OESTERREICHISCHE NATIONALBANK
EUROSYSTEM

WORKING PAPER 253

Multivariate assessment of interviewer-related errors in a cross-national economic survey

Lukas Olbrich, Elisabeth Beckmann, Joseph W. Sakshaug

The *Working Paper series of the Oesterreichische Nationalbank* is designed to disseminate and to provide a platform for discussion of either work of the staff of the OeNB economists or outside contributors on topics which are of special interest to the OeNB. To ensure the high quality of their content, the contributions are subjected to an international refereeing process. The opinions are strictly those of the authors and do in no way commit the OeNB.

The Working Papers are also available on our website (http://www.oenb.at) and they are indexed in RePEc (http://repec.org/).

Publisher and editor *Oesterreichische Nationalbank*
Otto-Wagner-Platz 3, 1090 Vienna, Austria
PO Box 61, 1011 Vienna, Austria
www.oenb.at
oenb.info@oenb.at
Phone (+43-1) 40420-6666
Fax (+43-1) 40420-046698

Editor *Martin Summer*

Cover Design *Information Management and Services Division*

DVR 0031577

ISSN 2310-5321 (Print)
ISSN 2310-533X (Online)

Multivariate assessment of interviewer-related errors in a cross-national economic survey

Lukas Olbrich*[1,2], Elisabeth Beckmann[3], Joseph W. Sakshaug[1, 2, 4]

[1]*Institute for Employment Research (IAB)*
[2]*LMU Munich*
[3]*OeNB*
[4]*University of Mannheim*

February 2024

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-related error and single-country cross-sectional surveys. We extend this literature by investigating the influence of interviewers 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 first analyze the extent of interviewer variance in financial literacy measures and how these effects compare to other questionnaire items. Building on the previous literature, we also evaluate the stability of these estimates over time and across countries. Second, we apply several data quality indicators on various dimensions of interviewer-related error and investigate country-years with particularly exceptional patterns. Finally, 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.

Keywords: interviewer effects, survey data quality, multilevel modeling, interviewer falsification, interviewer variance

*Corresponding author (e-mail for correspondence: lukas.olbrich@iab.de). The authors would like to thank Melanie Koch, Katharina Allinger, Julia Woerz, Helmut Stix, Thomas Scheiber, participants of the WAPOR 2023 conference, and an anonymous reviewer for their helpful feedback and comments. This project was funded by a Klaus-Liebscher-Economic-Research-Scholarship (OeNB).

Non-technical summary

Interviewers play a key role in face-to-face surveys. Among others, they are involved with sampling respondents, contacting respondents and convincing them to participate, and conducting the interviews. Though interviewers are usually trained to fulfill these tasks and receive detailed instructions, interviewers may induce errors during sampling, contacting, and interviewing. These errors can be intentional (i.e., if interviewers administer only parts of the questionnaire and fill in the rest themselves) or unintentional (i.e., if respondents are influenced by the interviewers' appearance) in nature.

In this paper, we analyze interviewer-related errors in the OeNB Euro Survey, an annual cross-national face-to-face survey. We use data collected over ten years in ten Central, Eastern, and Southeastern European countries. These data are particularly valuable since they allow for investigating changes over time and differences across countries concerning interviewer-related errors.

We first estimate interviewer variability in the responses to financial literacy questions. As shown by previous research, these questions are particularly prone to interviewer influences as these questions have *correct* responses and interviewers may help respondents answer them. We replicate these results in a cross-country and longitudinal context. Second, we test for the prevalence of (near-)duplicates of interviews within country-years which may arise from interviewer influences or errors by higher-level staff. Third, we assess whether interviewers have exceptional daily workloads. Fourth, we analyze item nonresponse (i.e., the proportion of "Don't know" or "No response" responses) in each country-year and estimate corresponding interviewer variability. Fifth, we investigate the variability of responses within item batteries (i.e., sets of questions with identical response scales). Lastly, we approximate unit nonresponse bias by calculating the proportion of female respondents in gender-heterogeneous couples living in the same two-person household (the expected proportion being 0.5). This measure has been used to estimate interviewer influences on selection in previous studies. To combine all the analysis approaches, we rely on a multivariate tree-based outlier detection approach to flag either country-years or interviewers with exceptional results.

We find that all measures of interviewer errors vary substantially over time and across countries. We observe that financial literacy items are particularly prone to interviewer variability, which holds across countries. For the (near-)duplicate analysis, we find that one country has exceptional proportions of near-duplicates in several years. Further investigations reveal that these near-duplicates are restricted to regions assigned to two field supervisors which are the likely source of the irregularities. Regarding daily workloads, we find several implausible workloads in earlier rounds of the OeNB Euro Survey, likely driven by documentation errors. For item nonresponse and straightlining, we find substantial variation over time in several countries and sizeable interviewer variability. Lastly, the unit nonresponse estimates deviate from 0.5 in multiple country-years. The multivariate outlier detection approach serves as a viable method to efficiently identify suspicious interviewers and country-years.

The results highlight the necessity of quality controls in interviewer-administered surveys, in particular in cross-country surveys where multiple survey institutes are involved. For applied researchers, our results emphasize the importance of accounting for interviewer influences in their analyses.

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 statistics in multiple ways, with the sampling (if interviewers are involved), recruitment, and measurement stages being particularly prone to interviewer influences (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. Intentional interviewer deviations are labeled *interviewer falsification* (AAPOR, 2003) and may induce error at all three of these stages (DeMatteis et al., 2020).

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 influences into perspective, identifying countries that require more or less caution concerning interviewers, and identifying correlates of interviewer influences. These studies include research on interviewer variance in substantive questionnaire items and biomeasure collection (e.g., Waldmann et al., 2023; Zins & Burgard, 2020), interviewer variance in data quality indicators such as straightlining or interview duration (e.g., Loosveldt & Beullens, 2013a, 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 potential influences of interviewers.

In this study, we investigate the influences of interviewers 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 influences 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 variance (Crossley et al., 2021).

Our analysis proceeds in multiple steps. First, we replicate Crossley et al. (2021) and evaluate to which extent interviewers influence financial literacy measures and how these effects compare to other questionnaire items, develop over time, and differ across countries. Second, we apply several measures of data quality (straightlining, item nonresponse, daily interviewer workload, (near-)duplicates) and investigate country-years with particularly exceptional outcomes. Finally, we use isolation forests to flag country-years and interviewers with outlying values based on the data quality measures. Furthermore, we explore the extent to which interviewer-related errors can impact substantive analyses.

Our analysis approach and results can serve as a guide for both data users of the OeNB Euro Survey and researchers interested in evaluating data quality in other surveys. For data users, our results indicate for which items, countries, and years researchers should be cautious and may want to implement robustness checks concerning interviewer influences. Researchers studying other surveys may use the proposed analysis approach as a template to learn about potential sources of interviewer errors that exist in those surveys. The analysis approach can be easily adapted to other surveys.

2 Interviewers in cross-national surveys

As the literature on interviewers and their role in survey data collection is abundant (see West & Blom, 2017, for a review), we provide only a brief overview of previous studies using cross-national data. Note, however, that we repeatedly refer to studies on interviewer effects in other settings in the methods section.

Loosveldt and Beullens (2013a) focused on the interview length in the 4th and 5th rounds of the European Social Survey (ESS) and found that interviewers account for up to 90 percent and 60 percent of the variance, respectively, in both rounds. Loosveldt and Beullens (2013b) analyzed durations for several questionnaire sections in the 5th round

of the ESS. Across twelve countries, they find that interviewers account for considerable variation in the section durations (ranging from 2.5 to 24.8 percent). Loosveldt and Beullens (2017) investigated straightlining which is the tendency to give identical (or nearly identical) responses to same-scaled items in an item battery (Yan, 2008) in the 6th round of the ESS. They find that interviewers account for up to 20 percent of the variance in their straightlining indicators, although they observe substantial heterogeneity across countries. Lastly, Vandenplas et al. (2018) jointly consider 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. Given that proportions above 10 percent are rare in the survey literature (e.g., Groves, 2004), these studies show that interviewer effects in cross-national surveys can be substantial. Blasius and Thiessen (2021) also used ESS data to identify suspicious interviewers using Categorical Principal Components Analysis (CatPCA). They identify several countries and interviewers with anomalous response patterns and find a strong correlation between their data quality measures and a corruption index.

Another strand in the literature focuses on the role of interviewers on unit nonresponse bias in cross-national surveys. To measure unit nonresponse bias, most studies rely on internal criteria derived from the survey data that do not require external validation (Sodeur, 1997, see section 4.2.5 for a detailed description). Kohler (2007) used data from six cross-national surveys and found that correlates of higher bias (due to sampling method, backchecking procedures, substitutions) are driven by interviewer behavior. 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) showed that more interviewer involvement is correlated with higher bias and is a key mediator for the relationship between response rates and nonresponse bias.

Several studies also focus on the potential influences of higher-level employees and

survey institutes by investigating (near-)duplicates and find evidence for manipulations in several surveys (e.g., Blasius & Thiessen, 2015; Koczela et al., 2015; Kuriakose & Robbins, 2016; Slomczynski et al., 2017). Although highly similar interviews could occur for reasons unrelated to data quality (Simmons et al., 2016), accumulations of nearly identical interviews might indicate copy-pasting parts of, or entire, interviews.

In rare instances, the prevalence and detection of fraudulent interviewers in cross-national surveys is documented. For example, Yamamoto and Lennon (2018) report on fabricated data in the Programme for the International Assessment of Adult Competencies (PIAAC) and the Programme for International Student Assessment (PISA) and Bergmann et al. (2019) report on interviewer falsification in the Survey of Health, Ageing and Retirement in Europe (SHARE).

The above studies investigate the role of interviewers in face-to-face surveys (with a particular focus on the ESS) from a variety of perspectives and document substantial heterogeneity across countries. In the present study, we combine and extend several of these approaches and use them as inputs to multivariate analysis methods.

3 Data

We use data from ten rounds (2012-2021) of the OeNB Euro Survey which is commissioned by the Austrian central bank (for detailed information on the OeNB Euro Survey, see https://www.oenb.at/en/Monetary-Policy/Surveys/OeNB-Euro-Survey.html). 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, 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 of the countries, an Austrian survey organization subcontracts opinion poll

[1]Slovakia was included in the survey until 2008.

institutes to conduct the survey. Samples are repeated cross-sections and over time the same regions and cities are covered. 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 the majority of 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 interviews in PAPI mode. 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 interviewers to more than 100 interviewers with an average of 70 interviewers (see A1 in the Appendix).

Nonresponse varies across countries and survey rounds. AAPOR RR1 response rates (AAPOR, 2016) are reported in Table B1 in the Appendix. 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 often coinciding with changes in the survey institute.

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.

4 Methods

In the forthcoming analysis, we evaluate data quality at the country-year and interviewer level. The latter analysis is conducted year-by-year to facilitate the application of the approach in future rounds of the OeNB Euro Survey. The subsequent sections describe the indicators used, their motivation, and their respective aggregation to the country-year and interviewer level.

4.1 Interviewer variance

Assuming random assignment of respondents to interviewers, respondents interviewed by the same interviewer might be more similar due to interviewers recruiting 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. The key variables for our interviewer variance analysis are financial literacy questions (Lusardi & Mitchell, 2008; Reiter & Beckmann, 2020). Using data from the German Panel on Household Finances, Crossley et al. (2021) document substantially higher interviewer variance for financial literacy questions than other survey items. Interviewers likely know the correct response to financial literacy questions and thus have additional capacity for influencing respondents. Previous research showed that interviewer knowledge can influence respondents' answers (Kerwin & Ordaz Reynoso, 2021). The OeNB Euro Survey contains four financial literacy questions[2] on inflation, interest rates, risk diversification, and exchange rates (see Table 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 additional

[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.

questions, 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. In line with Crossley et al. (2021)'s findings, we expect the interviewer variance for the financial literacy questions to be higher than for other variables.

The established approach to estimate interviewer variance is multilevel modeling (e.g., Brunton-Smith et al., 2017; Davis & Scott, 1995; O'Muircheartaigh & Campanelli, 1998; Schnell & Kreuter, 2005; Sturgis et al., 2021). The model for a continuous survey measure y is

$$y_{ijk} = \beta_0 + \theta_j + \mu_k + \varepsilon_{ijk} \tag{1}$$

where y_{ijk} is observed for each respondent i nested in interviewer j and primary sampling units (PSUs) k, β_0 is a constant, θ_j is each interviewer's effect on y and assumed to follow a normal distribution with mean zero and variance σ^2_θ. Similarly, the PSU effects μ_k are assumed to follow a normal distribution with mean zero and variance σ^2_μ. The residuals ε_{ijk} follow a normal distribution with mean zero and variance σ^2_ε. The interviewer variance is usually denoted by the intra-class (or intra-interviewer) correlation coefficient (ICC) that is calculated as $\sigma^2_\theta(\sigma^2_\theta + \sigma^2_\mu + \sigma^2_\varepsilon)^{-1}$ 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 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 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 Appendix A for more information). Further, we extend equation 1 by adding 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 (Hox, 1994). A detailed description of the control variables is provided in Table F1 of the Appendix. Interviewer characteristics (age, gender, experience) are not available for all years. Thus, we will only provide a brief discussion of their significance in Appendix E.

For binary outcome variables we fit multilevel logistic regression models denoted as:

$$log\left\{\frac{P(y_{ijk}=1)}{P(y_{ijk}=0)}\right\} = \beta_0 + \sum_m \beta_m x_m + \theta_j + \mu_k \tag{2}$$

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).

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

Table 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.

comparability, we only include the items for which all countries fulfill our criteria 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 error 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.

4.2 Indicators of data quality

4.2.1 (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. Note that high similarities within an interviewer's workload indicate interviewer-related 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). 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.

4.2.2 Daily interviews per interviewer

Another previously used indicator for flagging suspicious interviewers is the interviewers' daily number of successful interviews (Bushery et al., 1999). An interviewer's maximum number of successful interviews is restricted by the duration of interviews, unit nonresponse, the distance between respondents, and their daily working hours. Exceptionally high values indicate suspicious behavior.

Country-year level. We use the maximum number of daily interviews per interviewer within the respective country-year as an indicator.

Interviewer level. On the interviewer level, we rely on the maximum number of daily interviews per interviewer.

4.2.3 Item nonresponse

Item nonresponse 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 on 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 as a strategy to reduce effort (Crespi, 1945). Hence, both extremely high and low shares of item nonresponse are suspicious. 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 ($\mathbb{I}[item\,nonresponse > 0.05]$) and fit separate multilevel logistic regressions as denoted in Equation 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.

4.2.4 Straightlining

As the second satisficing indicator, we use straightlining which is a widely used indicator of data quality and has been applied to assess both respondent satisficing and interviewer effects (e.g., Kim et al., 2019; Krosnick, 1991; Loosveldt & Beullens, 2017; Olbrich et al., 2023). Lower data quality is indicated by a lack of, or low, variance of responses across same-scaled items since either respondents do not differentiate between item content or interviewers only ask a few items and fill the rest in themselves. 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. We apply a binary indicator of straightlining (1 = no variation in responses, 0 = variation in responses) since the complete lack of variation represents the most severe response style. 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.

4.2.5 Internal unit nonresponse bias indicator

While the previous indicators focused on measurement error, the next indicator is an internal criterion of 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 criteria 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. 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 both cultural and methodological sources of nonresponse bias (e.g., Eckman & Koch, 2019; Kohler, 2007; Menold, 2014; Rybak, 2023). In our analysis, we refrain from identifying correlates of nonresponse bias, but use the deviation from 50 percent as an indicator for interviewer error. 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).

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.

4.3 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.

5 Results

5.1 Interviewer variance

Figure 1 depicts the estimated ICCs for the four financial literacy questions and the financial literacy score. The estimated ICCs with 95 percent confidence intervals based on 200 bootstrap replications are provided in Tables G1 to G5. The ICCs for other survey items are depicted by the grey dots. 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).[3] We observe highly heterogeneous ICCs across financial literacy questions, countries, and years. We briefly discuss the most noteworthy developments below. In Croatia in 2013, the ICCs are substantially lower than in any other country and year for all variables. In the fall of 2013, the institute conducting the

[3] Figure I1 in the Appendix depicts boxplots of the ICC estimate for each country-year to ease the observation of trends and outliers.

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 inexperienced and 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 survey 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.[4] The ICCs for Hungary are among the highest for all countries. In particular, the ICCs for the interest rate question are always above 0.6, indicating substantial interviewer variance. Such ICCs are driven by a large proportion of interviewers collecting either only correct or only wrong answers to the respective financial literacy question. Poland is the only country in which we observe a steady decrease in ICCs over time. In Romania, we find a similar but less extreme reduction of ICCs in 2016 as for Croatia in 2013. This drop was also driven by a change in subcontractors. In 2016, the international survey organization collecting the data in Romania was a candidate for also taking over data collection in Bulgaria from 2017 onwards. Therefore, the quality of data collected in Romania in 2016 was undergoing intense scrutiny.

Compared to other survey 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 (out of 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 32 country-years, 28 country-years for the interest rate question, and 22 out of 80 country-years for the

[4]See information on methodology at OeNB Euro Survey.

risk diversification question; note, however, that the bootstrapped confidence intervals are relatively wide due to the small number of observations in each country-year.[5]

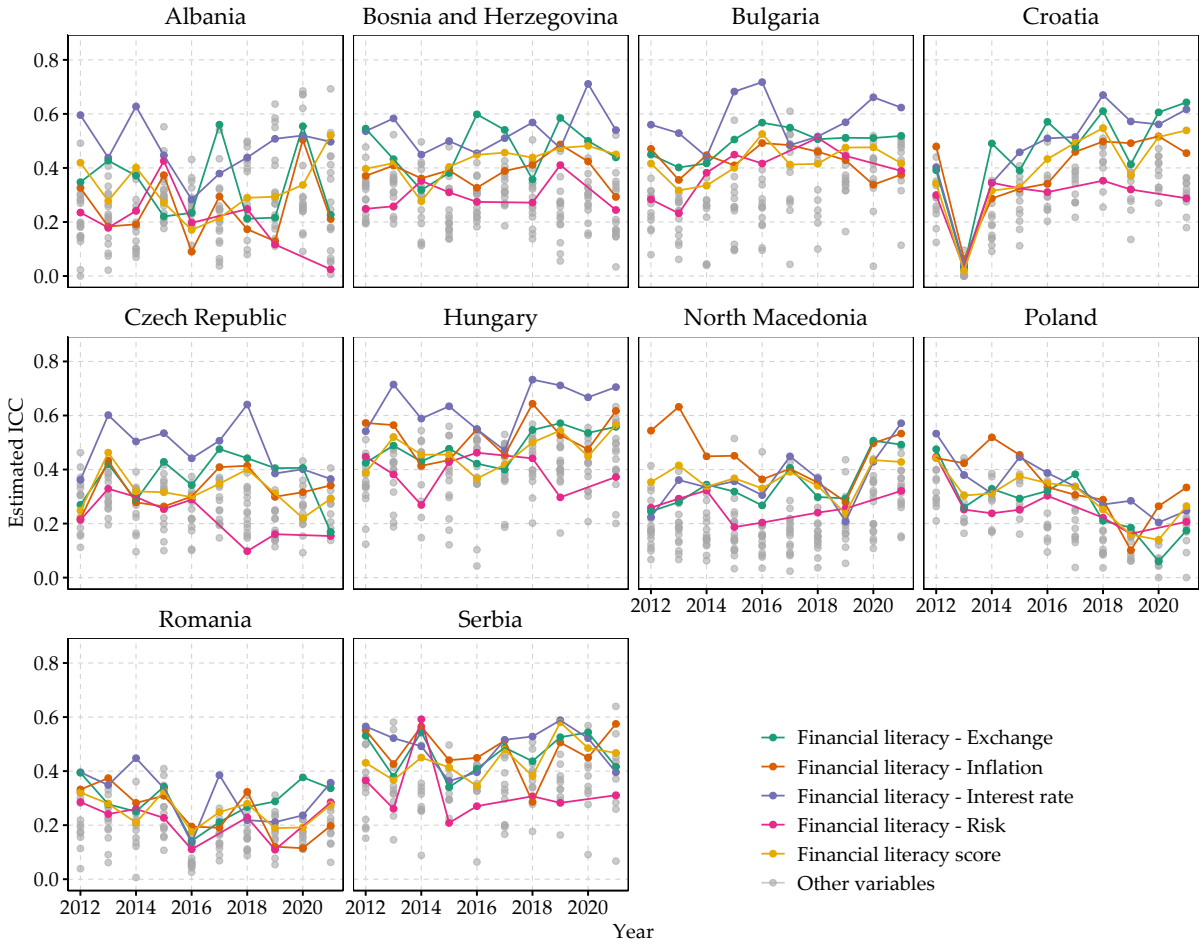

Figure 1: Estimated ICCs for financial literacy and other variables across countries and years.

Concerning the role of regional homogeneity, Figure H1 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 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).

[5]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).

5.2 Indicators of data quality

5.2.1 (Near-)duplicate analysis

For the (near-)duplicate analysis, Table 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. These results indicate that for several years in Albania, large proportions of interviews share an unusually high number of responses with other interviews.

Table 2: Proportion of observations flagged by mixture modeling approach

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)
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)
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)
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)

Notes: Proportion above 0.85 in parentheses.

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 2 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 (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 2c). 14 interviewers were involved in the largest component (green dots in Figure 2c). 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 2) show that high similarities only occurred from 2016 onwards, coinciding with the change to a new survey 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. 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 mostly conducted in PAPI mode in these years, such cases are likely driven by ex-post filling-in the respective data or a lack

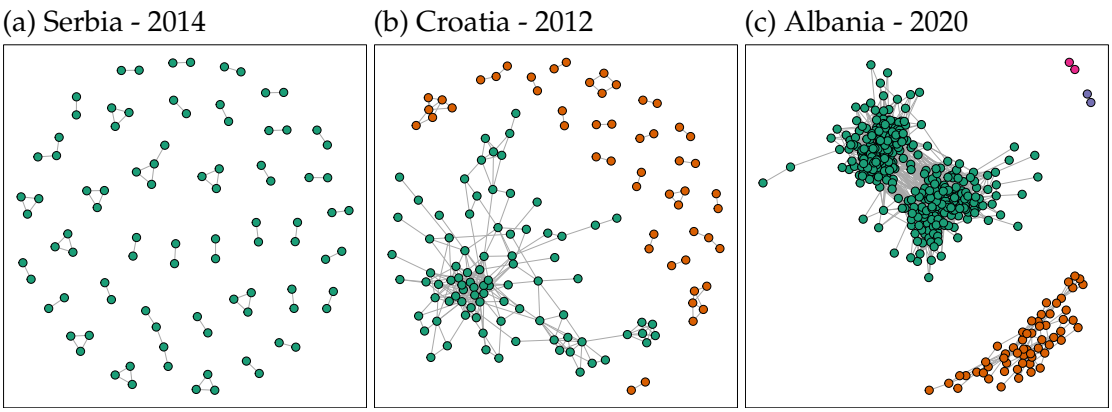

Figure 2: Fruchterman-Reingold layout plots. Each point corresponds to an interview that has at least one match exceeding the mixture modeling-based threshold. Interviews that are a match are connected by grey lines.

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.

5.2.2 Interviewer workload

The distribution of the daily number of successful interviews per interviewer varies substantially over time and countries. 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.

5.2.3 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 C1 in the Appendix). 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 C3 of the Appendix. With some exceptions, the estimated ICCs vary between 20 and 60 percent. These consistently high values indicate substantial interviewer variance. 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 again 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 survey items (see section 5.1).

5.2.4 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 C2 in the Appendix). 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.

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, 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.

5.2.5 Internal unit nonresponse bias indicator

Replicating Kohler (2007), the proportion of female respondents in gender-heterogeneous households for each country in each year is shown in Figure 3. 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 survey institute. In Bosnia and Herzegovina, values are close to 50 percent from 2012 to 2016, drop to around 35 percent in 2017 and 2018, and increase again to above 50 percent afterward. In Croatia, the proportions increase up to 68 percent in 2021, indicating an increase in bias. In North Macedonia, the proportions are beyond critical values only in the years 2016 to 2019 with values around 60 percent. From 2019 to 2020, the survey institute changed and the proportions went back to around 50 percent again.

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 recruitment. Given that interviewers play the most important role in random route surveys, interviewers are likely the main drivers behind these deviations from 50 percent. Furthermore, the estimates point to the importance of the survey institutes as switching survey institutes can lead to substantial changes.[6]

[6]We also assessed whether the proportion of female interviewers is related to the proportion of female respondents in gender-heterogeneous households and find a moderate association (Pearson correlation coefficient: 0.435), which is in line with the "liking" hypothesis that posits similarities in socio-demographic characteristics increase cooperation rates (Durrant et al., 2010; Groves et al., 1992).

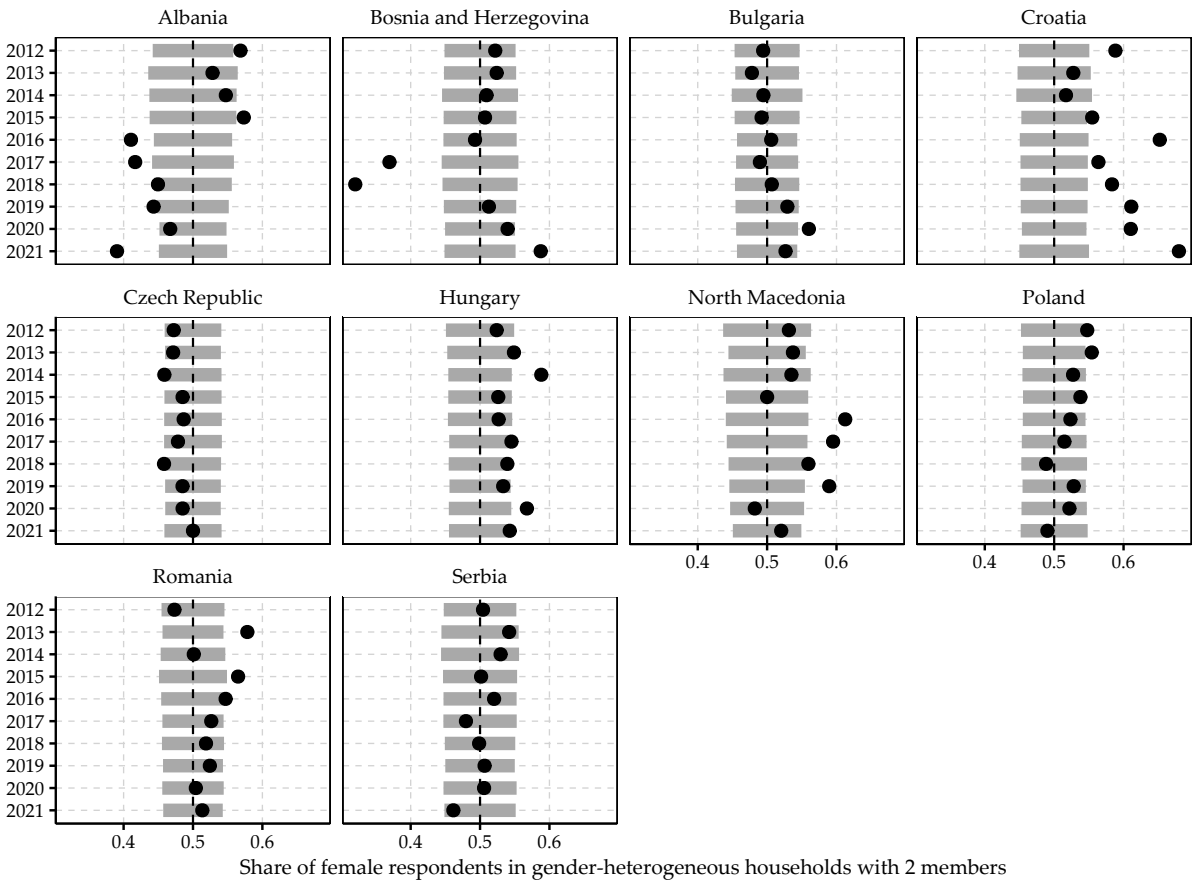

Figure 3: Share of female respondents in gender-heterogeneous households with two members across countries and years with confidence interval.

5.3 Isolation forests for outlier detection

The previous sections provide a detailed description of interviewer variance and 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.

5.3.1 Country-year level analysis

The country-year level analysis is based on nine indicators. In total, ten countries have scores above the outlier threshold of 0.5 (see Figure J1 in the Appendix). 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 4. 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 is flagged due to the ICCs in 2013. For Poland in 2020, Romania in 2016, and Croatia in 2021, several indicators contribute to the outlying score. Like Croatia in 2013, Romania in 2016 is flagged due to the low ICCs.

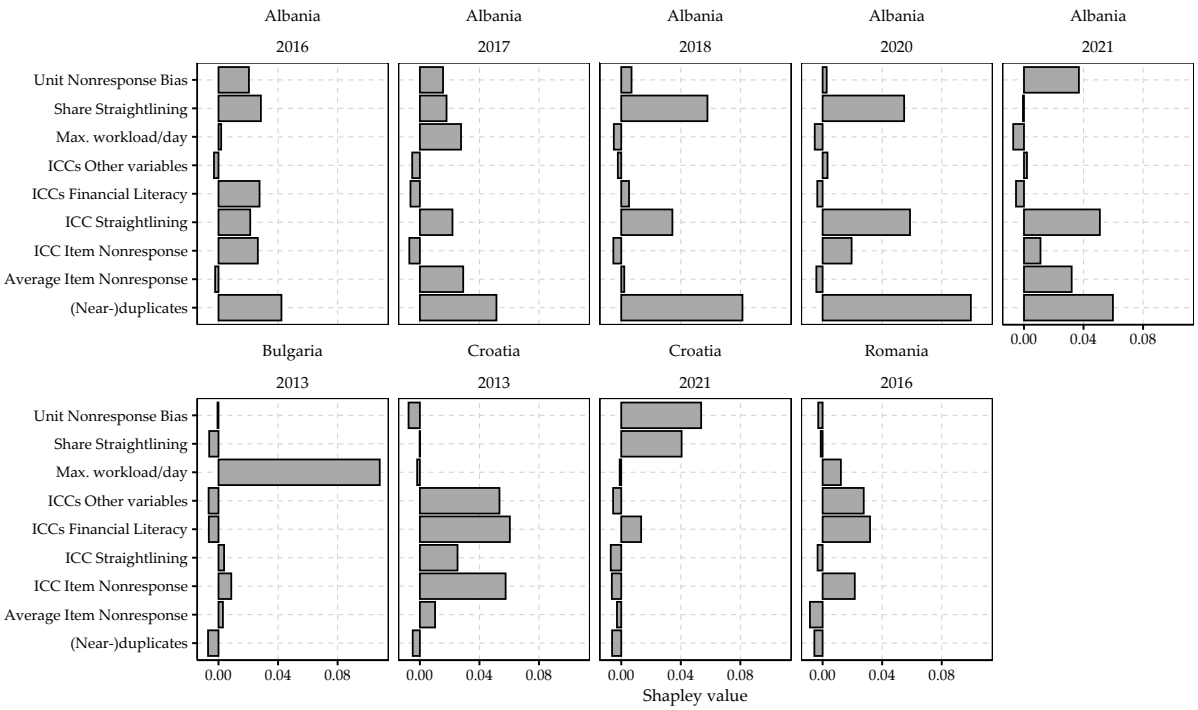

Figure 4: Shapley values for countries with scores above 0.5.

5.3.2 Interviewer-level analysis

For the interviewer-level analysis, we use six indicators and estimate an isolation forest for each year separately (see Figure J2 in the Appendix for the distribution of the scores in each year). As reported in Table 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 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 5 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.

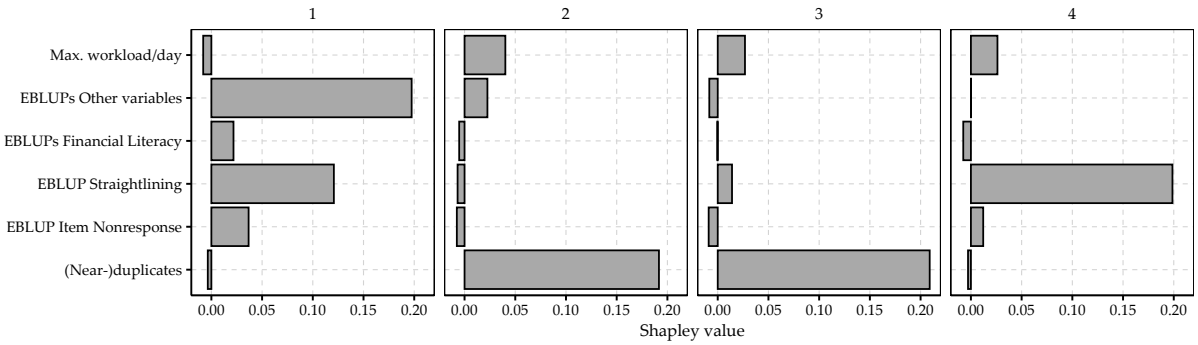

Figure 5: Shapley values for four interviewers with highest scores in 2020.

6 Impacts on substantive analyses

To explore the extent to which the described interviewer variance and 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 briefly discuss how interviewer-level clustering affects inference for the most recent OeNB Euro Survey round (2021).

Table 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.

Table 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 4.1 and the data and results from 2021 for the financial literacy score, Table 5 reports the average interviewer workloads, the ICC, the *deff*, and the *deft*. 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 5: Interviewer effects on financial literacy score in 2021.

Country	N	Interviewer workload	ICC	deff	deft
Albania	1000	34.483	0.521	18.439	4.294
Bosnia and Herzegovina	979	16.317	0.450	7.896	2.810
Bulgaria	995	9.851	0.417	4.695	2.167
Croatia	1008	13.263	0.539	7.610	2.759
Czech Republic	1000	20.408	0.292	6.667	2.582
Hungary	996	10.375	0.565	6.300	2.510
North Macedonia	974	17.088	0.428	7.887	2.808
Poland	1000	11.111	0.265	3.676	1.917
Romania	1027	10.926	0.275	3.729	1.931
Serbia	1007	14.594	0.467	7.351	2.711

Notes: Interviewer workloads based on sample sizes for multilevel models.

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 effects from various perspectives in ten rounds of a cross-national survey. First, we estimated the interviewer variance in various survey items with a special focus on financial literacy questions and found that the latter are particularly prone to interviewer variability (in line with Crossley et al., 2021). Second, we implemented several data quality indicators related to interviewer error (near-duplicates, interviewer workloads, item nonresponse, straightlining, internal unit nonresponse bias measure) and identified multiple country-years with suspicious values. To facilitate the efficient identification of outlying cases, we combined the variance estimates and 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. Our follow-up analyses indicate that is a rather local issue related to supervisors. 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. 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-related 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 showed an improvement over time, while we observed no clear changes in other countries. Our results also highlight the importance of the contracted survey institutes (Blasius & Sausen, 2023; Blasius & Thiessen, 2015, 2021). In various analyses, we observed substantial changes when the contracted survey institute in the respective country changed. Some changes led to improved quality, while others resulted in a decline. Interviewer-related errors can only be reduced if interviewers are made aware of standardized guidelines, are thoroughly trained, 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-related errors are unavoidable and will lower data quality.

Our results are subject to several restrictions. 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 identify the source of suspicious data, i.e., whether the interviewer, supervisors, or the survey 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 survey 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.

Literature

AAPOR. (2003). Interviewer falsification in survey research: Current best methods for prevention, detection, and repair of its effects.

AAPOR. (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (tech. rep.).

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.

Blasius, J., & Sausen, L. (2023). Detecting Fabricated Interviews Using the Hamming Distance. *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 and Research*, *50*(2), 740–777.

Brooks, M., E., Kristensen, K., van Benthem, K., 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 Methods Section*, 316–320.

Christensen, R. H. B. (2022). Ordinal-Regression models for ordinal data.

Cortes, D. (2023). *Isotree: Isolation-based outlier detection*. Manual.

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.
DeMatteis, J. M., Young, L. J., Dahlhamer, J., Langley, R. E., Murphy, J., Olson, K., & Sharma, S. (2020). *Falsification in surveys*. AAPOR.
Durrant, G. B., Groves, R. M., Staetsky, L., & Steele, F. (2010). Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys. *Public Opinion Quarterly*, *74*(1), 1–36.
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.
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. https://doi.org/10.1002/spe.4380211102
Greenwell, B. (2021). *Fastshap: Fast approximate shapley values*. Manual.
Groves, R. M. (2004). *Survey errors and survey costs*. John Wiley & Sons, Inc.
Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding The Decision to Participate in a Survey. *Public Opinion Quarterly*, *56*(4), 475.
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.
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.
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. (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.

Lusardi, A., & Mitchell, O. S. (2008). Planning and Financial Literacy: How Do Women Fare? *American Economic Review*, *98*(2), 413–417.

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. (2023). Detecting Interviewer Fraud Using Multilevel Models. *Journal of Survey Statistics and Methodology*, smac036. https://doi.org/10.1093/jssam/smac036

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 (O. Scrivner, Ed.). *PLOS ONE*, *18*(3), e0283092.

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.

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.

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
    find PSU i with shortest distance to another PSU j
    merge i and j
    use midpoint between i and j as updated GPS coordinate
end while

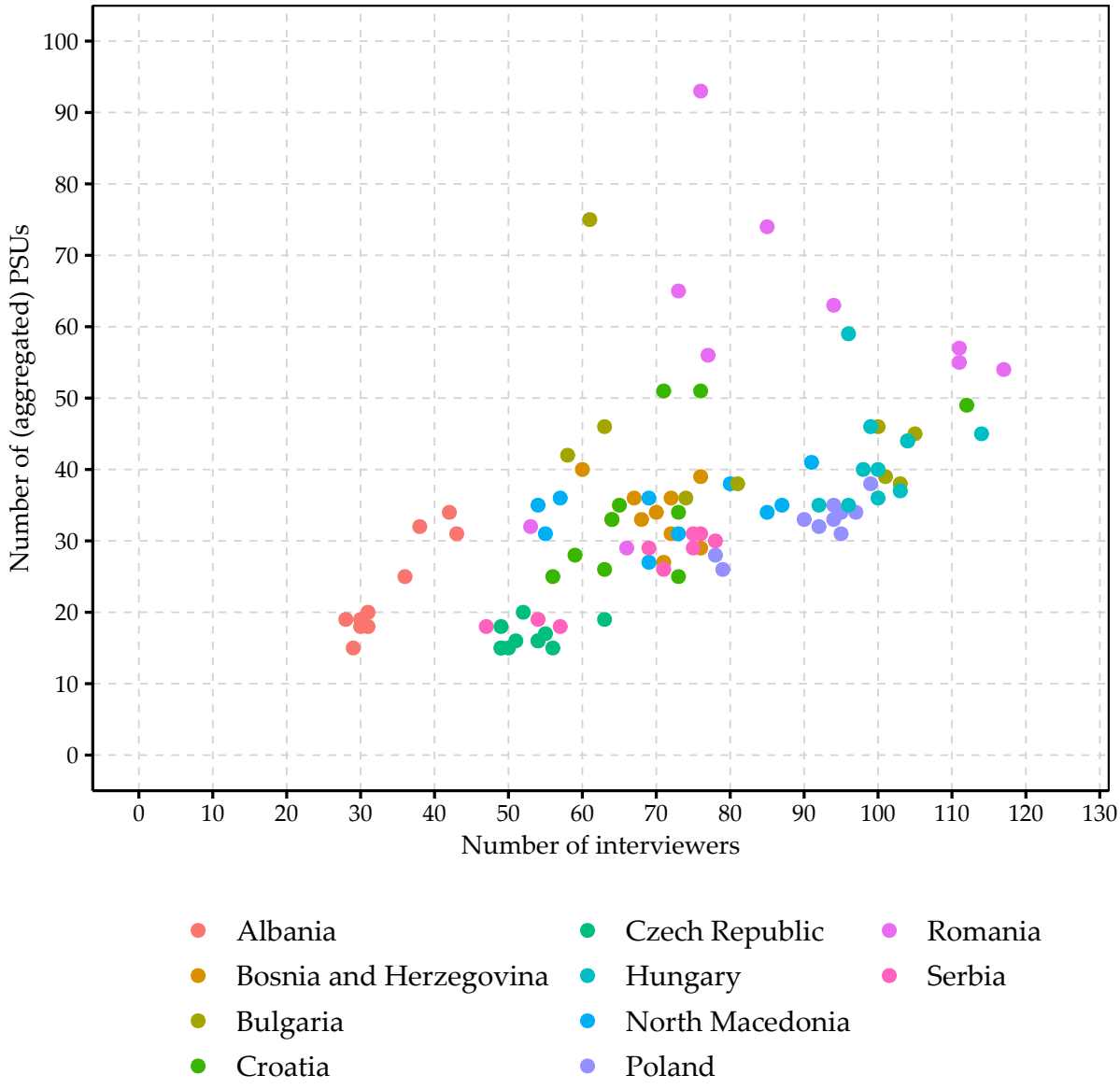

Figure A1: Number of interviewers and (aggregated) PSUs in each country-year.

B Response rates

Table 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.

C Item nonresponse and straightlining

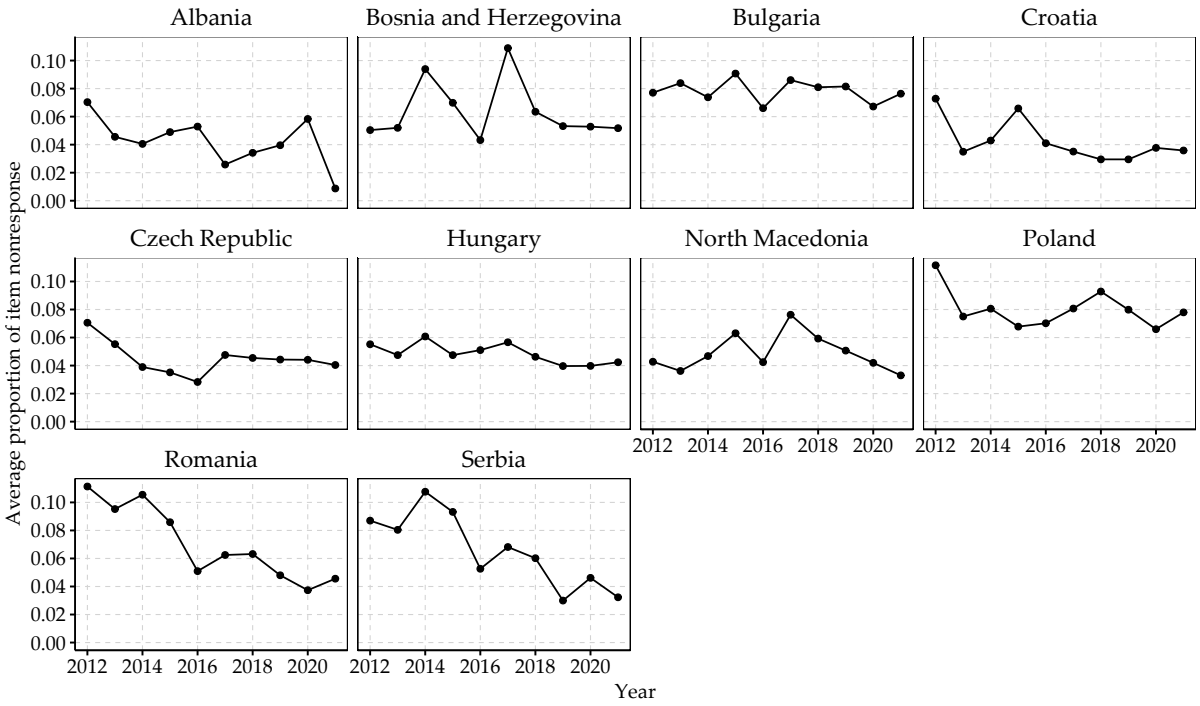

Figure C1: Average item nonresponse across countries and years.

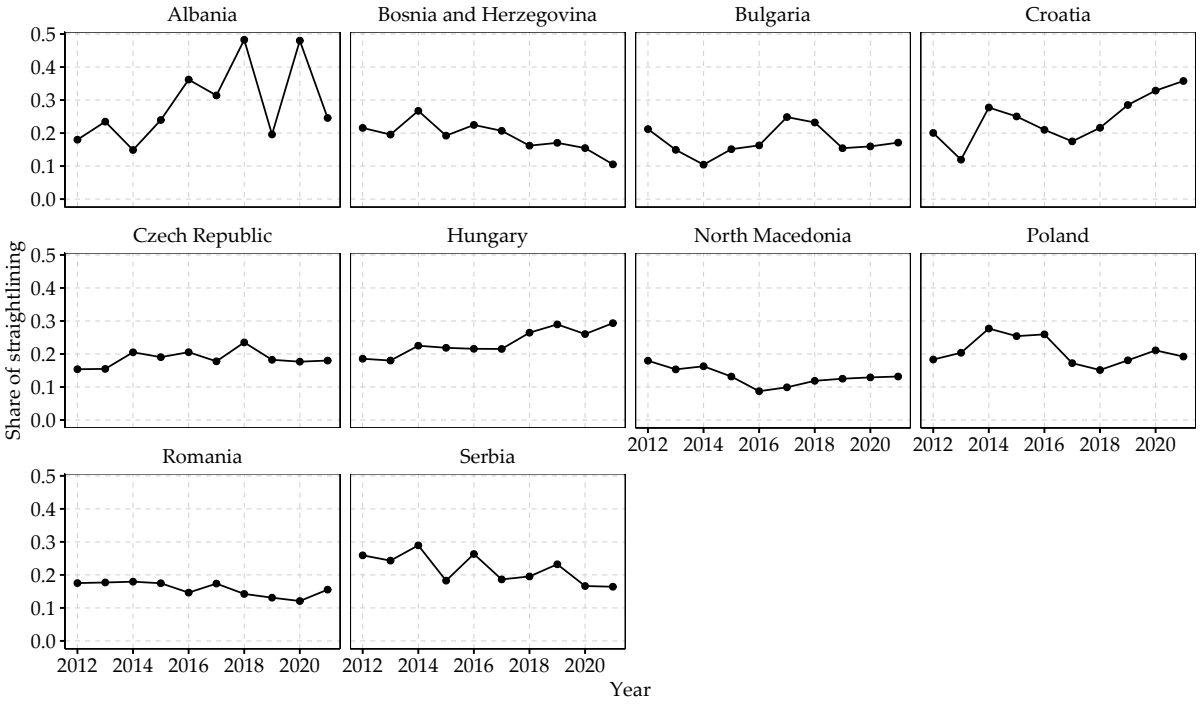

Figure C2: Average share of straightlining in the trust item battery across countries and years.

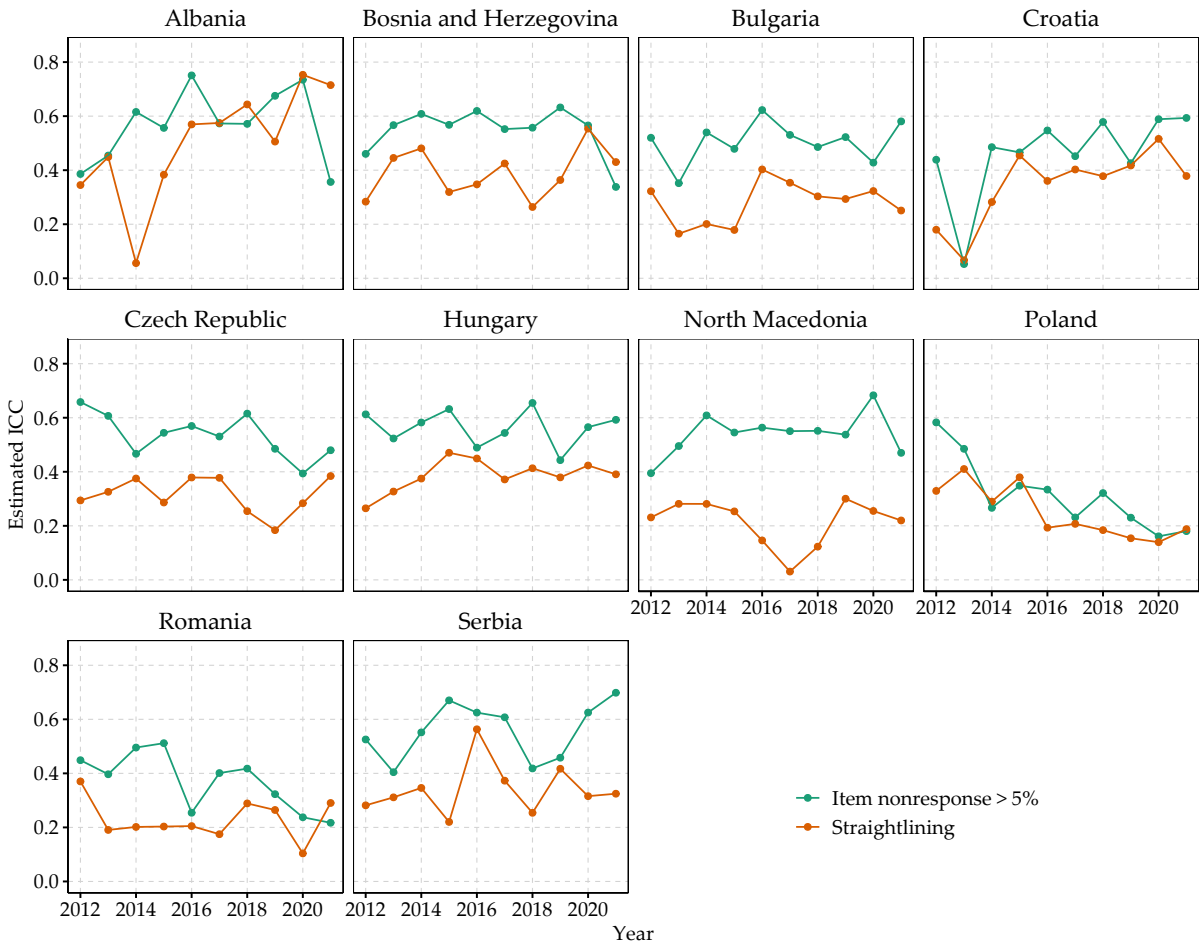

Figure C3: Estimated ICCs for item nonresponse and straightlining across countries and years.

D Change of survey 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 survey 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 survey 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.

D1 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 D1. 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 D1: Results for switching survey institutes.

Country	N pre	N post	N regions	N vars.	Mean ICC int	Mean 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.

E 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 E1 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 E2 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.

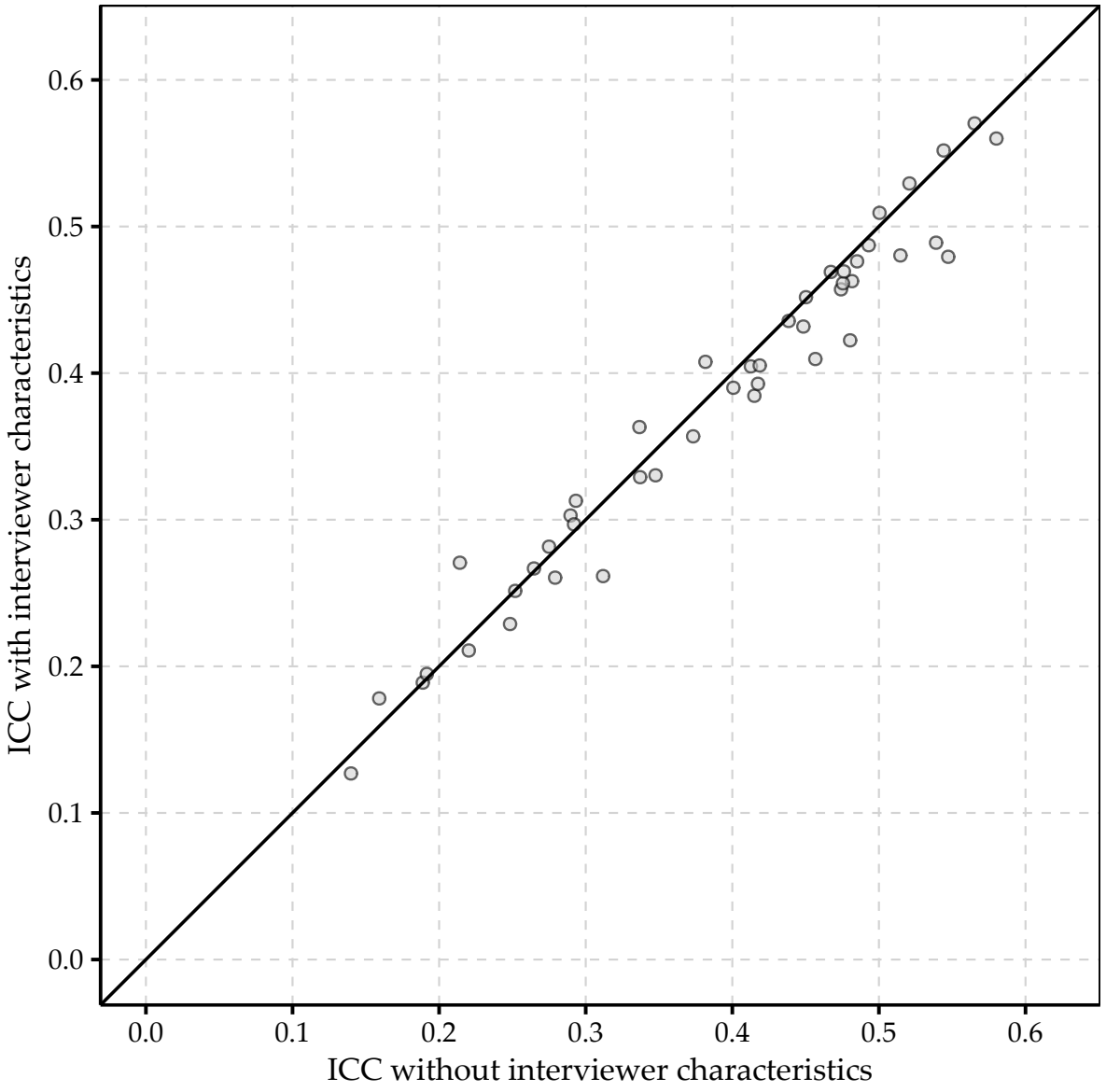

Figure E1: Estimated ICCs for financial literacy score with and without interviewer characteristics.

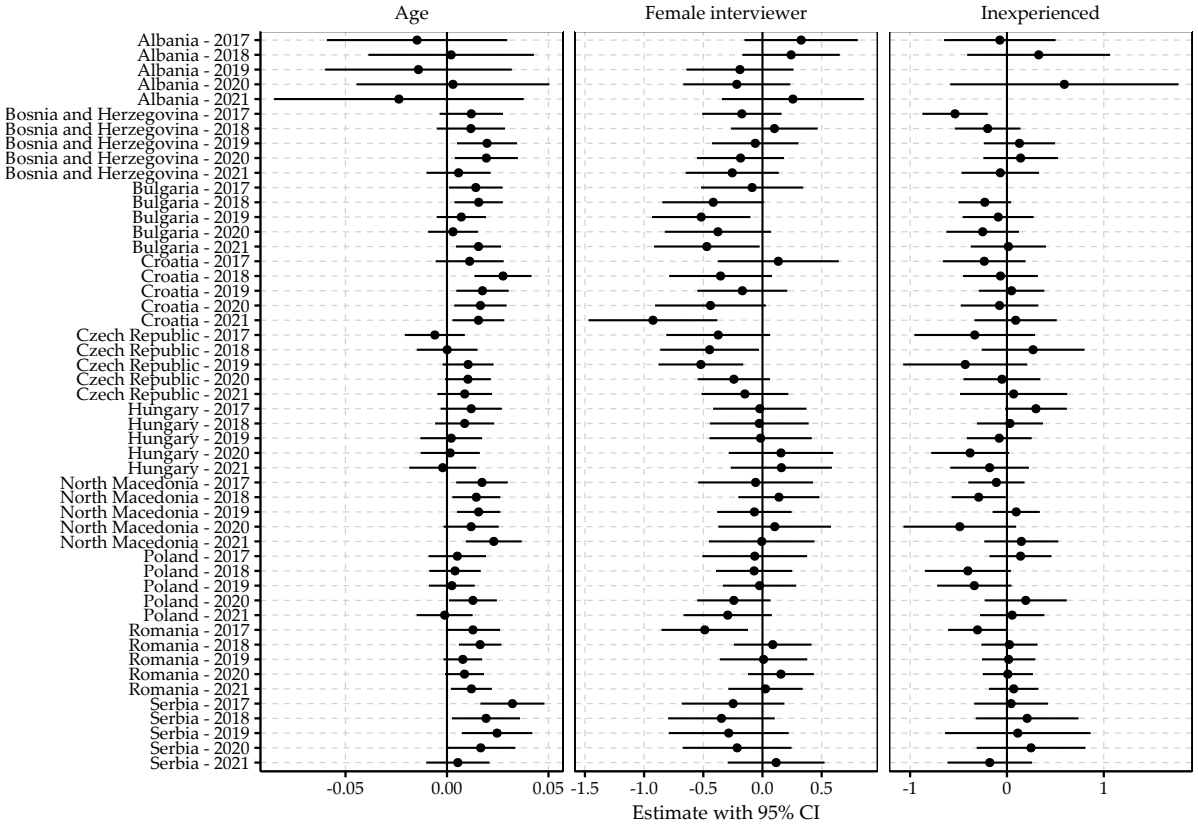

Figure E2: Estimated coefficients for interviewer characteristics in multilevel models with financial literacy score as the dependent variable.

F Control variables

Table F1: 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 $\geq$ 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 $\geq$ 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, $\geq$ 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

G ICC estimates with confidence intervals

Table G1: 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	0.536	0.583	0.449	0.499	0.454	0.510	0.568	0.473	0.711	0.540
Herzegovina	[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]
North Macedonia	0.224	0.361	0.335	0.356	0.305	0.449	0.367	0.208	0.429	0.571
	[0.104,0.367]	[0.206,0.466]	[0.186,0.472]	[0.221,0.480]	[0.191,0.390]	[0.305,0.556]	[0.237,0.471]	[0.095,0.328]	[0.293,0.539]	[0.391,0.679]
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]
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: Confidence intervals based on 200 bootstrap replications in brackets.

Table G2: 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	0.545	0.433	0.318	0.382	0.598	0.541	0.358	0.585	0.500	0.439
Herzegovina	[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]
North Macedonia	0.245	0.280	0.344	0.318	0.267	0.406	0.298	0.293	0.507	0.492
	[0.148,0.371]	[0.160,0.384]	[0.220,0.482]	[0.192,0.411]	[0.167,0.360]	[0.263,0.505]	[0.173,0.386]	[0.162,0.407]	[0.370,0.644]	[0.336,0.615]
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]
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: Confidence intervals based on 200 bootstrap replications in brackets.

Table G3: 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	0.371	0.409	0.361	0.391	0.327	0.389	0.411	0.487	0.424	0.293
Herzegovina	[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]
North Macedonia	0.544	0.632	0.449	0.451	0.364	0.403	0.352	0.280	0.497	0.533
	[0.402,0.723]	[0.500,0.722]	[0.309,0.589]	[0.314,0.545]	[0.237,0.476]	[0.263,0.505]	[0.211,0.449]	[0.161,0.409]	[0.353,0.607]	[0.379,0.635]
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]
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: Confidence intervals based on 200 bootstrap replications in brackets.

Table G4: 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]
North Macedonia	0.259	0.292	0.323	0.187	0.204	0.240	0.255	0.321
	[0.138,0.379]	[0.172,0.370]	[0.192,0.444]	[0.081,0.266]	[0.085,0.311]	[0.119,0.354]	[0.130,0.367]	[0.190,0.429]
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]
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: Confidence intervals based on 200 bootstrap replications in brackets.

Table G5: 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]	[0.058,0.267]	[0.094,0.316]	[0.149,0.424]	[0.169,0.451]	[0.184,0.472]	[0.332,0.620]
Bosnia and	0.397	0.418	0.277	0.404	0.449	0.457	0.438	0.474	0.482	0.450
Herzegovina	[0.291,0.488]	[0.311,0.523]	[0.176,0.405]	[0.283,0.496]	[0.343,0.543]	[0.323,0.550]	[0.327,0.541]	[0.363,0.569]	[0.396,0.572]	[0.344,0.552]
Bulgaria	0.416	0.317	0.335	0.400	0.526	0.413	0.415	0.475	0.476	0.417
	[0.317,0.523]	[0.196,0.409]	[0.222,0.428]	[0.287,0.512]	[0.435,0.610]	[0.326,0.510]	[0.323,0.510]	[0.382,0.558]	[0.376,0.562]	[0.333,0.500]
Croatia	0.342	0.020	0.316	0.329	0.433	0.493	0.547	0.373	0.515	0.539
	[0.221,0.453]	[0.001,0.079]	[0.200,0.435]	[0.207,0.425]	[0.325,0.542]	[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
	[0.139,0.361]	[0.334,0.571]	[0.225,0.418]	[0.172,0.416]	[0.182,0.422]	[0.221,0.451]	[0.292,0.508]	[0.166,0.392]	[0.124,0.310]	[0.184,0.386]
North Macedonia	0.354	0.415	0.336	0.368	0.330	0.391	0.340	0.237	0.435	0.428
	[0.236,0.456]	[0.304,0.513]	[0.233,0.438]	[0.268,0.435]	[0.222,0.408]	[0.317,0.473]	[0.229,0.435]	[0.149,0.330]	[0.310,0.538]	[0.310,0.511]
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]	[0.395,0.617]	[0.362,0.537]	[0.348,0.544]	[0.244,0.481]	[0.296,0.539]	[0.394,0.606]	[0.428,0.632]	[0.340,0.568]	[0.481,0.647]
Poland	0.443	0.304	0.312	0.375	0.349	0.337	0.252	0.159	0.140	0.265
	[0.315,0.531]	[0.212,0.421]	[0.216,0.422]	[0.292,0.465]	[0.264,0.446]	[0.219,0.443]	[0.166,0.329]	[0.081,0.233]	[0.073,0.212]	[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
	[0.226,0.411]	[0.206,0.364]	[0.126,0.295]	[0.217,0.413]	[0.098,0.253]	[0.151,0.354]	[0.183,0.389]	[0.103,0.276]	[0.107,0.269]	[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
	[0.255,0.523]	[0.243,0.496]	[0.267,0.573]	[0.244,0.506]	[0.252,0.436]	[0.356,0.564]	[0.234,0.514]	[0.465,0.655]	[0.350,0.559]	[0.356,0.549]

Notes: Confidence intervals based on 200 bootstrap replications in brackets.

H Interviewer and PSU effects

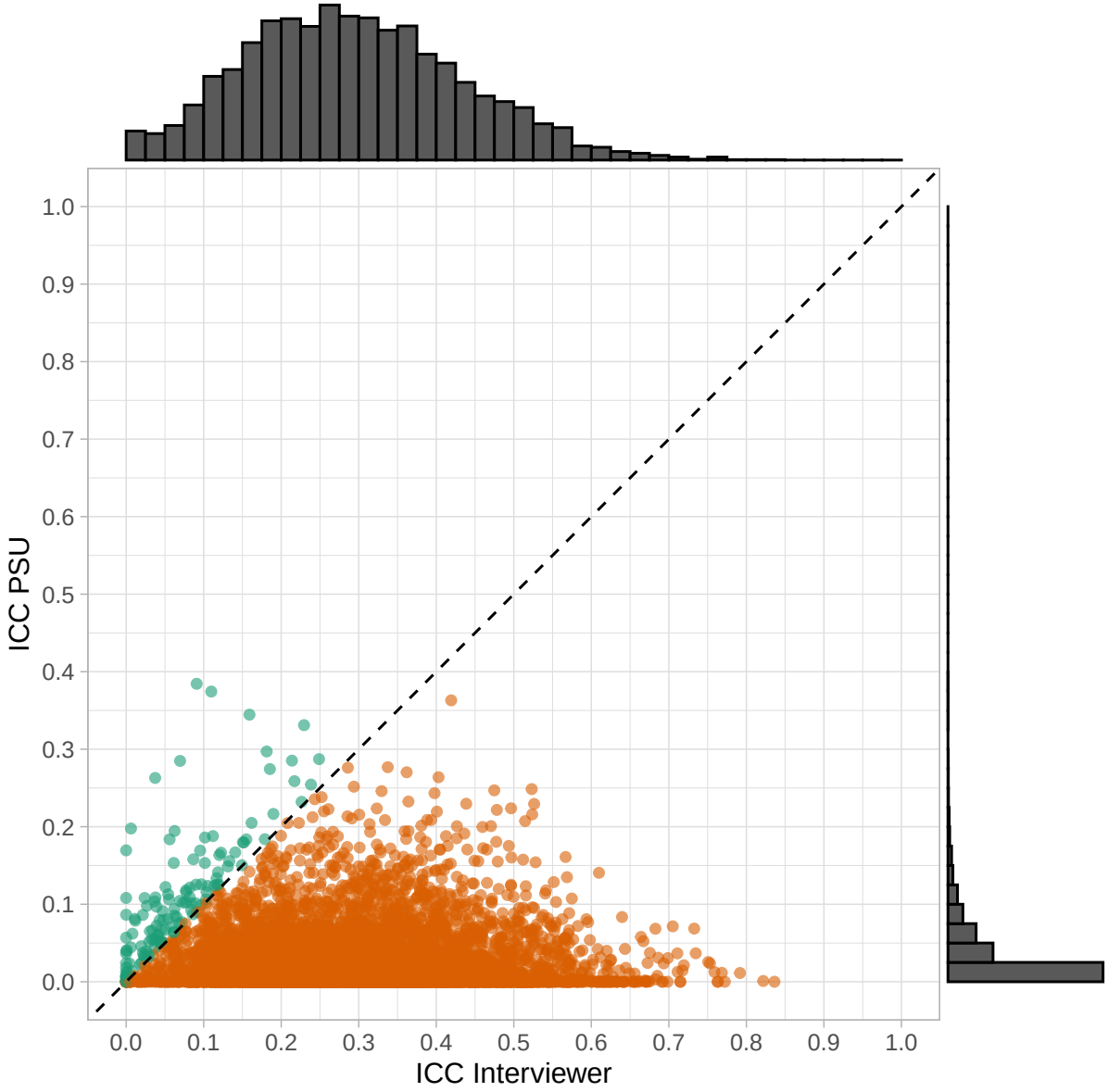

Figure H1: Interviewer and PSU ICCs for all estimated models across all country-years.

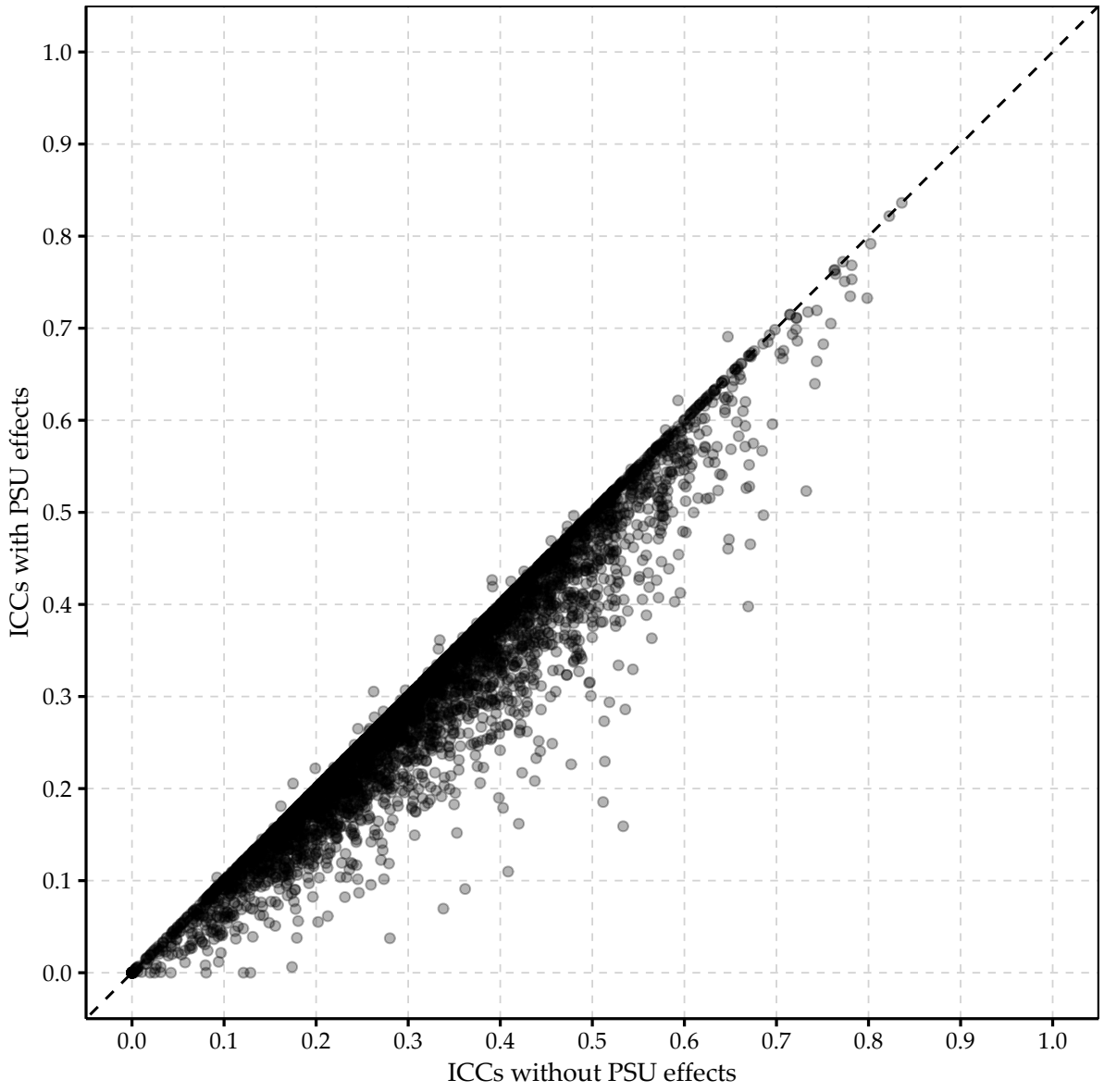

Figure H2: ICCs in models with and without PSU random effects for all estimated models across all country-years.

I Boxplot of ICCs

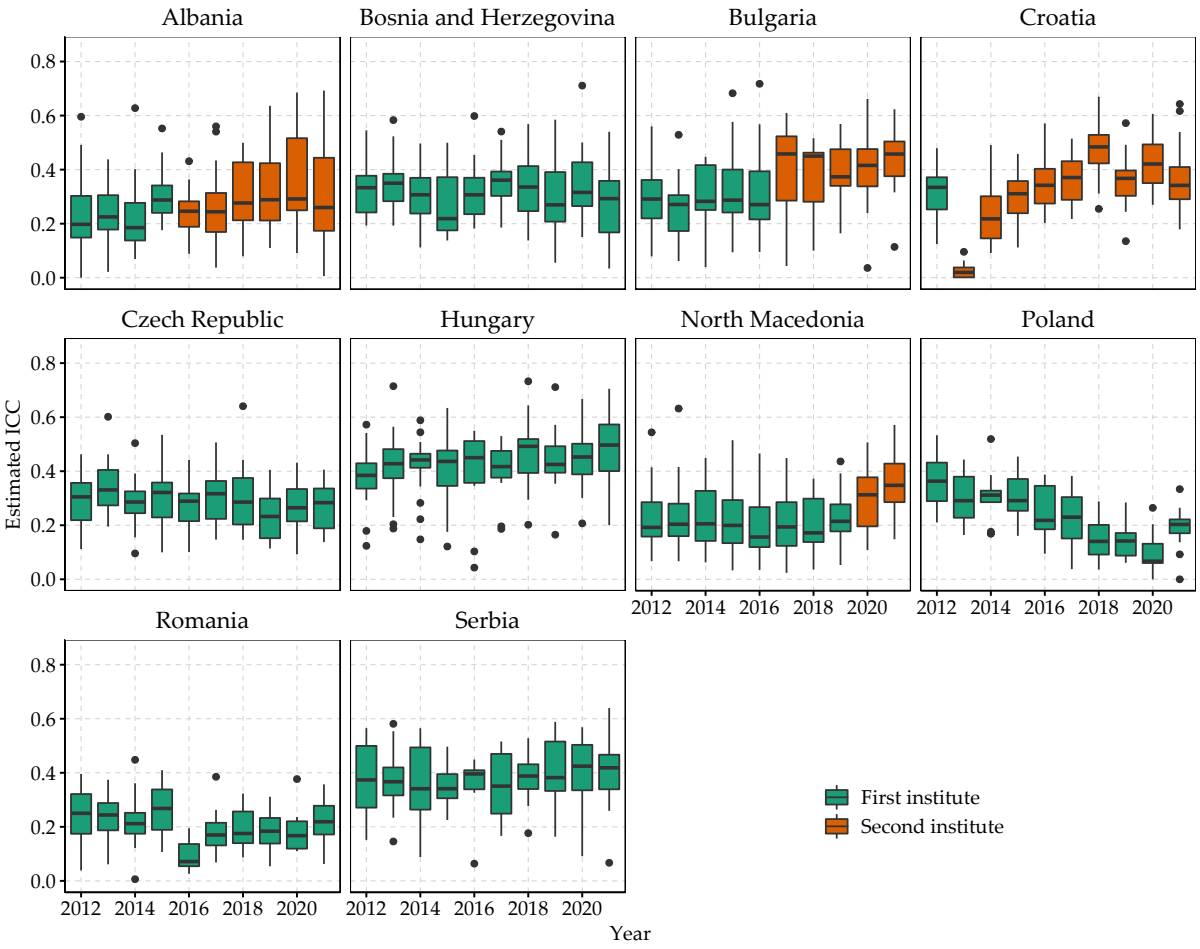

Figure I1: Boxplots of ICCs, restricted to variables for which ICCs were estimated in all years in the respective country.

J Distribution of isolation forest outlier scores

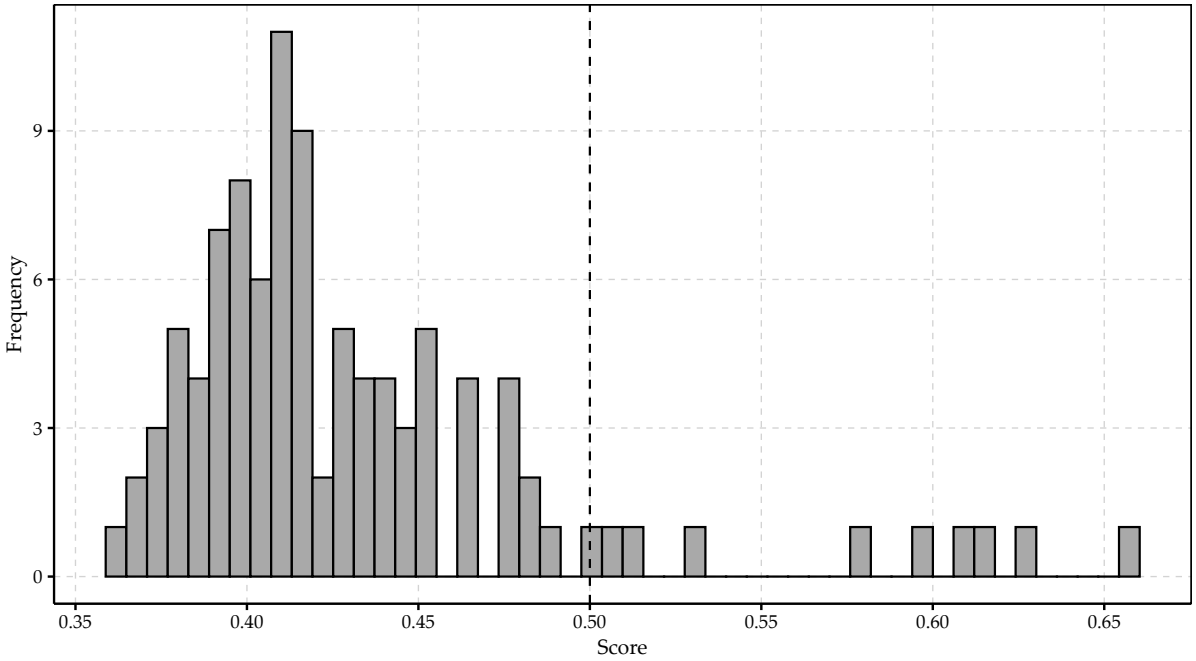

Figure J1: Distribution of isolation forest outlier scores, country-year analysis.

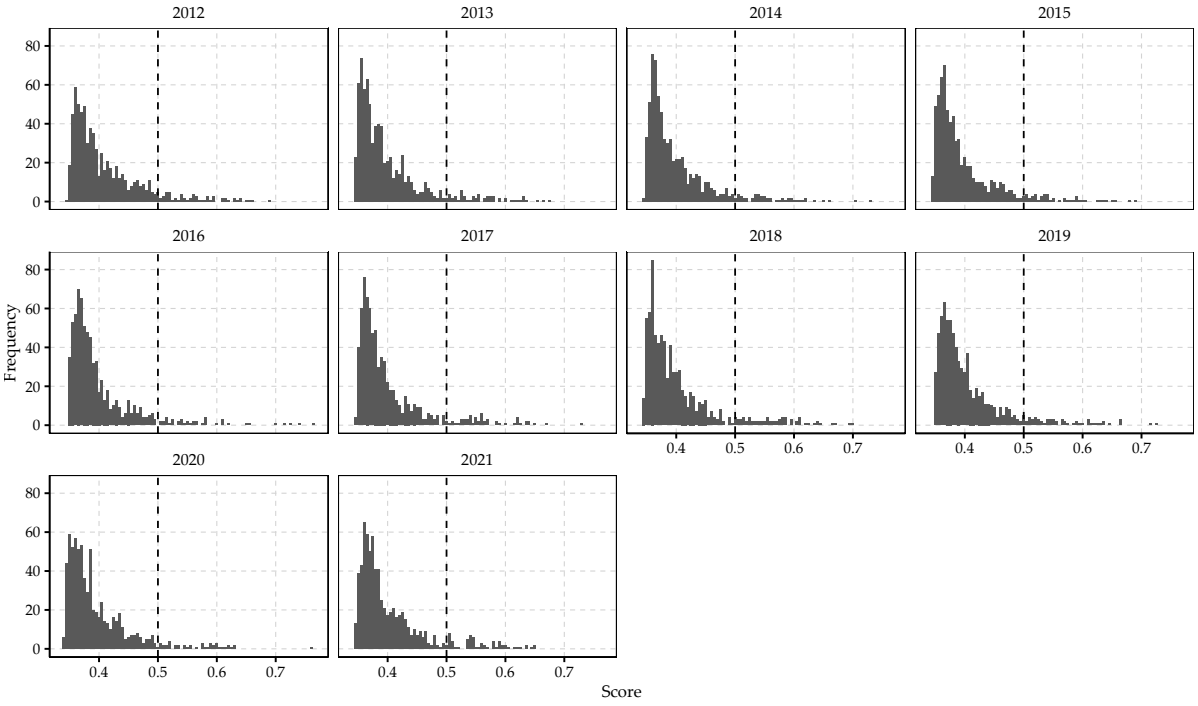

Figure J2: Distribution of isolation forest outlier scores, interviewer-level analysis.

Index of Working Papers:

January 13, 2021	Maximilian Böck, Martin Feldkircher, Burkhard Raunig	233	A View from Outside: Sovereign CDS Volatility as an Indicator of Economic Uncertainty
May 20, 2021	Burkhard Raunig	234	Economic Policy Uncertainty and Stock Market Volatility: A Causality Check
July 8, 2021	Thomas Breuer, Martin Summer, Branko Uroševic	235	Bank Solvency Stress Tests with Fire Sales
December 14, 2021	Michael Sigmund, Kevin Zimmermann	236	Determinants of Contingent Convertible Bond Coupon Rates of Banks: An Empirical Analysis
February 14, 2022	Elisabeth Beckmann, Christa Hainz, Sarah Reiter	237	Third-Party Loan Guarantees: Measuring Literacy and its Effect on Financial Decisions
February 16, 2022	Markus Knell, Reinhard Koman	238	Pension Entitlements and Net Wealth in Austria
May 9, 2022	Nicolás Albacete, Pirmin Fessler, Peter Lindner	239	The Wealth Distribution and Redistributive Preferences: Evidence from a Randomized Survey Experiment
June 20, 2022	Erwan Gautier, Cristina Conflitti, Riemer P. Faber, Brian Fabo, Ludmila Fadejeva, Valentin Jouvanceau, Jan-Oliver Menz, Teresa Messner, Pavlos Petroulas, Pau Roldan-Blanco, Fabio Rumler, Sergio Santoro, Elisabeth Wieland, Hélène Zimmer	240	New Facts on Consumer Price Rigidity in the Euro Area
June 29, 2022	Svetlana Abramova, Rainer Böhme, Helmut Elsinger, Helmut Stix	241	What can CBDC designers learn from asking potential users? Results from a survey of Austrian residents

July 1, 2022	Marcel Barmeier	242	The new normal: bank lending and negative interest rates in Austria
July 14, 2022	Pavel Ciaian, Andrej Cupak, Pirmin Fessler, d'Artis Kancs	243	Environmental-Social-Governance Preferences and Investments in Crypto-Assets
October 18, 2022	Burkhard Raunig, Michael Sigmund	244	The ECB Single Supervisory Mechanism: Effects on Bank Performance and Capital Requirements
April 5, 2023	Norbert Ernst, Michael Sigmund	245	Are zombie firms really contagious?
May 8, 2023	Richard Sellner, Nico Pintar, Norbert Ernst	246	Resource Misallocation and TFP Gap Development in Austria
September 5, 2023	Katharina Allinger, Fabio Rumler	247	Inflation Expectations in CESEE: The Role of Sentiment and Experiences
October 16, 2023	Pietro Saggese, Esther Segalla, Michael Sigmund, Burkhard Raunig, Felix Zangerl, Bernhard Haslhofer	248	Assessing the Solvency of Virtual Asset Service Providers: Are Current Standards Sufficient?
October 20, 2023	Pirmin Fessler, Severin Rapp	249	The subjective wealth distribution: How it arises and why it matters to inform policy?
October 27, 2023	Burkhard Raunig, Michael Sigmund	250	Watching over 21,000 Billion Euros: Does the ECB Single Supervisory Mechanism Affect Bank Competition in the Euro Area?
December 5, 2023	Markus Knell	251	Housing and the secular decline in real interest rates
December 14, 2023	Niko Hauzenberger, Florian Huber, Thomas O. Zörner	252	Hawks vs. Doves: ECB's Monetary Policy in Light of the Fed's Policy Stance
February 28, 2024	Lukas Olbrich, Elisabeth Beckmann, Joseph W. Sakshaug	253	Multivariate assessment of interviewer-related errors in a cross-national economic survey