distance-based approach. On the contrary, 64.3 percent of respondents assigned to Cluster 3 in the distance-based approach are assigned to Cluster 3 in the model-based approach, whereas the remaining majority (28.6 percent) is assigned to the “mixed” Clusters 4 and 5.

Given the disagreement between both approaches, we evaluate which cluster solution performs better with regard to separating straightlining from non-straightlining. We fit separate logistic regressions with straightlining indicators as dependent variables and the cluster assignment variables for each item battery and cluster solution. Across all item batteries, the BIC suggests a better model fit for the distance-based approach (results are reported in Appendix 7.H). Hence, the distance-based approach performs better with regard to disentangling likely attentive respondents from inattentive respondents.

7.5.2 Commitment pledge

We estimate the effect of receiving a commitment pledge on various data quality indicators. Table 7.3 reports the effects on failing the attention check and any item nonresponse. There is no statistically significant difference between the experimental groups for any of the indicators. Similarly, there is no significant difference with regard to break-offs (24.4 percent for the no pledge group; 25.0 percent for the pledge group). Appendix 7.I provides more insights into the break-offs in Study 2.

Table 7.3: Differences in data quality measures across pledge conditions.

	No pledge	Pledge	Difference	z-statistic	p-value	N
Failed attention check	0.421	0.440	0.020	1.473	0.141	5520
Any item nonresponse	0.466	0.458	-0.008	-0.594	0.553	5520

Figure 7.3 shows the differences between the predicted proportion of straightlining for respondents who received the pledge and those who did not receive the pledge for 16 screen-level item batteries (e.g., for the first item battery, the proportion of straightlining is 2 percentage points higher among respondents who received the commitment pledge compared to respondents in the control group). Contrary to expectations, we find slightly higher proportions of straightlining (around 2 to 3 percent) among respondents who received the commitment pledge for most item batteries. Given that we do not even observe a positive pledge effect with regard to data quality, fading-out effects are not present.

Lastly, we investigate the pledge effect on screen durations. Figure 7.4 shows the estimated difference between the pledge and no pledge respondents, with the earlier screen asking for the respondent’s age (2 screens before the commitment pledge) as the reference category. Respondents who receive the pledge are slightly faster ($\sim$ 1 second) for the screen succeeding the pledge screen. This effect does not persist for later screens. A rationale for this unexpected difference is that respondents who received the pledge are already in a state of speeding and are thus also faster on the next screen. Generally, respondents are very fast on the pledge screen with a median duration of 3.82 seconds with 13.8 percent of respondents taking less than two seconds. The estimated minimum time to read the commitment pledge based on the number of characters (Andreadis, 2021) is 6.9 seconds. Hence, many respondents likely did not read the (entire) commitment pledge, which might explain the lack of pledge effects.

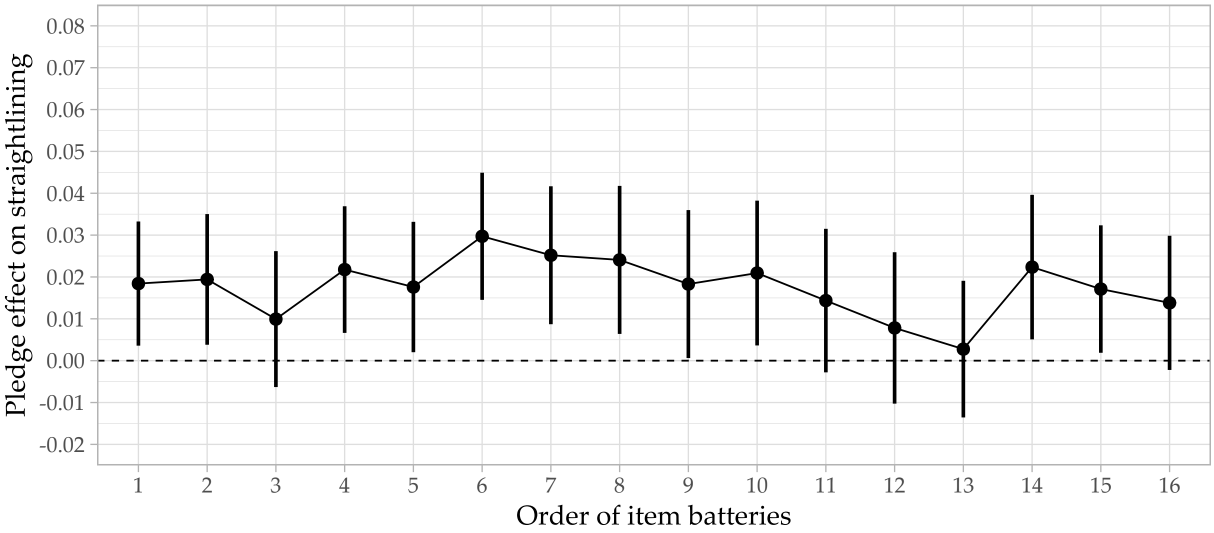


Figure 7.3: Pledge effects on straightlining over the questionnaire (with 95 percent confidence intervals).

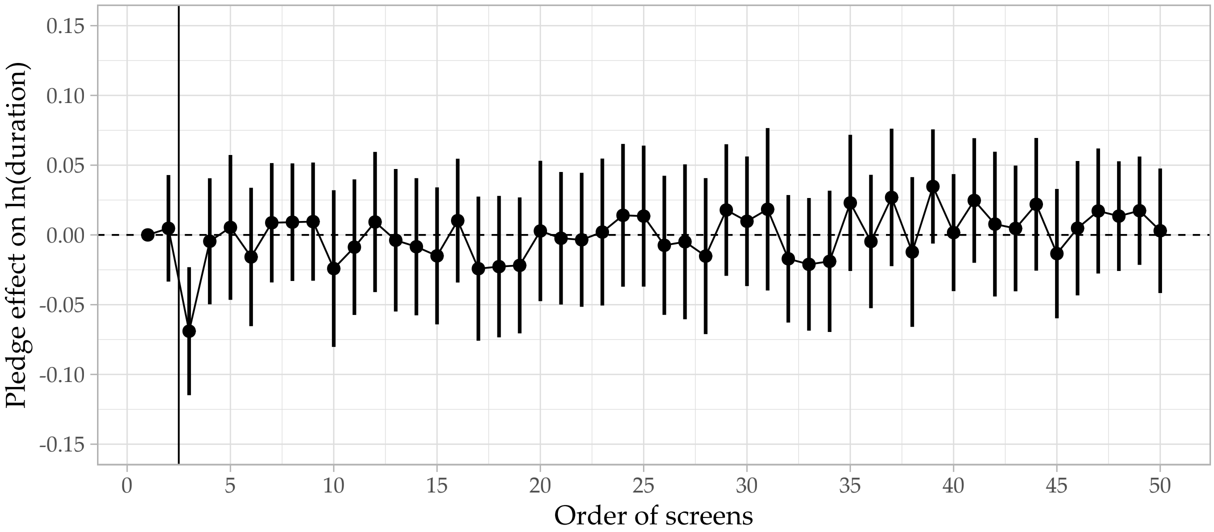


Figure 7.4: Pledge effects on screen durations over the questionnaire (with 95 percent confidence intervals). The vertical line denotes the position of the pledge.

7.5.3 Instructed response item (IRI)

For the analysis of the IRIs, we exclude respondents who skipped any item on the preceding screen (4.2 percent) to ensure that these respondents do not bias the results, as skipping items may lead to respondents being falsely flagged as inattentive for the Response-IRI group and attentive in the Blank-IRI group. Table 7.4 reports the proportion of respondents in both IRI groups who fail the respective IRI. 46.0 percent failed to select the requested response option, whereas 62.8 percent failed to leave the response blank.[2] Assuming that the true proportion of inattentive respondents is 62.8 percent and that inattentive respondents provide random responses, the expected proportion of inattentive respondents selecting the correct requested response option on a 5-point scale by chance is $1/5 \times 62.8 \approx 12.6$ percent. The observed difference of 16.8 percentage points is only slightly higher than this expected value. Hence, the results are in line with expectations for random responding by inattentive respondents. Furthermore, for respondents who failed the Blank-IRI, the proportion of respondents selecting "Very much" on the eight remaining items in the battery where the IRI experiment is placed varies between 16.9 and 19.6 percent (i.e., close to 20 percent). Again, these response patterns are consistent with random responding.

Concerning the median time spent on the screen of the IRI experiment, we find that the control group (9.0 seconds), the failing respondents in the Response-IRI group (8.0 seconds), and the failing respondents in the Blank-IRI group (8.0 seconds) have similar durations. However, the respondents who passed the check in the Response-IRI group (18.0 seconds) and the respondents who passed in the Blank-IRI group (23.4 seconds) differ by more than five seconds in their median durations. A potential explanation is that the Blank-IRI is more difficult to understand and thus requires more time. In this case, we would also expect that some respondents noticed the Blank-IRI but failed it due to its difficulty which would imply that failing respondents took longer in the Blank-IRI group, which is not what we find. Hence, the screen durations support the notion that some respondents who pass the Response-IRI do so by chance.

Table 7.4: Differences failing rates across experimental conditions.

	Response-IRI	Blank-IRI	Difference	z-statistic	p-value	N
Failed attention check	0.460	0.628	0.168	10.142	0.000	3523

Concerning spillover effects, Figures 7.J1 and 7.J2 in the Appendix depict the differences between the treatment groups and the control group with regard to straightlining and screen durations. In both figures, the item battery or screen duration preceding the experiment screen are used as reference categories and the depicted estimates report the difference between the respective treatment group and the control group. For straightlining, we do not observe any impact on succeeding item batteries. For the durations, we find that both treatment groups take slightly longer on the screen after the attention check, however, this effect does not prevail for further screens.

7.6 Impacts on substantive analyses

To analyze the impacts of inattentive responding on univariate and regression analyses, we rely on the Study 1 data and the cluster analysis results. For the impacts on analyzing survey experiments, we reanalyze the data used by Read et al. (2022).

[2] If respondents who skipped any item on the preceding screen are not excluded, the percent who failed the respective IRI is 46.9 in the Response-IRI group and 61.8 in the Blank-IRI group.

7.6.1 Univariate analysis

To assess the influence of the likely inattentive respondents on substantive research results, we calculate average values for all Likert-scaled items and items with Yes-No response options by cluster. As for the timestamps analysis, we exclude items subject to preceding filters. The response options are re-scaled ranging from zero to one to enable comparisons across differently scaled items. Figure 7.5 depicts these averages item-by-item for the distance-based 7-cluster solution. The same figure for all other cluster approaches and solutions is provided in Appendix 7.K.

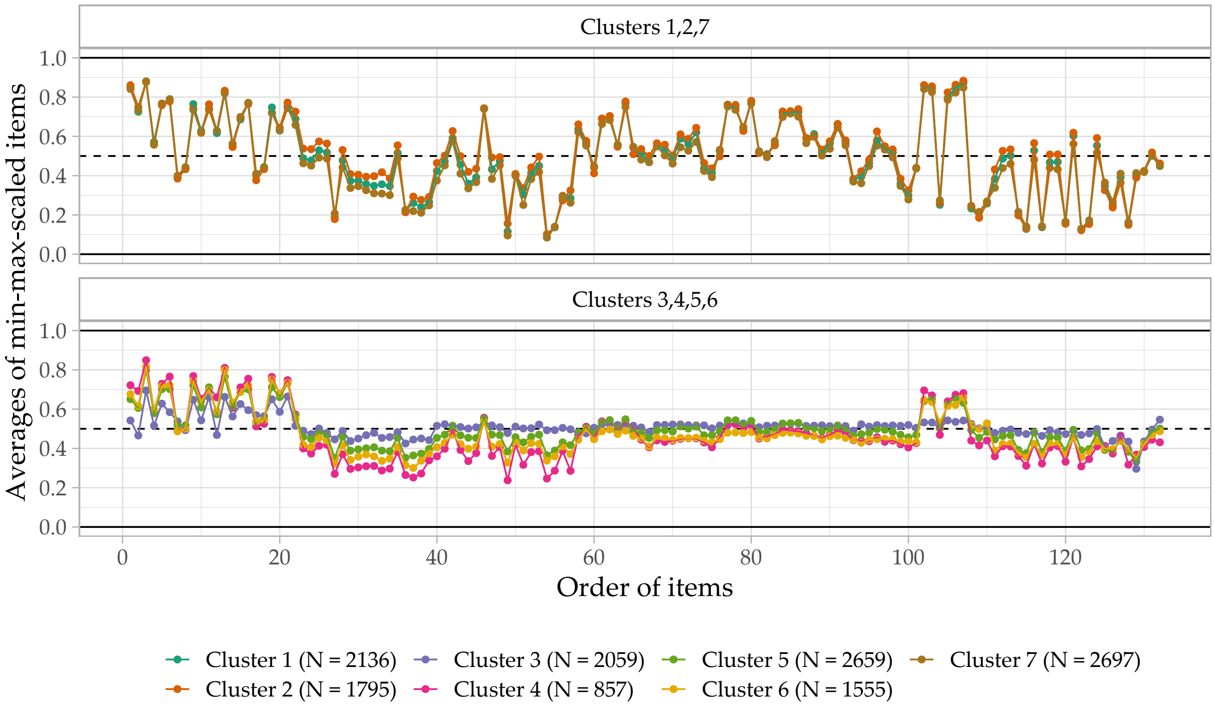


Figure 7.5: Development of scaled average responses by cluster over the questionnaire.

The average responses differ widely across clusters. For the likely attentive clusters (i.e., 1, 2, 7), the averages vary across items, with some averages close to the highest response option and others close to the lowest response option. These patterns are consistent across clusters. For the likely inattentive clusters (i.e., 3, 4, 5, 6), however, the averages show less variation and vary around 0.5, i.e., the mean of the response options. Furthermore, this lack of variance across items differs across clusters as Cluster 3 deviates from 0.5 by more than 5 percentage points in only 18.2 percent of all items, whereas the other clusters show substantially more variation (e.g., 79.5 percent in Cluster 1). For items with Yes-No response options, averages close to 0.5 indicate random responding. On Likert-scaled items, averages close to the scale midpoint can result from response strategies, such as random responding and middle responding. In our case, middle responding plays only a minor role. For example, for Cluster 3 (i.e., the most extreme cluster) the proportion of respondents selecting the midpoint on 5-point scales is 35.2 percent. Instead, a mixture of inattentive response strategies (i.e., straightlining of varying scale points, random responding) results in an equal distribution of responses over the response options. These findings also do not support the argument that inattentive respondents try to manipulate substantive outcomes, for example, by intentionally stating that they do not believe in climate change. To illustrate this point, while only 32.2 percent of respondents in Cluster 3 stated that they are moderately or very sure climate change is happening (compared to 76.0 percent in the likely attentive clusters), 62.3 percent of respondents in Cluster 3 rather contradictorily stated that their feelings about climate change affect their daily life in a moderately, very much, or extremely negative way (compared to 32.5 percent in the likely attentive clusters).

7.6.2 Regression analysis

To show the potential impacts of inattentive responding on regression coefficients, we investigate socio-demographic predictors of climate change belief in the data (see Hornsey et al., 2016). Table 7.5 reports the results of a simple logistic regression model with a binary dependent variable on whether the respondent is sure that climate change is happening (= 1 if the respondent is moderately or very sure it is happening; = 0 if the respondent is very, moderately, or slightly sure it is not happening, doesn't know, or is slightly sure it is happening) and the respondent's age, area, education, gender, and party identification as explanatory variables. The model is fitted for the full sample and the sample without the likely inattentive Clusters 3, 4, 5, and 6. We observe multiple differences with regard to the size of the coefficients and their statistical significance. Although the coefficient for age is small, its size increases slightly. Excluding the likely inattentive clusters leads to statistically significant differences between areas. Concerning education, the coefficient for the highest-educated respondents is no longer statistically significant. The difference between males and females is reduced by more than 50 percent and the difference between Republicans and Democrats increases substantially. Notably, despite reducing the sample size by more than 50 percent, confidence intervals remain similar in size.

Table 7.5: Average marginal effects for logistic regression with certainty about climate change as dependent variable, study 1 data.

	Full sample	Without clusters 3,4,5,6
Age	-0.002 [-0.005, 0.001]	-0.007 [-0.011,-0.004]
Area: Suburban	-0.010 [-0.030, 0.011]	0.024 [0.000, 0.047]
Area: Urban	-0.006 [-0.028, 0.017]	0.028 [0.002, 0.055]
High school or less	-0.058 [-0.081,-0.035]	-0.063 [-0.091,-0.035]
Some graduate/master/doctoral degree	-0.060 [-0.098,-0.021]	0.022 [-0.028, 0.073]
Vocational/some college	-0.004 [-0.028, 0.021]	0.001 [-0.027, 0.029]
Female	0.089 [0.072, 0.106]	0.032 [0.010, 0.053]
Party: Independent	-0.087 [-0.108,-0.067]	-0.119 [-0.141,-0.097]
Party: Other/No response	-0.135 [-0.159,-0.110]	-0.138 [-0.167,-0.109]
Party: Republican	-0.271 [-0.294,-0.248]	-0.375 [-0.405,-0.345]
N	13737	6626

Notes: 95 percent confidence intervals in brackets. Reference category for area is rural, for education it is Associates or bachelor's degree, for party it is Democrat.

7.6.3 Survey experiments

To illustrate the consequences of inattentive responding for the analysis of survey experiments, we follow Read et al. (2022) and Kane et al. (2023) who suggest stratifying treatment effects by categories of inattention. We use the same data as Read et al. (2022) available online (Read et al., 2021) to obtain a comparison with their modeling approach and because their data contains the famous "Asian disease" experiment by Tversky and Kahneman (1981) (see Table 7.L1 for the questionnaire text and Druckman, 2001, for an in-depth discussion of the experiment and effect sizes). We use all screen durations before the actual experiment to avoid post-treatment bias (30 of 40 screens). Four attention check questions are employed as external indices which suggest either a 4- or 10-cluster solution (see Figures 7.M1, 7.M2, and 7.M3). Figure 7.6 depicts the CATEs by cluster for both solutions. The green line denotes the estimated effect for the full sample. These results replicate the finding of highly heterogeneous effects by Read et al. (2022). However, our distance-based clustering approach allows for a finer distinction between respondent types which leads to larger differences between clusters (for Read et al., 2022, the CATEs are 0.36, 0.21, and 0.29). As we still use the majority of screens, we apply a simulation approach to test whether fewer screens would still allow for identifying the heterogeneity just as well. Figure 7.M4 in the Appendix shows the CATEs based on 100 variable draws of size 5, 10, 15, and 20, where similar changes between clusters can be seen, as in Figure 7.6. Hence, our proposed clustering approach may serve the same purpose as the mock vignettes developed by Kane et al. (2023) without increasing response burden and requiring researchers to design appropriate mock vignettes for their experiments.

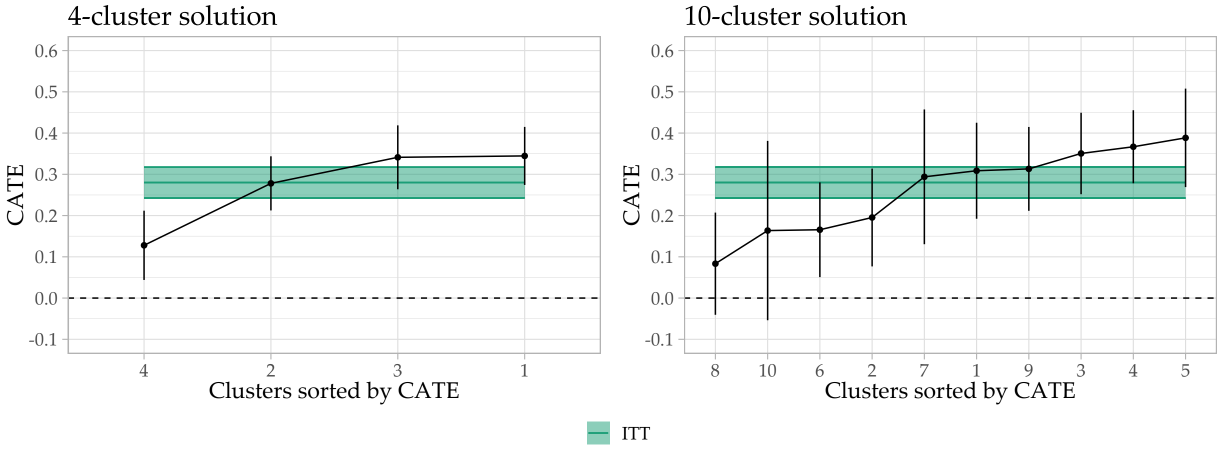


Figure 7.6: Conditional average treatment effects by cluster (with 95 percent confidence intervals), data from Read et al. (2022).

7.7 Discussion

This study carried out multiple experiments and analyses on preventing and detecting inattentive respondents in web surveys. We found that requesting respondents to commit to providing high-quality responses had no effects on data quality, which is in contrast to previous research (Cibelli, 2017; Clifford & Jerit, 2015; Conrad et al., 2017; Hibben et al., 2022). As identified by an analysis of the time spent on the commitment pledge screen, a potential explanation is that respondents do not even read the commitment pledge text. Furthermore, we showed that the proportion failing an instructed response item (IRI) depends on its specific instruction and may lead to substantial differences in flagged respondents. In our case, instructing respondents to select a specific response option led to 16.8 percentage points fewer respondents being flagged compared to when the instruction was to leave the item blank, suggesting that many respondents who passed the IRI instructing a specific response did so by chance. We further developed a timestamp-based clustering approach that allows for identifying likely inattentive respondents and classifying respondents who differ in their response behavior, such as those who speed up over the course of the questionnaire (Bowling et al., 2021). The proposed distance-based clustering approach performed better than a mixture modeling approach similar to previously proposed methods in the literature with regard to multiple straightlining indicators. Lastly, our results showed that likely inattentive respondents introduce biases in univariate, regression, and experimental analyses. The magnitude of the biases in univariate and regression analyses exceeded those found in previous studies (e.g., Anduiza & Galais, 2017; Greszki et al., 2015; Gummer et al., 2021).

These results have several implications for survey practice. In particular, in surveys prone to inattentive responding solely adding commitment pledges may not be enough to prevent inattention. To ensure that respondents at least read the pledge, practitioners may add warnings when respondents proceed too fast on the respective screen (Conrad et al., 2017). Given that the IRI results varied substantially across instructions, we recommend instructing respondents not to provide any response to the attention check item to ensure that random responding does not introduce false negatives. However, other potential problems associated with attention checks, such as increased response burden, deliberate defiance, and signaling distrust of respondents should be kept in mind when implementing attention checks (Silber et al., 2022). The proposed clustering method requires no such considerations. As the collection of timestamps throughout the questionnaire is well-established in web surveys and the method does not require sophisticated preprocessing or modeling steps, it can be efficiently applied after data collection to identify and assess the prevalence of likely inattentive responding (in our case, up to 52 percent in the Study 1 data).

Given that we have shown that the industry standard threshold (below one-third of the median completion time) is inappropriate as it flags significantly fewer inattentive respondents (in our case, only 0.08% and 0.9% of the respective study samples), we suggest that the proposed clustering method be applied by survey vendors after data collection to enhance their data quality controls. While using internal criteria (e.g., cluster validity indices) is likely to lead to only a small number of clusters being identified (separating the fast and slow respondents), external criteria (e.g., straightlining) can provide more detailed insights and identify more nuanced response behaviors. If no external criteria are available, practitioners may exploratively set the number of clusters higher to infer whether particularly suspicious clusters or clusters with noteworthy response patterns are likely to emerge. The data visualizations introduced in this study may aid in identifying such patterns. We note, however, that implementing the method in "real-time" during the field period may be limited by the sample composition. As the cluster analysis is based on *relative* screen durations, a respondent's cluster assignment may change over the course of the field period if, for example, more inattentive respondents respond later in the field. However, the most extreme cases can still be identified early on as they will be assigned to the most suspicious clusters throughout the field period.

Our analyses are not without limitations. First, the experiments were implemented in a non-probability web survey for a younger population that is prone to inattention, which may limit external validity. For the commitment pledge experiment, the findings do not necessarily invalidate the use of commitment pledges per se as they might still be effective in other settings less prone to inattention (e.g., with different populations and sampling strategies, see Hibben et al., 2022). Similarly, the difference between the IRIs depends on the true proportion of inattentive respondents. As approximately 20 percent (for a 5-point scale) of inattentive respondents might pass an IRI that requests a specific response by chance, the difference between the investigated IRI versions decreases with a decreasing proportion of inattentive respondents. Second, the proposed cluster analysis method is an unsupervised algorithm and thus does not output clusters with definitive labels. Hence, it is still up to the researcher to decide which clusters should be deemed inattentive. However, the internal and external criteria we used coupled with the proposed data visualization tool can greatly assist with this decision, while still providing leeway on how strict one wants to be with regard to flagging respondents. Third, for the impact on substantive results, we lack true population values and can only estimate differences between likely attentive and inattentive respondents. However, the results are in line with expectations for inattentive respondents who utilize random responding behaviors, which validates the findings. Lastly, the detected prevalence of inattentive responding might be inflated by fatigue effects that increase over the questionnaire. While this is an issue for IRIs, the proposed cluster analysis approach can detect clusters of respondents who get faster as the survey progresses, and thus even alert researchers to potential problems with questionnaire length and design.

Future research may want to replicate our experiments and apply the proposed cluster analysis to data collected from different populations and sampling schemes. Applied researchers who conduct substantive analyses using data collected from similar sources as used here should carefully assess the quality of those data, as non-probability surveys are not only prone to biases due to selection but also potentially substantial amounts of inattentive response behavior as we have illustrated here.

Literature

Andreadis, I. (2021). Web survey response times: What to do and what not to do. *Proceedings of the Survey Research Method Section, American Statistical Association*, 1774–1782.

Anduiza, E., & Galais, C. (2017). Answering without reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, *29*(3), 497–519.

Aronow, P. M., Baron, J., & Pinson, L. (2019). A note on dropping experimental subjects who fail a manipulation check. *Political Analysis*, *27*(4), 572–589.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., & Lavrakas, P. J. (2010). Research synthesis: AAPOR report on online panels. *Public Opinion Quarterly*, *74*(4), 711–781.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*(3), 739–753.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2016). Can we turn shirkers into workers? *Journal of Experimental Social Psychology*, *66*, 20–28.

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, *24*(4), 718–738.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. *Sociological Methodology*, *12*(1981), 389–437.

Chang, L., & Krosnick, J. A. (2010). Comparing oral interviewing with self-administered computerized questionnaires: An experiment. *Public Opinion Quarterly*, *74*(1), 154–167.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust : An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*(6), 1–36.

Cibelli, K. L. (2017). *The effects of respondent commitment and feedback on response quality in online surveys* [Doctoral dissertation, University of Michigan]. https://deepblue.lib.umich.edu/bitstream/2027.42/136981/1/kcibelli_1.pdf

Clifford, S., & Jerit, J. (2015). Do attempts to improve respondent attention increase social desirability bias? *Public Opinion Quarterly*, *79*(3), 790–802.

Conrad, F. G., Couper, M. P., Tourangeau, R., & Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods*, *11*(1), 45–61.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36.

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

Couper, M. P. (2017). New developments in survey data collection. *Annual Review of Sociology*, *43*, 121–145.

Druckman, J. N. (2001). Evaluating framing effects. *Journal of Economic Psychology*, *22*, 91–101.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*(3), 224–247.

Geisen, E. (2022). *Improve data quality by using a commitment request instead of attention checks*. Market Research. Retrieved September 9, 2023, from https://www.qualtrics.com/blog/attention-checks-and-data-quality/

Gomila, R. (2021). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, *150*(4), 700–709.

Greszki, R., Meyer, M., & Schoen, H. (2015). Exploring the effects of removing "too fast" responses and respondents from web surveys. *Public Opinion Quarterly*, *79*(2), 471–503.

Gummer, T., Roßmann, J., & Silber, H. (2021). Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research*, *50*(1), 238–264.

Hauser, D. J., & Schwarz, N. (2015). It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *SAGE Open*, *5*(2), 1–6.

Hauser, D. J., Sunderrajan, A., Natarajan, M., & Schwarz, N. (2016). Prior exposure to instructional manipulation checks does not attenuate survey context effects driven by satisficing or gricean norms. *methods, data, analyses*, *10*(2), 195–220.

Hibben, K. C., Felderer, B., & Conrad, F. G. (2022). Respondent commitment: Applying techniques from face-to-face interviewing to online collection of employment data. *International Journal of Social Research Methodology*, *25*(1), 15–27.

Höhne, J. K., Schlosser, S., Couper, M. P., & Blom, A. G. (2020). Switching away: Exploring on-device media multitasking in web surveys. *Computers in Human Behavior*, *111*, 106417.

Hornsey, M. J., Harris, E. A., Bain, P. G., & Fielding, K. S. (2016). Meta-analyses of the determinants and outcomes of belief in climate change. *Nature Climate Change*, *6*(6), 622–626.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828–845.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129.

Kane, J. V. (2024). More than meets the ITT: A guide for anticipating and investigating non-significant results in survey experiments. *Journal of Experimental Political Science*, 1–16. https://doi.org/10.1017/XPS.2024.1

Kane, J. V., Velez, Y. R., & Barabas, J. (2023). Analyze the attentive and bypass bias: Mock vignette checks in survey experiments. *Political Science Research and Methods*, *11*, 293–310.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail – web mixed-mode surveys. *Social Science Computer Review*, *37*(2), 214–233.

Kreuter, F., Couper, M., & Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. *Proceedings of the Survey Research Method Section, American Statistical Association*, 282–296.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Leiner, D. J. (2019). Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods*, *13*(3), 229–248.

Liu, M., & Wronski, L. (2018). Trap questions in online surveys: Results from three web survey experiments. *International Journal of Market Research*, *60*(1), 32–49.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83.

Matjašič, M., Vehovar, V., & Manfreda, K. L. (2018). Web survey paradata on response time outliers: A systematic literature review. *Advances in Methodology and Statistics*, *15*(1), 23–41.

McPhee, C., Barlas, F., Brigham, N., Darling, J., Dutwin, D., Jackson, C., Jackson, M., Kirzinger, A., Little, R., Lorenz, E., Marlar, J., Mercer, A., Scanlon, P. J., Weiss, S., & Wronski, L. (2022). *Data quality metrics for online samples: Considerations for study design and analysis.* AAPOR.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455.

Melipillan, E. R. (2019). *Careless survey respondents: Approaches to identify and reduce their negative impact on survey estimates.*

Mercer, A. W., Kreuter, F., Keeter, S., & Stuart, E. A. (2017). Theory and practice in nonprobability surveys. *Public Opinion Quarterly*, *81*, 250–279.

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, *62*(3), 760–775.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872.

Read, B., Wolters, L., & Berinsky, A. J. (2021). Replication data for: Racing the clock: Using response time as a proxy for attentiveness on self-administered surveys. *Harvard Dataverse*, (V1). https://doi.org/10.7910/DVN/6OGTHL

Read, B., Wolters, L., & Berinsky, A. J. (2022). Racing the clock: Using response time as a proxy for attentiveness on self-administered surveys. *Political Analysis*, *30*(4), 550–569.

Sendelbah, A., Vehovar, V., Slavec, A., & Petrovčič, A. (2016). Investigating respondent multitasking in web surveys using paradata. *Computers in Human Behavior*, *55*, 777–787.

Shamon, H., & Berning, C. C. (2020). Attention check items and instructions in online surveys: Boon or bane for data quality? *Survey Research Methods*, *14*(1), 55–77.

Silber, H., Danner, D., & Rammstedt, B. (2019). The impact of respondent attentiveness on reliability and validity. *International Journal of Social Research Methodology*, *22*(2), 153–164.

Silber, H., Roßmann, J., & Gummer, T. (2022). The issue of noncompliance in attention check questions: False positives in instructed response items. *Field Methods*, *34*(4), 346–360.

Ternovski, J., Orr, L., Kalla, J., & Aronow, P. (2022). A note on increases in inattentive online survey-takers since 2020. *Journal of Quantitative Description: Digital Media*, *2*, 1–35.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(30), 453–458.

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & Von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593–619.

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & Von Davier, M. (2024). Using response times for joint modeling of careless responding and attentive response styles. *Journal of Educational and Behavioral Statistics*, *49*(2), 173–206.

Ulitzsch, E., Shin, H. J., & Lüdtke, O. (2024). Accounting for careless and insufficient effort responding in large-scale survey data—development, evaluation, and application of a screen-time-based weighting procedure. *Behavior Research Methods*, *56*(2), 804–825.

Welz, M., & Alfons, A. (2024). *When respondents don't care anymore: Identifying the onset of careless responding.* arXiv: 2303.07167 [stat].

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, *5*(2), 175–211.

Yan, T. (2008). Nondifferentiation. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 520–521). SAGE Publications, Inc.

Appendix

7.A Validity of screen durations

As timestamp data are process-generated data and delivered from a survey vendor, we implemented several checks with regard to the validity of these data. Potential threats to data validity are the speed of the internet connection, overwriting timestamps when going back and forth, and refreshing questionnaire pages. We tested the influence of the internet connection by conducting test interviews and using Google Devtools, which allows for setting the network speed to slow 3G and fast 3G. We conducted three interviews in each condition and clicked through the questionnaire without providing any responses. Influences of the internet connection should result in different screen durations. Figure 7.A1 shows the durations over the screens for all interviews. We find no evidence that the speed of the internet connection counts toward screen durations and thus does not pose a threat to the validity of the timestamp data. In further test interviews, we validated that going back and forth does not overwrite screen durations, but found that refreshing does overwrite screen durations. In the latter case, any provided responses are deleted as well. Thus, even if respondents refresh a screen they still have to provide a response which should take some time. In addition, such cases should emerge for single screens and not repeatedly throughout the interview, and thus the influences of refreshing screens should be minimal. In sum, our tests suggest the timestamp data sent by the survey institute are valid.

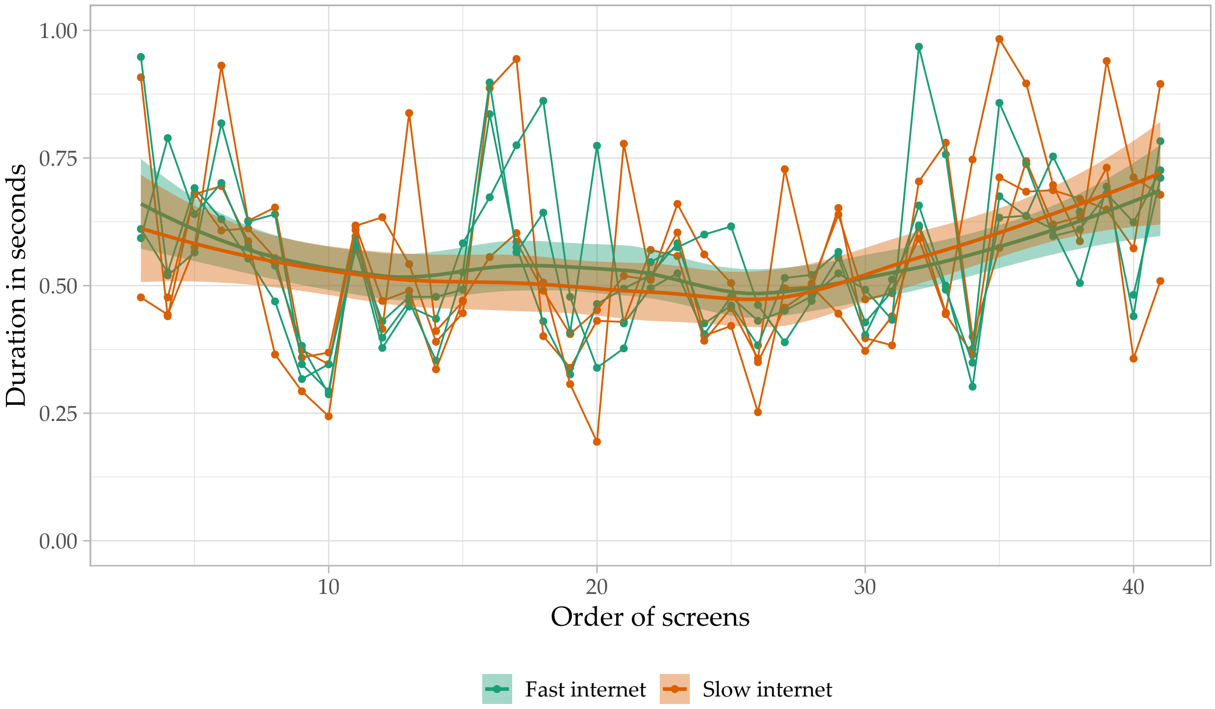


Figure 7.A1: Screen durations for 6 test interviews (3 with slow 3G, 3 with fast 3G).

7.B Cluster analysis results for desktop respondents

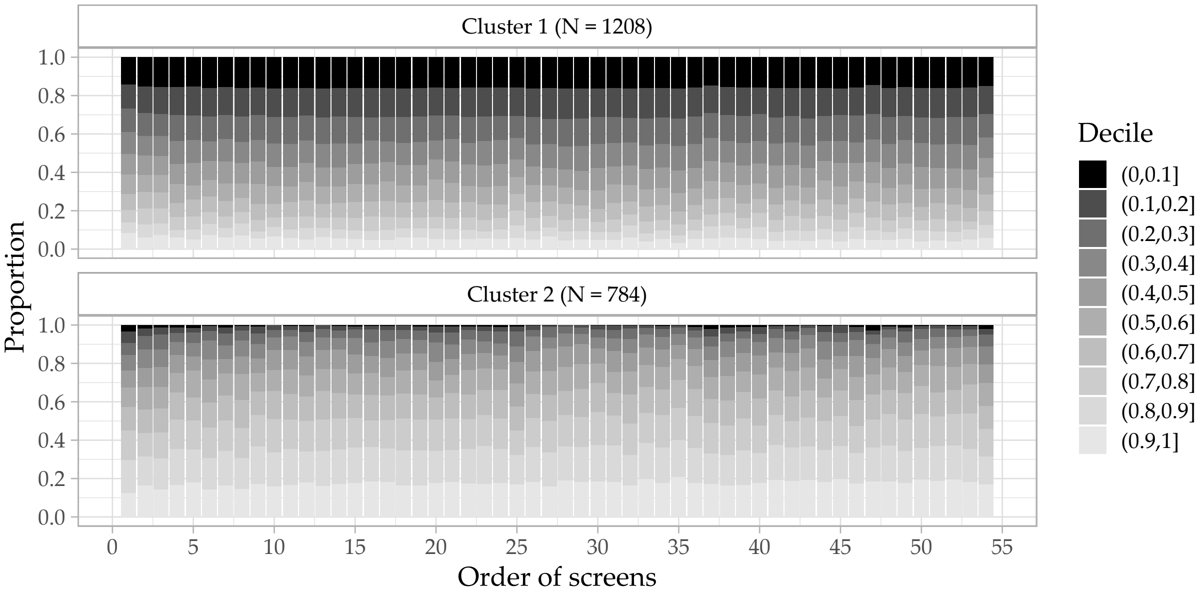


Figure 7.B1: Duration decile composition of 2-cluster solution, desktop respondents.

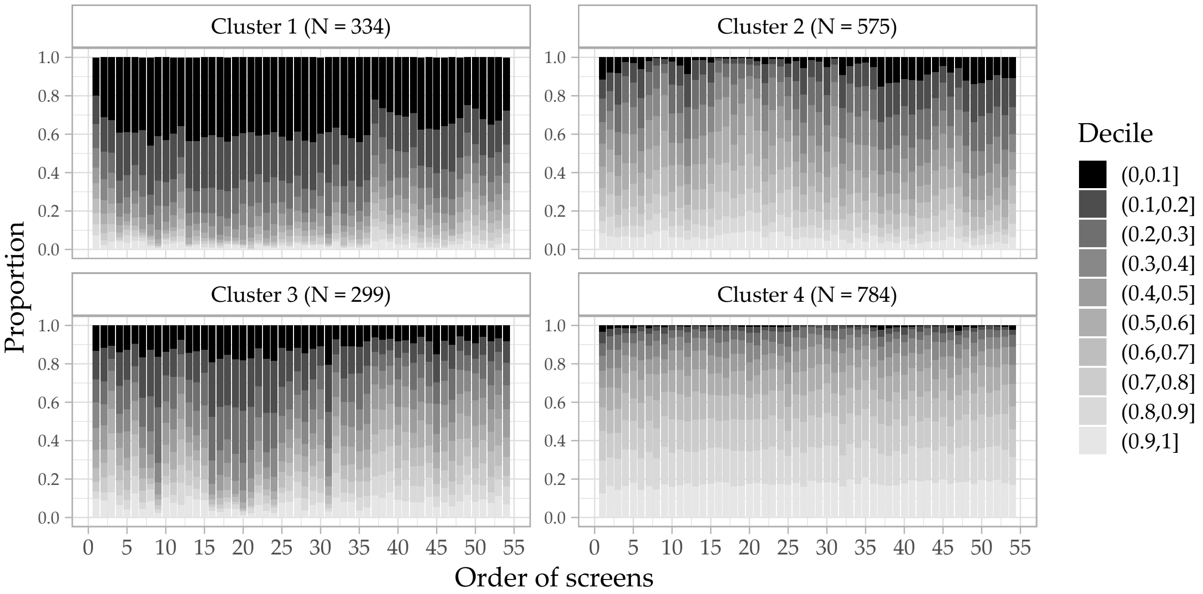


Figure 7.B2: Duration decile composition of 4-cluster solution, desktop respondents.

Table 7.B1: Proportion of straightlining across clusters, distance-based approach for desktop respondents.

Solution	Cluster	Q5	Q7	Q18	Q19
2	Cluster 1 (N = 1208)	0.152	0.177	0.157	0.184
2		[0.132,0.173]	[0.156,0.200]	[0.137,0.179]	[0.162,0.207]
2	Cluster 2 (N = 784)	0.075	0.085	0.069	0.117
2		[0.058,0.096]	[0.067,0.107]	[0.052,0.089]	[0.096,0.142]
4	Cluster 1 (N = 334)	0.240	0.302	0.284	0.281
4		[0.195,0.290]	[0.254,0.355]	[0.237,0.336]	[0.234,0.333]
4	Cluster 2 (N = 575)	0.047	0.049	0.045	0.071
4		[0.031,0.068]	[0.033,0.070]	[0.030,0.066]	[0.052,0.095]
4	Cluster 3 (N = 299)	0.254	0.284	0.231	0.291
4		[0.206,0.307]	[0.234,0.339]	[0.184,0.283]	[0.240,0.346]
4	Cluster 4 (N = 784)	0.075	0.085	0.069	0.117
4		[0.058,0.096]	[0.067,0.107]	[0.052,0.089]	[0.096,0.142]

Notes: 95 percent confidence intervals in brackets.

Table 7.B2: Proportion of straightlining across clusters, model-based approach for desktop respondents.

Solution	Cluster	Q5	Q7	Q18	Q19
2	Cluster 1 (N = 1061)	0.104	0.120	0.106	0.146
2		[0.086,0.124]	[0.101,0.141]	[0.088,0.126]	[0.125,0.169]
2	Cluster 2 (N = 931)	0.142	0.165	0.142	0.171
2		[0.120,0.166]	[0.142,0.191]	[0.120,0.166]	[0.147,0.197]
3	Cluster 1 (N = 478)	0.211	0.253	0.224	0.272
3		[0.176,0.251]	[0.215,0.295]	[0.187,0.264]	[0.233,0.314]
3	Cluster 2 (N = 907)	0.137	0.163	0.137	0.164
3		[0.115,0.161]	[0.140,0.189]	[0.115,0.161]	[0.141,0.190]
3	Cluster 3 (N = 607)	0.028	0.020	0.021	0.058
3		[0.016,0.044]	[0.010,0.034]	[0.011,0.036]	[0.040,0.079]

Notes: 95 percent confidence intervals in brackets.

7.C Cluster validity indices

Table 7.C1: Internal cluster validity indices.

Index	Optimal number of clusters
KL	2
CH	2
CCC	2
Cindex	2
DB	2
Silhouette	2
Ratkowsky	2
PtBiserial	2
McClain	2
Dunn	2
SDindex	2
Hartigan	3
Scott	3
TraceW	3
Friedman	3
Rubin	3
Ball	3
Frey	3
Marriot	6
TrCovW	6
SDbw	11
Beale	13

7.D Item batteries for straightlining in Study 1

Table 7.D1: Q5.

How much, if at all, does climate change make you feel any of the following?	
Sad	- not at all - a little - moderately - very - extremely
Helpless	- not at all - a little - moderately - very - extremely
Anxious	- not at all - a little - moderately - very - extremely
Afraid	- not at all - a little - moderately - very - extremely
Optimistic	- not at all - a little - moderately - very - extremely
Angry	- not at all - a little - moderately - very - extremely
Guilty	- not at all - a little - moderately - very - extremely
Ashamed	- not at all - a little - moderately - very - extremely
Hurt	- not at all - a little - moderately - very - extremely
Depressed	- not at all - a little - moderately - very - extremely
Despair	- not at all - a little - moderately - very - extremely
Grief	- not at all - a little - moderately - very - extremely
Powerless	- not at all - a little - moderately - very - extremely
Indifferent	- not at all - a little - moderately - very - extremely
Numb	- not at all - a little - moderately - very - extremely
Isolated/lonely	- not at all - a little - moderately - very - extremely

Table 7.D2: Q7.

How much, if at all, does climate change make you think the following?	
I'm hesitant to have children	- not at all - a little - moderately - very much - extremely
Humanity is doomed	- not at all - a little - moderately - very much - extremely
The future is frightening	- not at all - a little - moderately - very much - extremely
I won't have access to the same opportunities my parents had	- not at all - a little - moderately - very much - extremely
My, or my family's, security will be threatened (such as economic, social, physical)	- not at all - a little - moderately - very much - extremely
The things I most value will be destroyed	- not at all - a little - moderately - very much - extremely
People have failed to take care of the planet	- not at all - a little - moderately - very much - extremely
My life will be worse because of climate change	- not at all - a little - moderately - very much - extremely
My choice of where to live will be influenced by the impact of climate change	- not at all - a little - moderately - very much - extremely
My feelings about climate change cause problems in my relationships	- not at all - a little - moderately - very much - extremely
I should focus on enjoying myself now instead of planning for the future	- not at all - a little - moderately - very much - extremely
It's hard for me to be motivated to succeed in a job, career, or vocation.	- not at all - a little - moderately - very much - extremely
I question whether the work I put into my education will matter.	- not at all - a little - moderately - very much - extremely
My plans for the future will be impacted by climate change	- not at all - a little - moderately - very much - extremely
My life will be better because of climate change	- not at all - a little - moderately - very much - extremely
Climate change will bring more positive than negative changes to the world	- not at all - a little - moderately - very much - extremely
Warmer weather in some places will be a welcome change	- not at all - a little - moderately - very much - extremely

Table 7.D3: Q18.

In relation to climate change, do you believe that the US government is:	
Taking your concerns seriously enough	yes - no
Doing enough to avoid a climate catastrophe	yes - no
Dismissing people's distress	yes - no
Acting in line with climate science	yes - no
Protecting you, the planet and/or future generations	yes - no
Trustworthy	yes - no
Lying about the effectiveness of the actions they're taking	yes - no
Failing young Americans	yes - no
Betraying you and/or future generations	yes - no

Table 7.D4: Q19.

When you think about how the US government is responding to climate change, how much, if at all, do you feel:	
Anguished	not at all - a little - moderately - very - extremely
Abandoned	not at all - a little - moderately - very - extremely
Afraid	not at all - a little - moderately - very - extremely
Hopeful	not at all - a little - moderately - very - extremely
Reassured	not at all - a little - moderately - very - extremely
Angry	not at all - a little - moderately - very - extremely
Valued	not at all - a little - moderately - very - extremely
Ashamed	not at all - a little - moderately - very - extremely
Belittled	not at all - a little - moderately - very - extremely
Protected	not at all - a little - moderately - very - extremely
Ignored	not at all - a little - moderately - very - extremely
Proud	not at all - a little - moderately - very - extremely
Thankful	not at all - a little - moderately - very - extremely
Dismissed	not at all - a little - moderately - very - extremely

7.E Model-based approach with top-coded screen durations

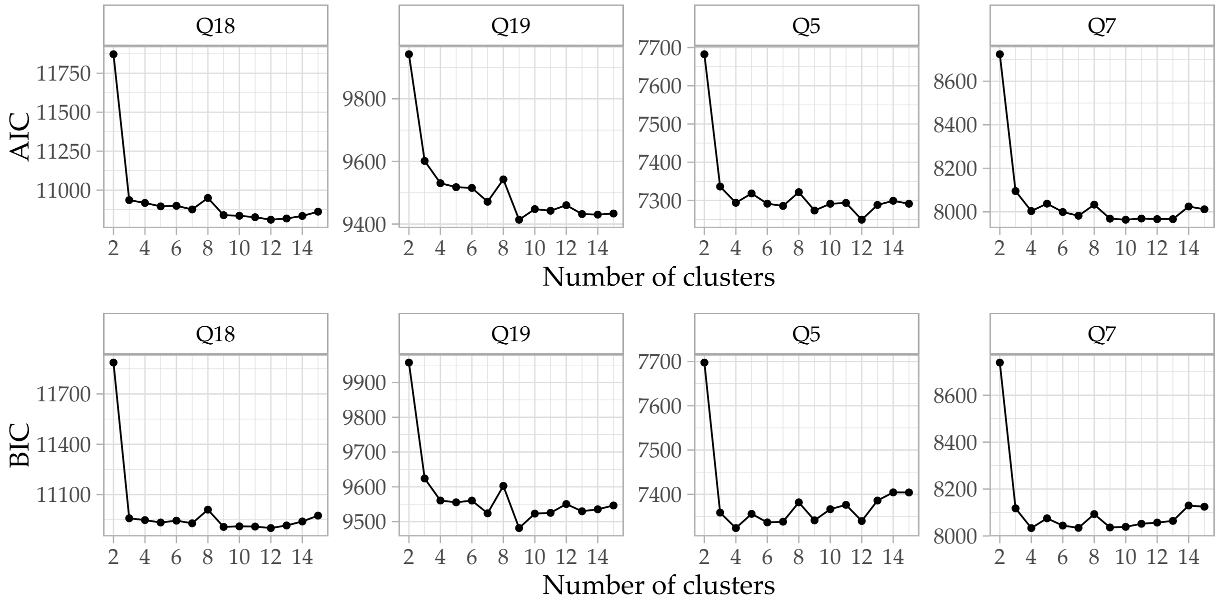


Figure 7.E1: AIC and BIC for model-based clusters, top-coded screen durations.

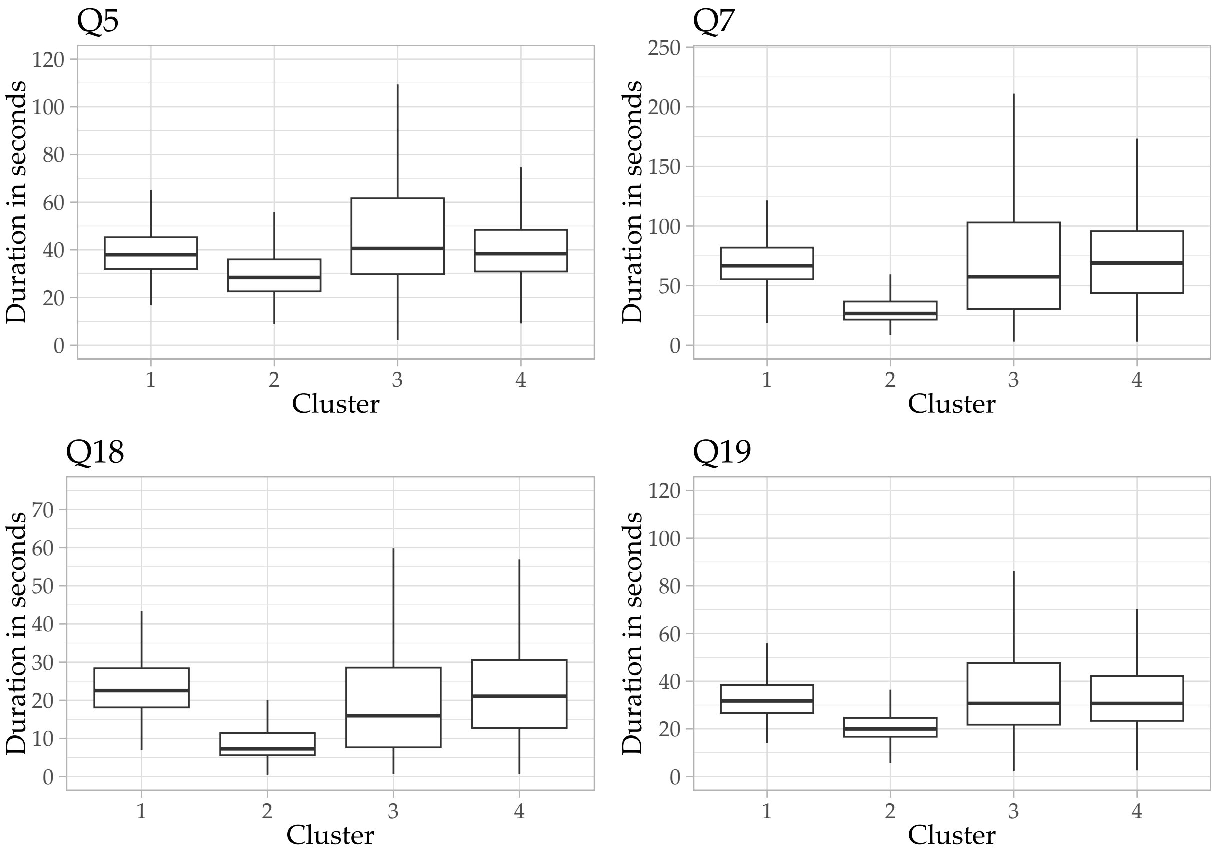


Figure 7.E2: Boxplots of durations for item batteries by cluster, top-coded screen durations.

Table 7.E1: Proportion of straightlining across clusters, model-based approach with top-coding.

Solution	Cluster	Q5	Q7	Q18	Q19
3	Cluster 1 (N = 4474)	0.026	0.021	0.040	0.048
3		[0.022,0.031]	[0.017,0.026]	[0.034,0.046]	[0.042,0.055]
3	Cluster 2 (N = 3461)	0.137	0.182	0.283	0.181
3		[0.125,0.149]	[0.170,0.196]	[0.268,0.298]	[0.168,0.194]
3	Cluster 3 (N = 5823)	0.089	0.104	0.171	0.134
3		[0.081,0.096]	[0.096,0.112]	[0.161,0.181]	[0.126,0.143]
4	Cluster 1 (N = 3084)	0.014	0.006	0.024	0.029
4		[0.010,0.019]	[0.003,0.009]	[0.019,0.030]	[0.023,0.035]
4	Cluster 2 (N = 3295)	0.135	0.182	0.282	0.182
4		[0.124,0.147]	[0.169,0.195]	[0.266,0.297]	[0.169,0.195]
4	Cluster 3 (N = 3817)	0.100	0.117	0.191	0.144
4		[0.091,0.110]	[0.107,0.128]	[0.179,0.204]	[0.133,0.156]
4	Cluster 4 (N = 3562)	0.066	0.075	0.118	0.109
4		[0.058,0.075]	[0.067,0.084]	[0.108,0.129]	[0.099,0.119]

Notes: 95 percent confidence intervals in brackets.

7.F Item battery with IRI experiment in Study 2

Table 7.F1: Q11.

How much, if at all, do these factors contribute to your feelings about climate change:	
Severe weather events in my area/region	- not at all - a little - moderately - very much - extremely
Unseasonable or unusual weather in my area/region	- not at all - a little - moderately - very much - extremely
News about climate change or weather events on social media or in mainstream media	- not at all - a little - moderately - very much - extremely
Current response of governments of poor countries	- not at all - a little - moderately - very much - extremely
Current response of the US government	- not at all - a little - moderately - very much - extremely
Current response of governments of other wealthy countries	- not at all - a little - moderately - very much - extremely
Current actions of corporations	- not at all - a little - moderately - very much - extremely
To show you have read this sentence please mark “Very much”	- not at all - a little - moderately - very much - extremely
To show you have read this sentence please leave the question blank	- not at all - a little - moderately - very much - extremely
Current actions of my family and families like mine	- not at all - a little - moderately - very much - extremely

7.G Cluster analysis for Study 2

Durations are available on the screen level in milliseconds. Note, however, that several item batteries are split across multiple screens, such that each screen contains either informational text, a single item, or parts of an item battery. Following the approach for Study 1, we exclude screens that could be over-filtered and exclude screens subject to experimental manipulation. In total, the sample consists of 276,000 durations for 50 screens and 5,520 respondents.

The cluster validity indices suggest a 2-cluster solution, and the external criterion (failing the attention check) suggests a 9-cluster solution (see Figure 7.G1). We show the results for the latter.

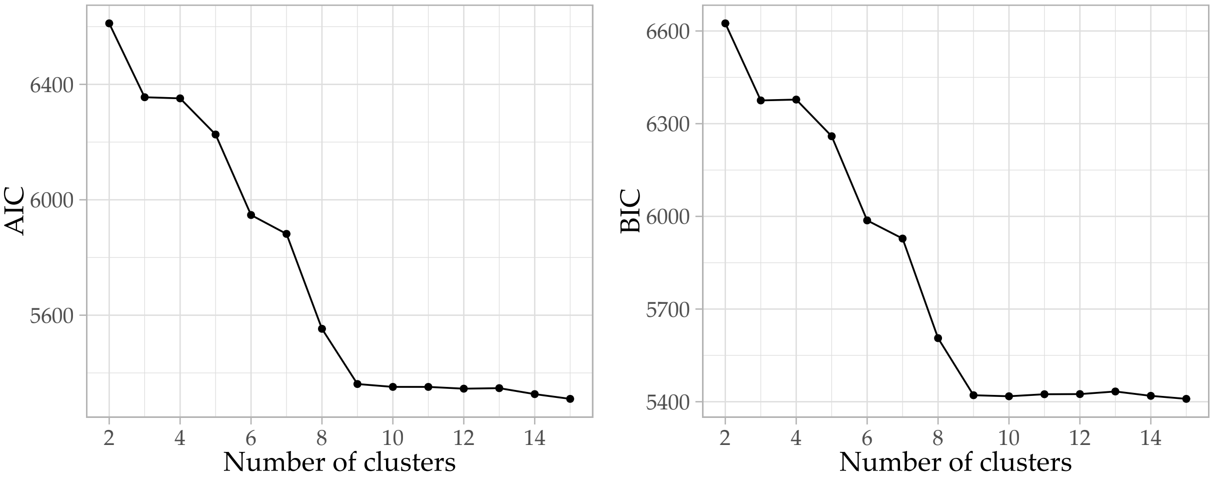


Figure 7.G1: AIC and BIC for distance-based clusters for Study 2 data.

Figure 7.G2 shows the cluster composition across duration deciles. Cluster 2 is the fastest cluster with around 40 percent of durations in the first decile for most items. In Cluster 7, respondents speed up in the first ten screens and are slower for the later screens (i.e., the socio-demographic questions). Clusters 1, 4, and 6 show similar developments, though less pronounced. Clusters 4 and 5 show less variation over the questionnaire but have substantial shares in the fastest deciles. Clusters 3 and 8 are relatively fast at the beginning and at the end (i.e., the socio-demographic questions), whereas Cluster 9 is slow throughout the entire interview. Among the respondents flagged by the industry standard threshold (i.e. below one-third of the median duration), 49 lie within Cluster 2 and 1 respondent belongs to Cluster 7.

Table 7.G1 lists the proportion of respondents failing the attention check across clusters. In Clusters 3, 8, and 9, only around 5 percent fail the attention check. We observe shares above 70 percent for Clusters 2, 6, and 7 and between 40 and 60 percent for Clusters 1, 4, and 5.

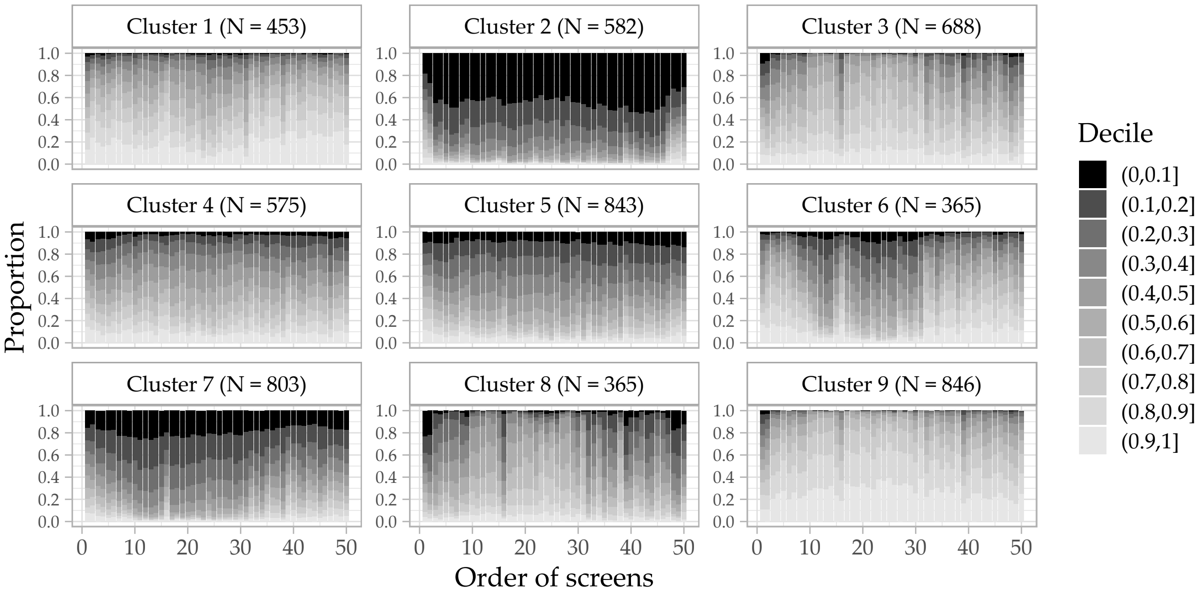


Figure 7.G2: Duration decile composition of 9-cluster solution for Study 2 data.

Table 7.G1: Proportion of respondents failing the attention check across clusters, distance-based approach.

Cluster	Proportion failed
Cluster 1 (N = 453)	0.519 [0.472,0.566]
Cluster 2 (N = 582)	0.716 [0.678,0.753]
Cluster 3 (N = 688)	0.057 [0.041,0.077]
Cluster 4 (N = 575)	0.424 [0.384,0.466]
Cluster 5 (N = 843)	0.569 [0.535,0.603]
Cluster 6 (N = 365)	0.781 [0.735,0.822]
Cluster 7 (N = 803)	0.760 [0.729,0.789]
Cluster 8 (N = 365)	0.044 [0.025,0.070]
Cluster 9 (N = 846)	0.060 [0.045,0.079]

Notes: 95 percent confidence intervals in brackets

7.H Information criteria for straightlining for cluster solutions

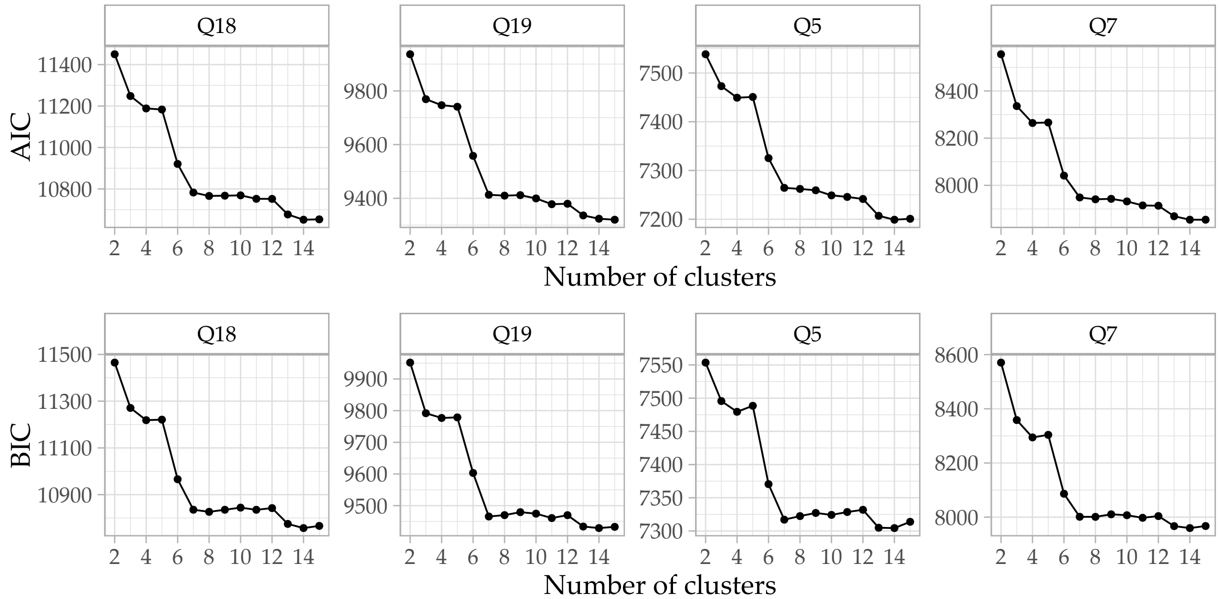


Figure 7.H1: AIC and BIC for distance-based clusters.

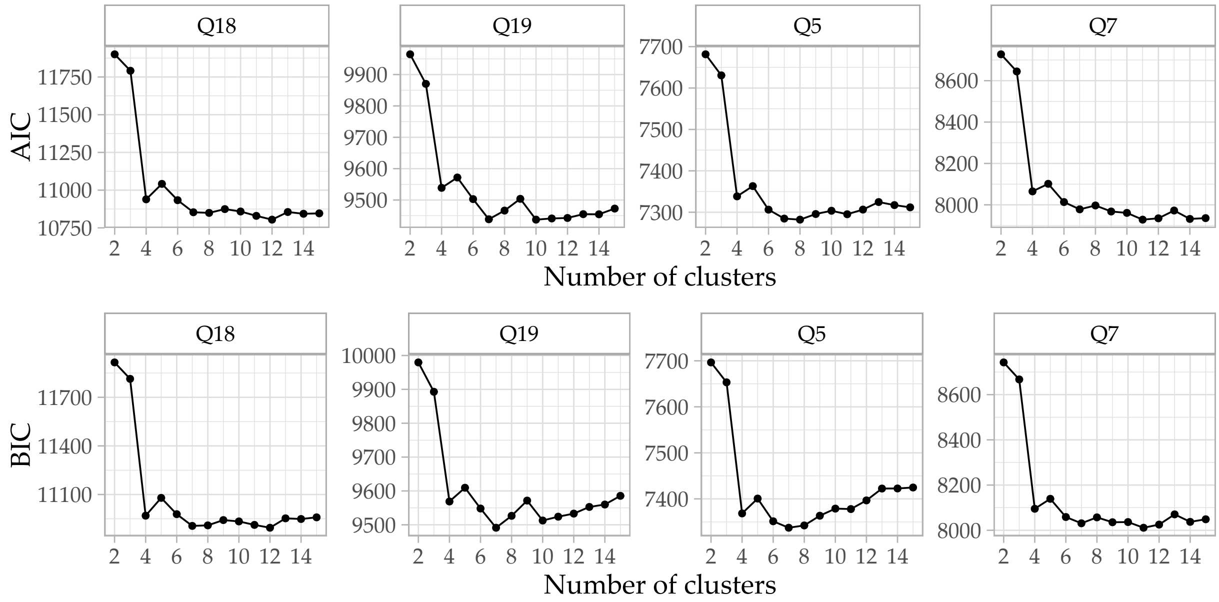


Figure 7.H2: AIC and BIC for model-based clusters.

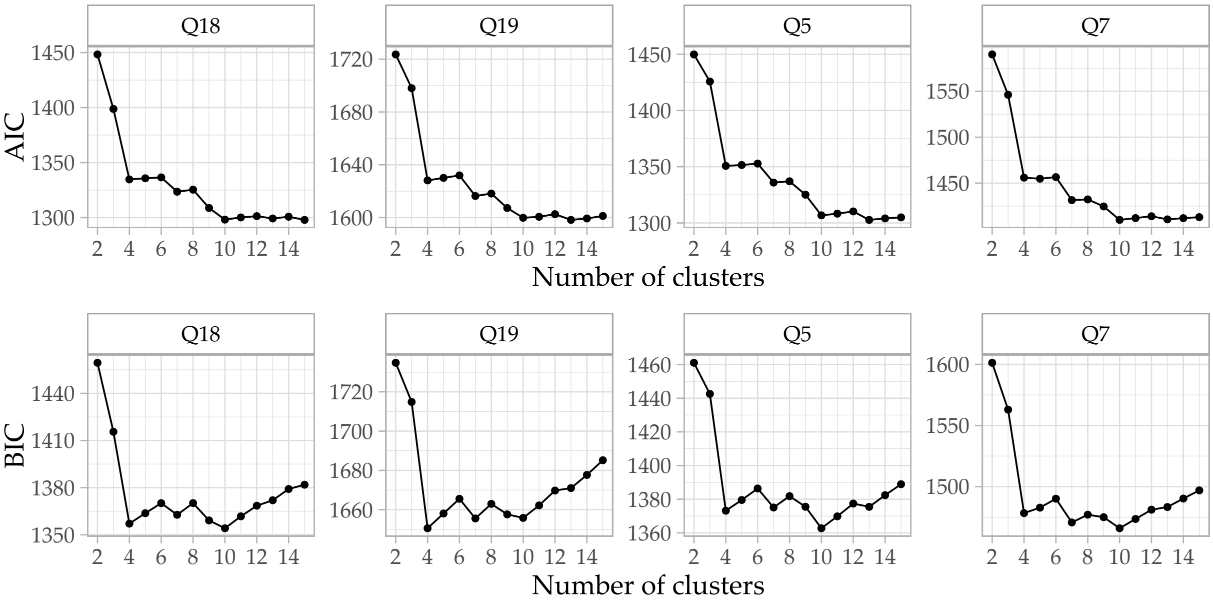


Figure 7.H3: AIC and BIC for distance-based clusters, desktop respondents.

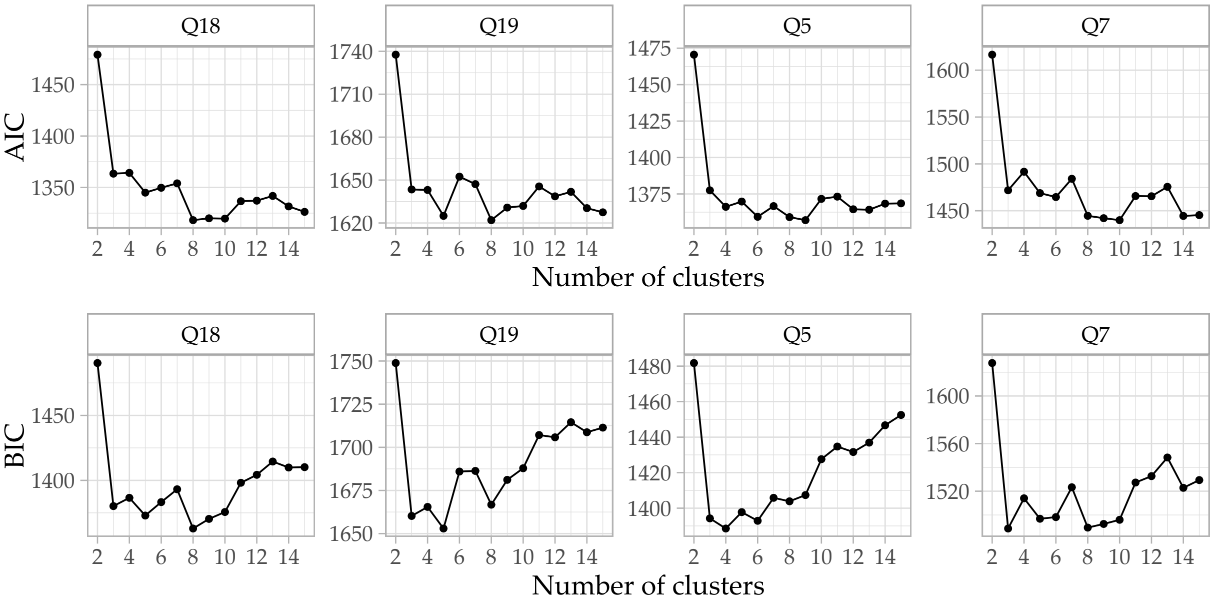


Figure 7.H4: AIC and BIC for model-based clusters, desktop respondents.

7.I Break-offs in Study 2

As described in the main text, the overall break-off rates do not differ across pledge conditions. Figures 7.I1 and 7.I2 show the Kaplan-Meier estimates for the entire sample and separately by the pledge conditions. The estimates denote the probability of staying in the survey at each questionnaire screen. We do not find that the pledge or the IRI experiments induce break-offs. We also estimate a Cox regression and find no statistically significant effect of the pledge ($exp(\beta) = 1.03$, $p - value = 0.51$).

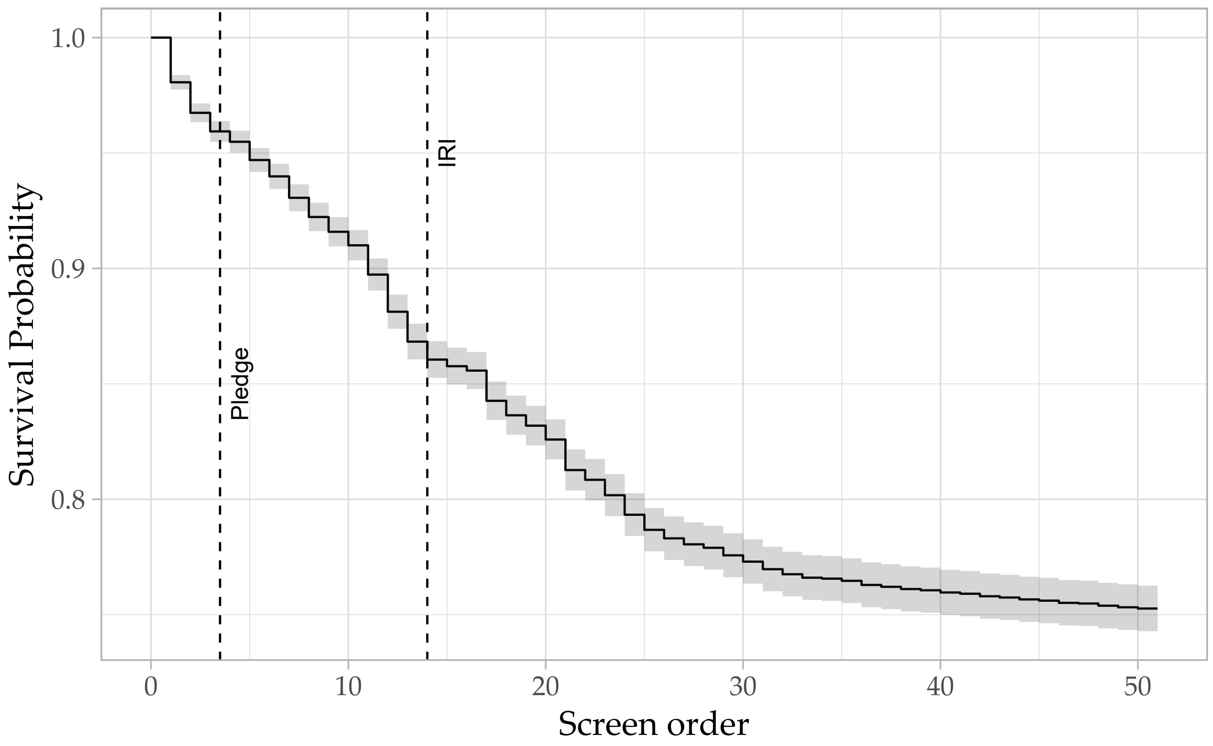


Figure 7.I1: Kaplan-Meier plot for break-offs (with 95 percent confidence intervals).

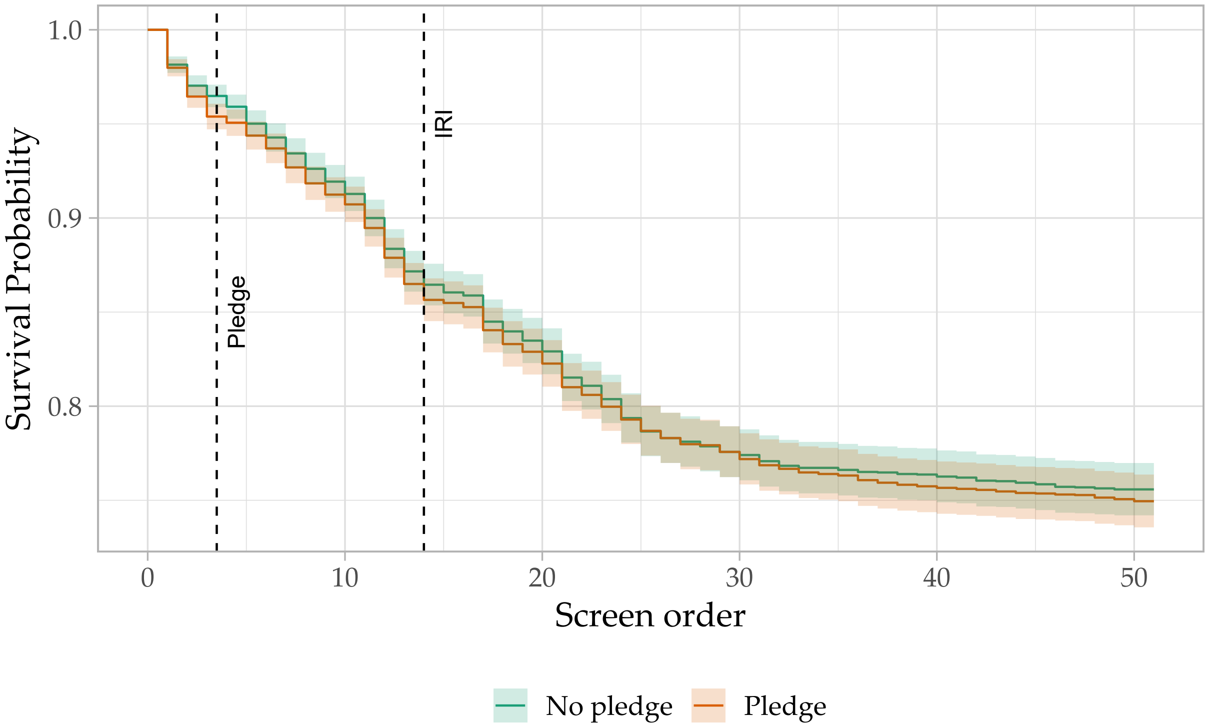


Figure 7.I2: Kaplan-Meier plot for break-offs by pledge condition (with 95 percent confidence intervals).

7.J Spillover effects of instructed response items

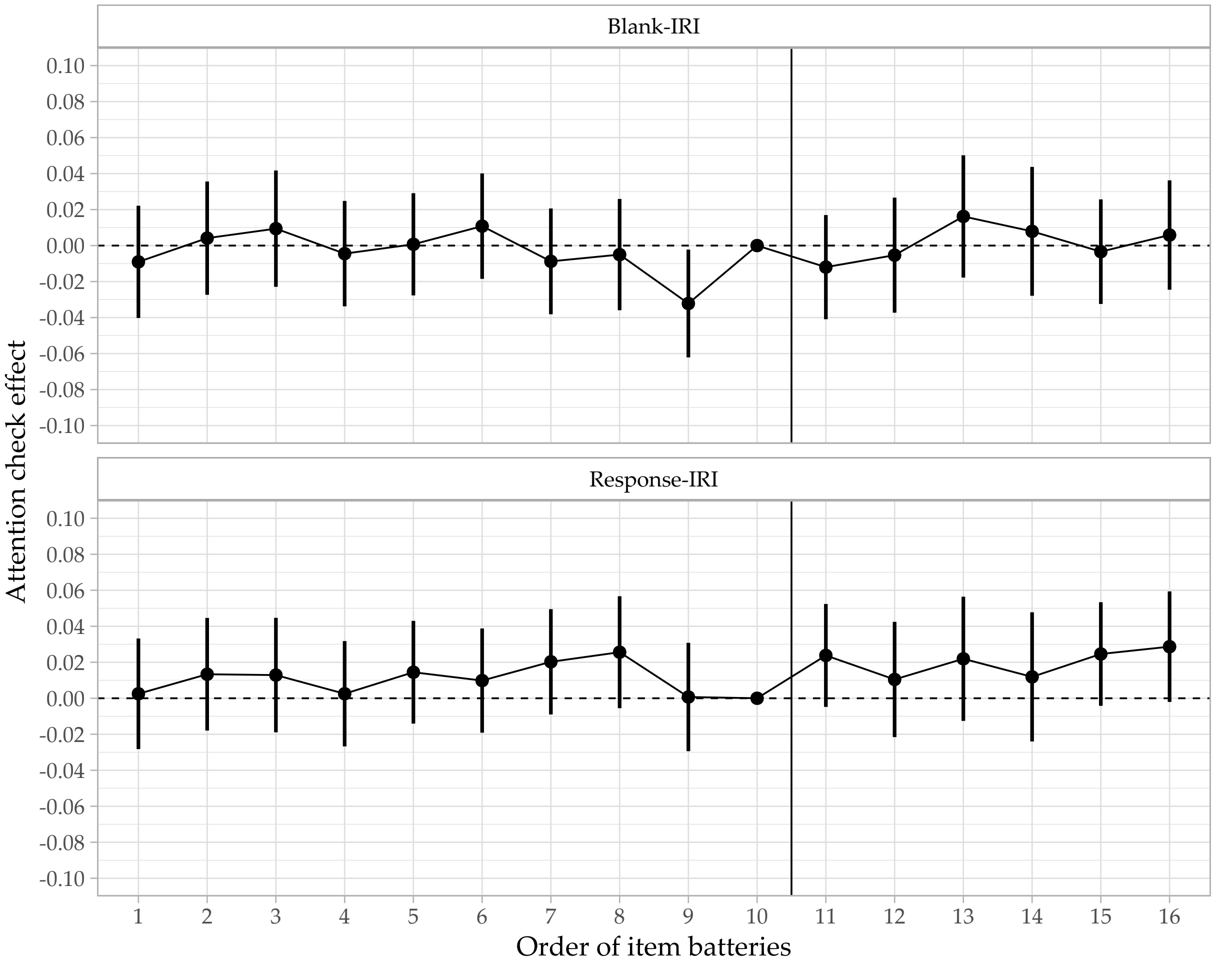


Figure 7.J1: Attention check effects on straightlining over the questionnaire (with 95 percent confidence intervals).

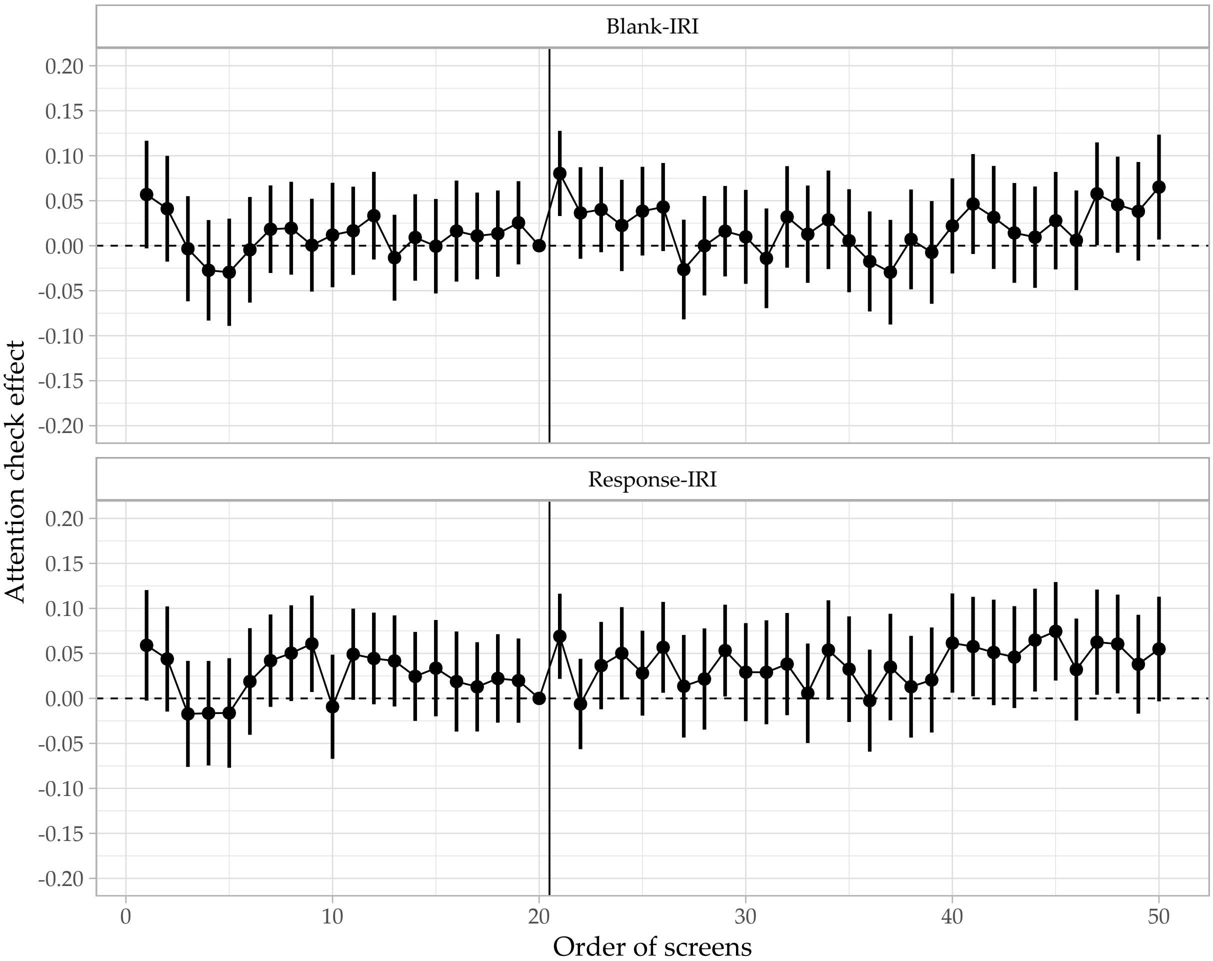


Figure 7.J2: Attention check effects on screen durations over the questionnaire (with 95 percent confidence intervals).

7.K Development of bias for all clustering approaches and devices

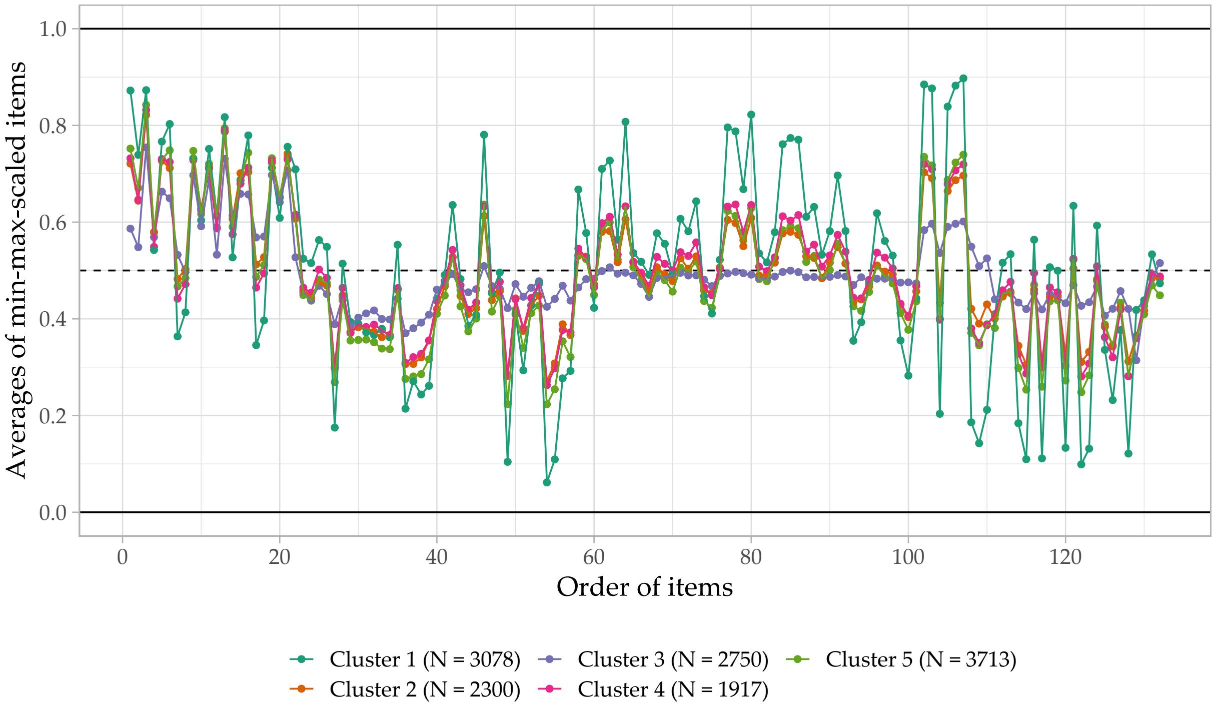


Figure 7.K1: Development of scaled average responses by cluster over the questionnaire, model-based approach.

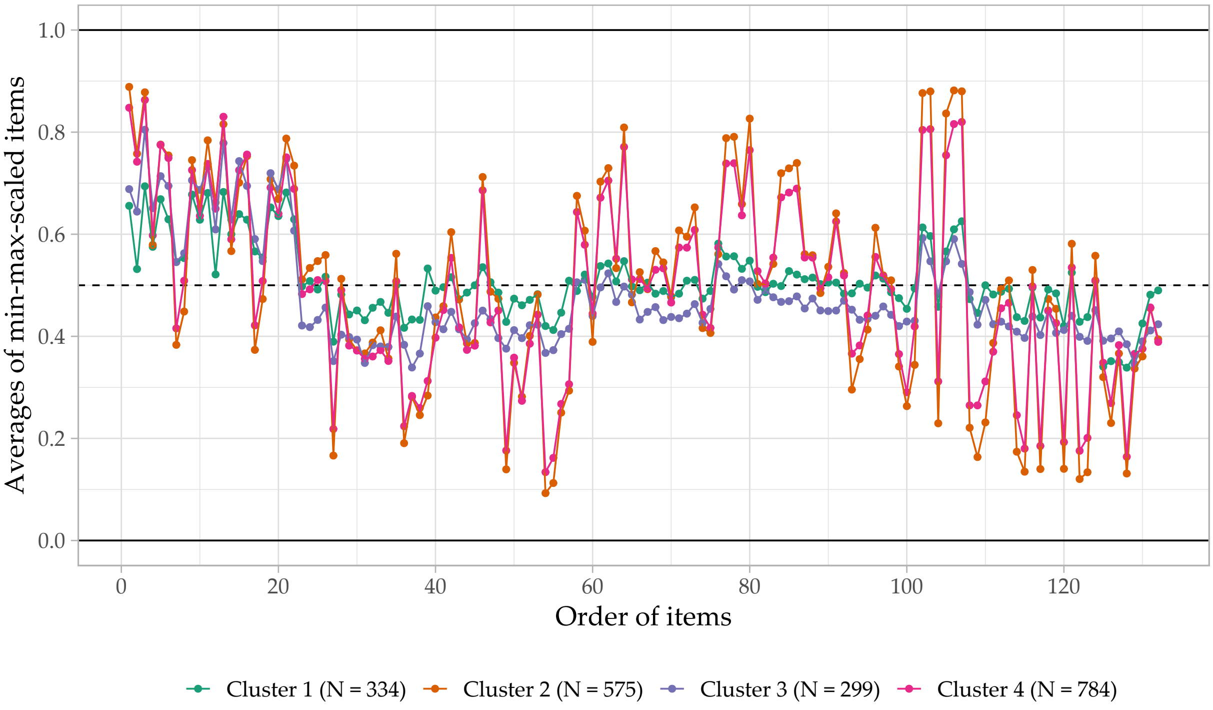


Figure 7.K2: Development of scaled average responses by cluster over the questionnaire, distance-based approach for desktop respondents.

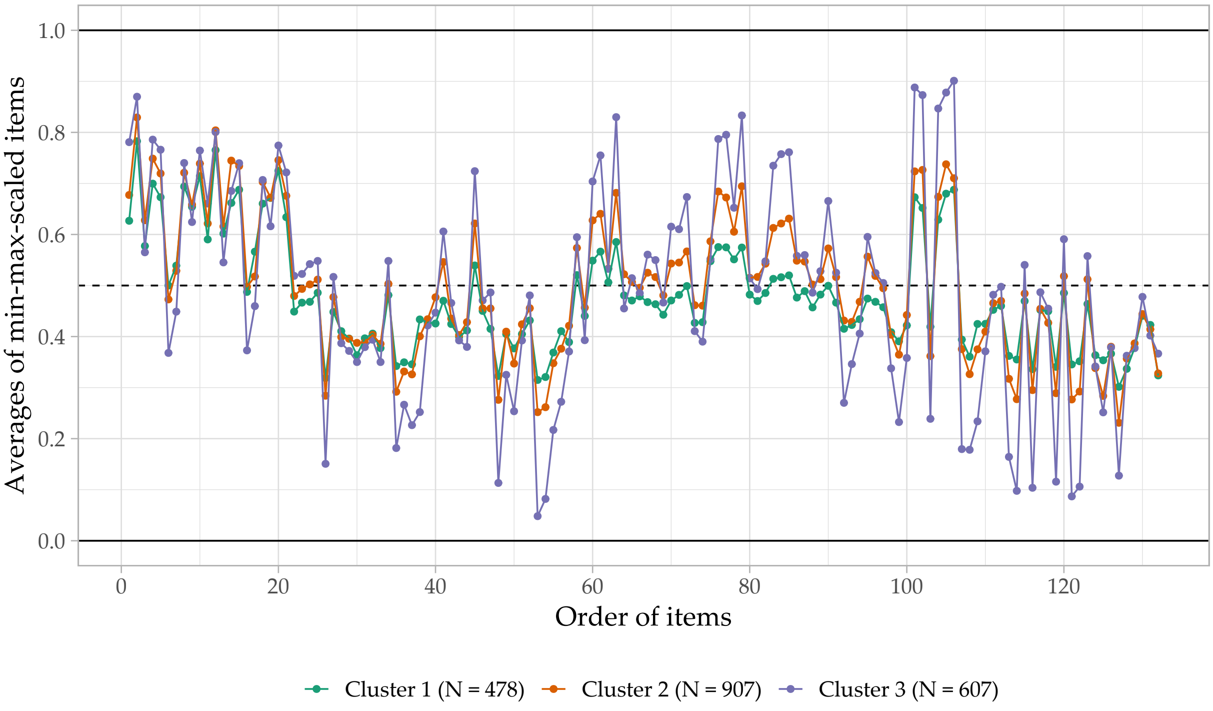


Figure 7.K3: Development of scaled average responses by cluster over the questionnaire, model-based approach for desktop respondents.

7.L Asian disease problem

Table 7.L1: Asian disease problem as developed by Tversky and Kahneman (1981).

Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:		
Gain	A	If Program A is adopted, 200 people will be saved.
	B	If Program B is adopted, there is 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved.
Loss	A	If Program A is adopted, 400 people will die.
	B	If Program B is adopted, there is 1/3 probability that no people will die, and a 2/3 probability that 600 people will die.
Which program do you favor?		

7.M Reanalyzing Read et al. (2022)

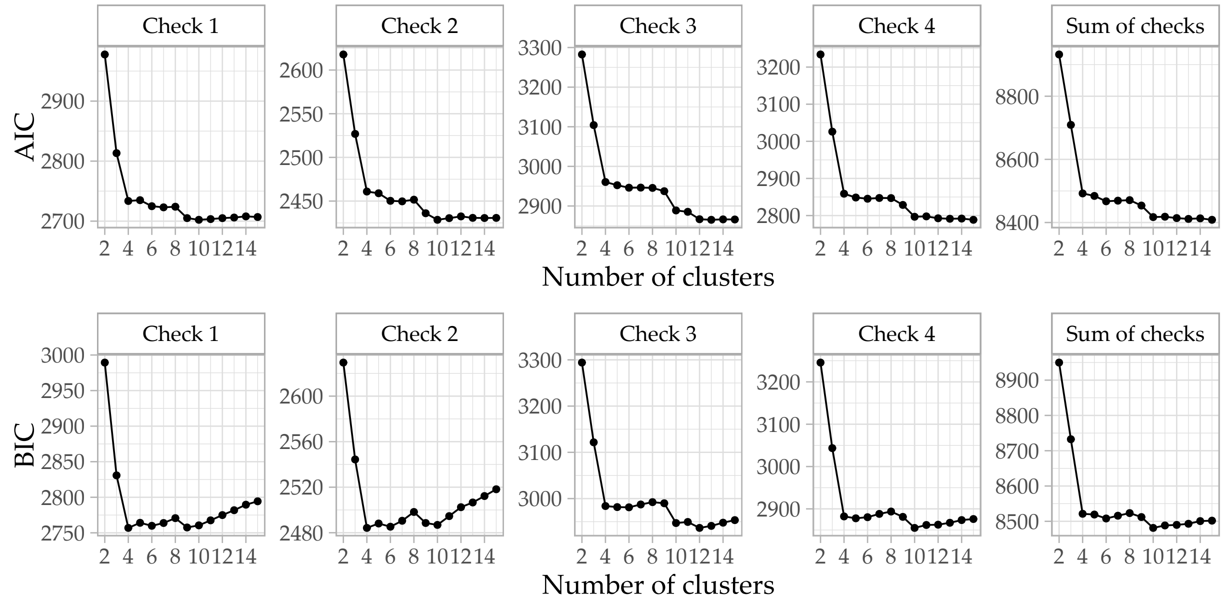


Figure 7.M1: AIC and BIC for distance-based clusters for Read et al. (2021) data.

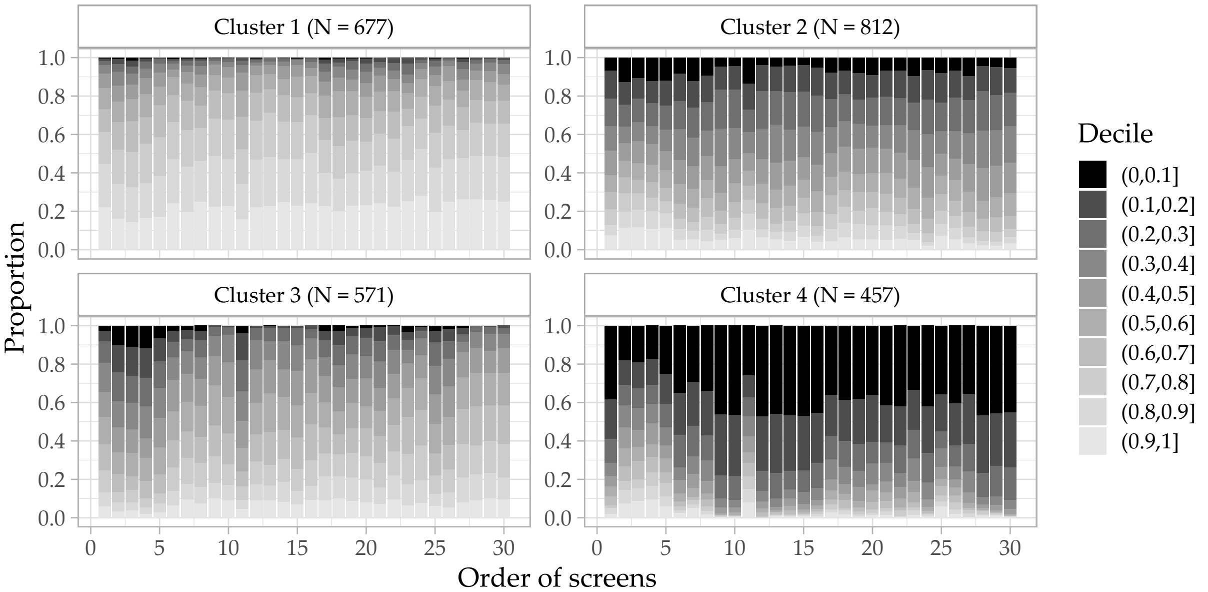


Figure 7.M2: Duration decile composition of 4-cluster solution for Read et al. (2021) data.

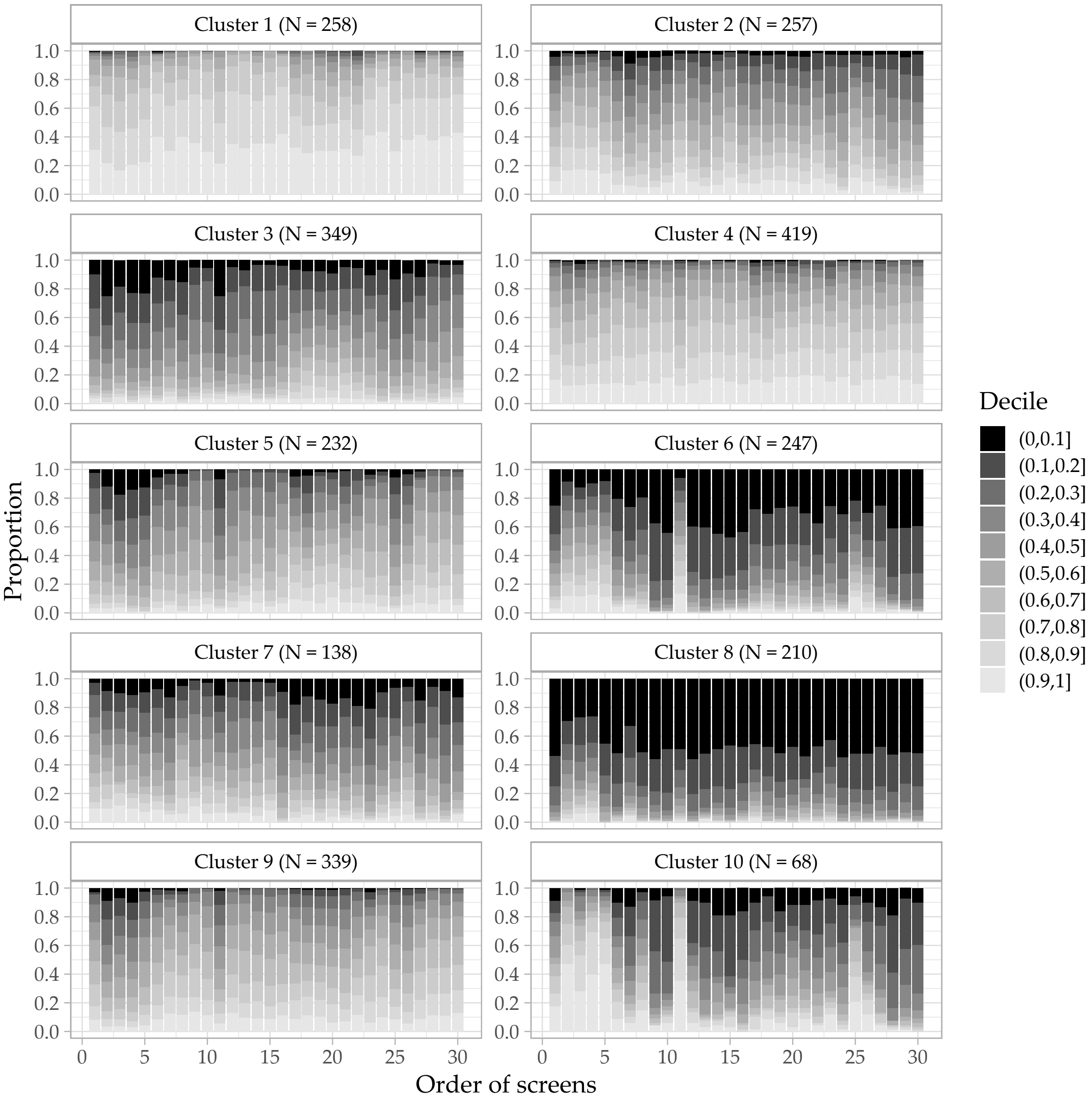


Figure 7.M3: Duration decile composition of 10-cluster solution for Read et al. (2021) data.

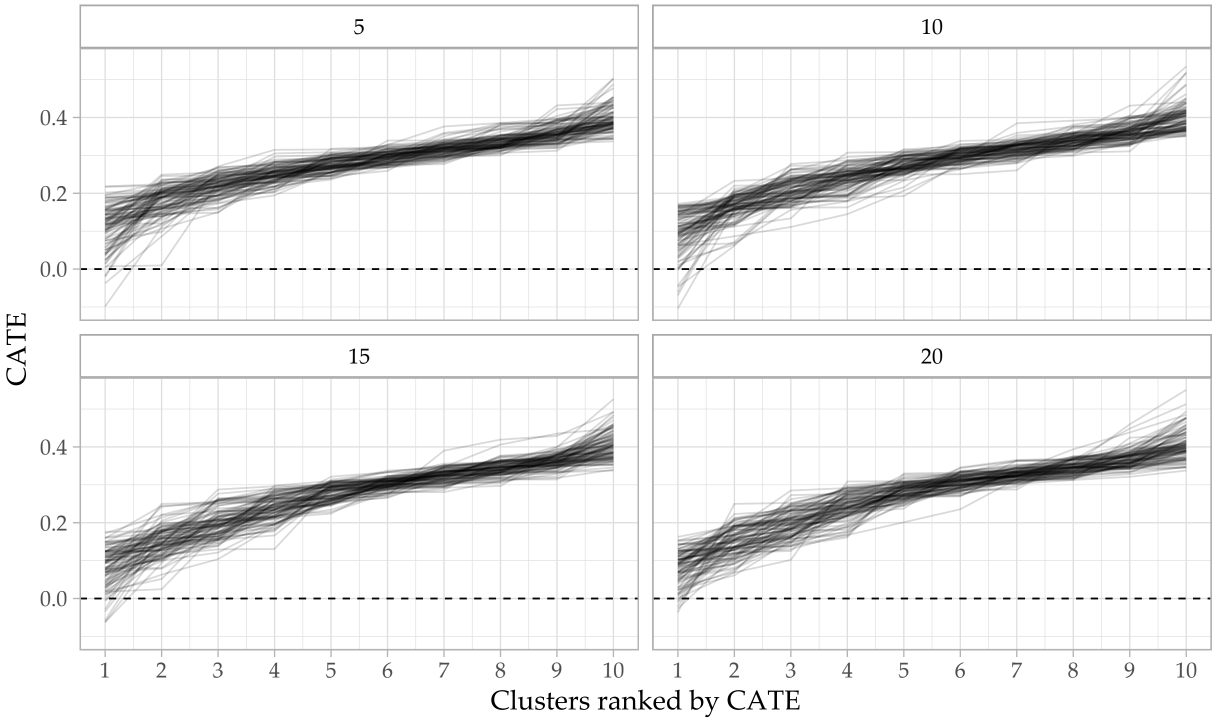


Figure 7.M4: CATEs for different subsets of screens (100 draws for each number of variables).

Part III

Concluding remarks

8 Concluding remarks

This dissertation consists of five articles on preventing and identifying deliberate errors in surveys. The implications and limitations of these are discussed in-depth in the respective chapters. This chapter provides a brief general conclusion and outlook on the remaining challenges concerning deliberate errors in surveys.

For face-to-face interviewers, the results of this dissertation show that deliberate errors are widely spread and can substantially affect substantive analyses. While preventive measures such as audio recordings can be effective, identification methods such as those developed in this dissertation are necessary and require further testing and validation. However, a lack of publicly available verified falsifications hinders the development of universally applicable detection approaches. To enable a transparent development and testing pipeline of identification methods, the research community would require a platform of datasets containing verified falsifications (Winker, 2016). This assumes general transparency about the occurrence of falsifications in surveys which is currently not always given. The difficulty of verifying falsifications is an additional complicating factor as statistical identification approaches only provide indications of deliberate interviewer errors and actual proof of such can only be provided by the fieldwork institute or the interviewers themselves. A further challenge is to define at which point interviewer behavior is too far from their instructions and guidelines. Should we only exclude interviewers who fabricate all their interviews or should we drop any interviewer who sometimes skips questions or parts of the questionnaire text (which would leave us with very few remaining interviewers) from the data? Such rules are still survey-specific and require in-depth knowledge about the target population and the questionnaire. Finally, researchers should be aware that many decisions they make during survey design (e.g., payment scheme, questionnaire length and difficulty) could in the end increase the burden on interviewers. Accounting for and reducing such burdens in the first place can substantially reduce the risk of deliberate errors, in particular during times of decreasing response rates and increasing scarcity of skilled personnel.

Concerning inattentive responding in web surveys, the main challenge that cannot be resolved is the lack of verified inattentive responding. Hence, identification methods as those developed and tested in this dissertation can always only provide indications. As a result, no statistical identification approach can unambiguously identify all inattentive respondents in a survey, instead, different approaches will lead to different results. Given the ubiquity of inattentive respondents in non-probability panels and the increasing use of these panels in academic research, reporting guidelines on how inattentive responding was handled and how it affects results should be established (Berinsky et al., 2024). Moreover, as with face-to-face surveys, researchers should be aware that their decisions – in particular with regard to the questionnaire – will influence respondent burden and the prevalence of inattentive responding.

Despite all the potential deliberate errors in surveys, researchers should always be grateful to interviewers and respondents for taking on all the challenges involved with collecting data for advancing social science research. In particular, because the majority of actors involved in surveys do not engage in behaviors discussed in this dissertation.

Literature

Berinsky, A. J., Frydman, A., Margolis, M. F., Sances, M. W., & Valerio, D. C. (2024). Measuring attentiveness in self-administered surveys. *Public Opinion Quarterly*, *88*(1), 214–241.

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

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, §8 Abs. 2 Pkt. 5)

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

München, den 22. Oktober 2024

Lukas Olbrich