mda

methods, data, analyses

Journal for Quantitative Methods and Survey Methodology

Volume 18, 2024 | 2

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.

Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany

Phone: + 49.621.1246526
E-mail: mda@gesis.org
https://mda.gesis.org

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession's knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.

All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. mda appears in two regular issues per year (January and July).

GESIS Leibniz Institute
for the Social Sciences

Layout: Bettina Zacharias (GESIS)
ISSN 2190-4936 (Online)
© of the compilation GESIS, Mannheim, July 2024

Content

RESEARCH REPORTS

RESEARCH NOTE

Combining Information from Multiple Data Sources to Improve Sampling Efficiency

Paul Burton, Sunghee Lee,
Trivellore Raghunathan & Brady T. West

University of Michigan, Survey Research Center (SRC)

Abstract

Many surveys target population subgroups that may not be readily identified in sampling frames. In the case study that motivated this study, the target population was households with children between the ages of 3 and 10 from two areas surrounding Cleveland, Ohio and Dallas, Texas. A standard approach is to sample households from these two areas and then screen for the presence of age-eligible children. Based on the estimated number of age-eligible households in these two areas, this approach would have required completing screening interviews with 5.4 to 5.7 households to find one eligible household. We developed a model-assisted sample design strategy to improve screening efficiency by attaching a measure of eligibility propensity to each household in the population. For this, we used a modeling and imputation strategy that combined information from several data sources: (1) the population of addresses for these two areas with demographic covariates from a commercial vendor, (2) external population data (from the American Community Survey and Census Planning Data) for these two areas, and (3) screening data from a large nationally representative survey. We first tested this sampling strategy in a pilot study and then implemented it in the main study. This strategy required 4.2 to 4.3 completed screeners to identify one eligible household. The proposed approach therefore improved the sampling efficiency by about 25% relative to the standard approach.

Keywords: address-based sampling, imputation, rare populations, commercial data, census data, address frame

The Housing and Children Study (H&C) is an evaluation study of housing voucher programs provided by the United States Department of Housing and Urban Development (HUD), specifically about their effect on the environment and experiences of children ages 3 to 10 years old in Dallas, TX and Cleveland, OH. These voucher programs assist low-income families and are operationalized in municipalities (e.g., Dallas) by local HUD branches known as Public Housing Authorities (PHA). H&C was designed to be an in-person survey with two sample components: one with individuals that have applied for housing vouchers ("the voucher sample") and the other with members of the general population ("the population sample"). The voucher sample was drawn from a well-specified sampling frame by PHAs. This paper focuses exclusively on the methodologies used for designing the population sample. Appendix 1 includes a list of acronyms used in this paper along with their definitions.

The population sample was designed as an area-probability sample from the two areas and used income level as a stratification factor. The main goal was to develop strategies to increase the sampling efficiency by reducing the number of households to be screened to identify eligible families and, thereby, reducing the field cost. Eligible families had at least one child ages 3 to 10 years old. This age eligibility rate was estimated nationally at 18.4% based on the American Community Survey (ACS) 2010-2014 5-year public use microdata sample (PUMS) and 17.5% based on the National Survey of Family Growth (NSFG) Cycle 8. Based on these rates, we would be required to screen roughly 5.4 to 5.7 households to identify one age-eligible household.

Due to declining response rates and increasing costs for population-based surveys, survey researchers have started examining the utility of auxiliary data to mitigate such difficulties (Smith, 2011). Commercial databases, typically purchased from sample vendors, are an example of this type of auxiliary data. Developed for commercial purposes, these databases provide a rich set of information at the individual address level, which may allow survey researchers to consider these databases as a means for improving sampling efficiency and nonresponse bias adjustment (e.g., Buskirk et al., 2014; English et al., 2019; Harter et al., 2016; Pasek et al., 2014; West, 2013; West, Wagner, Hubbard, & Gu, 2015).

Direct correspondence to
Sunghee Lee, University of Michigan, Survey Research Center (SRC),
426 Thompson St. Ann Arbor, MI 48104, USA
E-mail: sungheel@umich.edu

Sampling Rare Population Subgroups Using Commercial Databases

When surveys target specific population subgroups that are rare or small in number, a non-trivial amount of resources is required for screening eligible cases. Under this type of sampling scenario, if commercial databases include information relevant to the characteristics of target subgroups, it can be appended to the sampling frame and used for stratification (Kalton, Kali, & Sigman, 2014; Valliant, Hubbard, Lee, & Chang, 2014). A wide range of information is available from commercial databases, from socio-demographics to product purchase behaviors, donations, and voting records, and the amount of information varies by vendors (see Tables A1 and A2 in West et al., 2015). Valliant et al. (2014) also demonstrated a stratified sampling approach for the Health and Retirement Study (HRS), which is a longitudinal survey that targets a specific age cohort every six years using area-probability sampling. In 2016, HRS targeted households whose oldest member was born between 1960 and 1965 with an additional goal of oversampling ethnic and racial minorities. The sampling combined stratification at two levels: (1) stratification of geographic segments using aggregate level information from the decennial Census and ACS; and (2) stratification of addresses based on age and race/ethnicity information about people at the address obtained from commercial databases. With a disproportionate allocation, their design achieved cost savings under a variety of constraints. Similar gains in sampling efficiency were also demonstrated for a telephone survey, the National Immunization Survey (Barron et al., 2015)we assume that information is available at the sampling stage to stratify the general-population sampling frame into high-and low-density strata. Under a fixed constraint on the variance of the estimator of the domain mean, we make the optimum allocation of sample size to the several strata and show that, in comparison to proportional allocation, the optimum allocation requires (a, where landline telephone numbers were stratified by matched commercial data, enabling the targeting of households with a minor member.

Practical Limitations in Using Commercial Databases for Sampling

There are three issues with utilizing commercial databases for sampling rare population subgroups. First, not all sample addresses (or telephone numbers) may be matched to commercial data (Valliant et al., 2014), with matching rates potentially varying by vendors (West et al., 2015). Second, for the addresses successfully matched with commercial data, variables in the commercial data vary in terms of their missing rates, and this also varies by vendor (West et al., 2015). The third problem is the quality of the information in the commercial databases. The agreement rates between self-reported survey data and commercial data examined. For example, in a study that matched the 2010 U.S. decennial Census with commercial data, Rastogi and O'Hara (2012, Tables 23 and 24) showed varying agreement rates not only by vendors but also by characteristics. For example, on race/ethnicity, the agreement rates between the Census and commercial data was higher for Whites than for minority groups. The rate was around or above 95% for Whites but was around or below 10% for American Indian or Alaska Natives. Moreover, there are no standardized racial/ethnic categories across the commercial data vendors.

In sum, the third issue above deals with data accuracy, and the first two with data availability or completeness. Information incompleteness is directly a missing data issue, which has been discussed as a major limitation of using commercial data for sampling (Kalton et al., 2014; Roth, Han, & Montaquila, 2013), although a recent study reports some improvement (Roth et al. 2018). In addition to the varying missing rates across variables within a database, the missingness in the commercial databases itself appears not necessarily at random. For example, home ownership in the commercial databases is less likely to be missing among home owners than non-owners (Pasek et al., 2014, Table 3).

Imputing Missing Data in Commercial Databases

To maximize potential benefits of the existing commercial data while mitigating the practical limitations of missing data and poor accuracy, this study proposes a new method of using commercial data for sampling rare population subgroups by imputing missing data and using eligibility prediction models. These methods are then demonstrated via application to a case study. In the next section, we present the sampling design used for H&C, the imputation approaches applied to the commercial databases, and the sample design using predicted eligibility assisted by the imputed commercial data at the address level as well as external data aggregated at the geographic segment level. We then examine the accuracy and efficiency of the proposed method as observed in real fieldwork.

To meet the goal of improving screening efficiency on H&C, we used three data sources: (1) the population of addresses enhanced with commercial data for the sampled areas purchased from a vendor, (2) external population data (from ACS and Census Planning Data) for these two areas, and (3) screening data from a large nationally representative survey that includes information relevant to the eligibility in H&C. Using information from these three sources, we developed a two-stage sample design. The first stage involved sampling Census block groups (BGs), and the second stage then sampled addresses within the selected BGs using enhanced address lists. In both stages, we modelled and predicted eligibility using external data. For the first stage, we developed a model to estimate the number of households with at least one child between the ages of 3 and

10 years for each BG and used this as the measure of size (MoS) in the selection of the BGs. For the second stage, we predicted the probability for having an age eligible child for each address in the selected BGs and used this predicted eligibility as the MoS.

We first implemented this design in a pilot study before refining the strategy for the main study. The next two sections describe the H&C pilot and main study. Within each section, sampling methods and results are presented.

Pilot Study

Sampling Frame

The pilot study was conducted in Dallas, TX, using a sampling frame that included a total of 998 BGs, covering 70.5% of the ZIP codes where potential voucher applicants resided.

Sample Design

The sample design leveraged multiple external data sources: (1) the ACS 2010-2014 5-year summary file (SF); (2) the 2016 Census planning data; (3) a commercial database purchased from Marketing Systems Group (MSG: http://www.m-s-g.com/); and (4) household roster data from the 2011-2015 National Survey of Family Growth (NSFG), an area-probability national sample survey conducted by the Centers for Disease Control and Prevention. It should be noted that NSFG and MSG data are available at the address/household level, while ACS SF and Census planning data are aggregated at various levels of geography as fine as BGs. The availability of NSFG roster data was crucial for the H&C design, because it provided precise data on H&C age eligibility used in both stages of sampling.

Two-stage sampling as illustrated in Table 1 was used to select the sample. In the primary stage, BGs were sampled using a stratified probability proportionate to estimated size (PPeS) design. In the secondary stage, addresses/households were selected from sampled BGs also using a stratified PPeS design. The detail for each stage is described below.

Table 1 Description of Overall Sample Design, Housing and Children Study

	Primary Stage	Secondary Stage
Sampling unit	Census block groups (BG)	Addresses/Households
Measure of Size	Number of households with at least one child aged between 3 and 10 years old	Probability of having at least one child aged between 3 and 10 years old
Estimation Method	Model-based prediction	Model-based prediction
Prediction Model	Grouped logit model with - DV: Household-level age eligibility indicator from the NSFG roster data aggregated to the BG level - IVs: BG-level auxiliary data (ACS SF and Census planning data with the dimensions reduced through principal components analysis)	Individual logit model with - DV: Household-level age eligibility indicator from the NSFG roster data - IVs: Address-level commercial data (with missing data treated through sequential multiple imputation) + BG-level auxiliary data (ACS SF data with the dimensions reduced through principal components analysis)
Prediction	Multiply the proportion of eligible households for each BG in the H&C frame, predicted by fitting the grouped logit model, with the number of households for each BG	Predict the probability of being age eligible for each address in BGs sampled from the primary stage by fitting the individual logit model
Stratification Variable and Method	Proportion of households with an annual income less than $35,000 directly available from ACS SF	Household income from commercial data - If not missing, exact income values from commercial data - If missing, imputed income from sequential multiple imputation

Note. DV: Dependent variable; IV: Independent variable; H&C: Housing and Children Study; NSFG: National Survey of Family Growth; ACS SF: American Community Survey Summary File

Primary Stage Design

The primary stage focused on selecting 15 BGs from 998 BGs on the H&C pilot frame using the BG-level number of households with at least one child aged between 3 and 10 years old as the MoS. Note that this MoS is not readily available from any of the external data. We estimated the MoS as follows using NSFG and ACS SF data at the BG level. First, we created a dataset by aggregating the household-level H&C age eligibility in the NSFG roster data to the BG level and appending 160 variables from ACS SF relevant for this age eligibility (see Supplementary Table 1 at https://goo.gl/co4SuZ). Second, for the goal of estimating the proportion of H&C eligible households at the BG level, we fitted a grouped logit model of aggregated eligibility by regressing the aggregated BG-level eligibility rates from NSFG on ACS SF variables. Instead of selecting individual variables from ACS for this model, we used principal component analysis (PCA) to reduce the dimensionality from 160 ACS variables while retaining a similar amount of information. With the PCA suggesting 63 components that explained 95% of the variance in the original 160 variables, we modelled the aggregated BG-level eligibility from NSFG on these 63 components as well as 155 two-way interactions identified from a stepwise variable selection process. This model included a total of 1,909 BGs in NSFG and provided fair fit with an area under the ROC curve of 0.66 and a non-significant Hosmer–Lemeshow goodness-of-fit test (χ^2=7.90, *df*=8; *p*=.443).

The estimated model was applied to the 988 BGs on the H&C pilot frame, from which the BG-level proportion of H&C eligible households was predicted. With the counts of total households available from ACS SF, the predicted proportions were multiplied by the household counts, yielding the MoS at the BG level. The minimum MoS was set at 10 eligible households. BGs smaller than the minimum MoS were combined within income strata as described shortly.

BGs on the frame were stratified using the proportion of “low income” households from ACS SF defined as those with annual income less than $35,000. Specifically, we used the tertiles of this distribution as cutoff points, designating BGs into three strata: low (>37.4%; i.e., more than 37.4% of the households in BG with income less than $35,000), middle (19.3-37.4%) and high income (<19.3%). With the overall project goal being to select BGs at the ratio of 3:2:1 from low-, middle- and high-income strata, the pilot study selected 8, 5, and 2 low-, middle- and high-income BGs with PPeS within strata.

Secondary Stage Design

The secondary stage dealt with selecting addresses from the 15 sampled BGs using the predicted probability of a given address being H&C eligible as the MoS, which allowed us to improve our ability to target likely eligible households. With this information not directly available, we leveraged four external datasets through a prediction model, similar to the primary stage design. First, we concatenated all 61,085 addresses in the NSFG roster data with their H&C eligibility indicator and all 10,304 addresses in the 15 BGs sampled for the pilot study from the USPS delivery sequence file. For H&C addresses, the eligibility indicator was naturally missing. To these data, we merged address-level MSG data (15 variables in Table 2) and BG-level ACS SF and Census planning data (483 variables in Supplementary Table 2 at https://goo.gl/ERGWvy). The idea was to model the household-level eligibility as a function of the MSG variables and ACS/Census variables. This required treatments of the missingness in the MSG data and the large dimensionality of the ACS/Census data.

The large dimensionality was handled with PCA, similar to the procedure used for the primary stage. A total of 483 ACS/Census variables was reduced to 113 components that retained 95% of the total variance. The missing rates of MSG variables considered in the pilot study were as low as 17.6% and as high as 83.9% as reported in Table 2. To mitigate this issue, we applied sequential imputation using multivariable regression models through the software package IVEware (Raghunathan, Berglund, & Solenberger, 2018). For numeric variables, ordinary least squares regression models were used; for binary variables, logit models; and for categorical variables, multinomial logit models. The baseline imputation model included the 113 components from the PCA as predictors. We used multiple imputation in order to assess model fit and ascertain uncertainty associated with the random error in the imputation models, which single imputation does not allow. Repeating the imputation 10 times offers sufficient information about this uncertainty (Raghunathan et al., 2018). Because imputed values for the missing cases varied only minimally across imputations, we used the average of 10 imputed values.

Logistic regression was used to model the eligibility of 61,085 addresses in the NSFG roster data with the ACS/Census principal components and imputed commercial data. Across 10 imputations, the model fit was comparable with an area under the ROC curve ranging around 0.71-0.72. The estimated model was applied to the 10,304 H&C addresses in order to predict their probability of being age eligible. The predicted eligibility was around 24-25%, and this result was similar across the 10 imputations as shown in Table 3. The average of the 10 predicted eligibility probabilities was used as the MoS.

H&C addresses were stratified by the income variable in the MSG data whose missingness was treated with imputation as described above. The income strata were formed based on the tertiles of this income distribution. Addresses with income <$30,000 were assigned to the low-income stratum, $30,000-62,500 to the middle-income stratum and >$62,500 to the high-income stratum. Considering the target ratio of 3:2:1 for these income strata, 684 addresses were selected using PPeS with predicted eligibility as the MoS within income stratum for the screening interviews conducted from October to December 2016.

Table 2 Missing Rates of Variables in MSG Data Used for Address-Level Eligibility Prediction, Housing and Children Study

	Missing Rate	
Variable Description	Pilot Study *(n=71,389)**	Main Study *(n=135,716)**
Age of Person 1 in household	48.4%	47.1%
Education of Person 1 in household	28.6%	26.7%
Ethnicity of Person 1 in household	28.6%	26.7%
Gender of Person 1 in household	17.6%	15.1%
Total household Income	17.6%	15.1%
Marital Status of Person 1 in Household	26.4%	27.2%
Flag for Asian Surname of Person 1 in Household	17.6%	15.1%
Flag for Hispanic Surname of Person 1 in Household	17.6%	15.1%
Flag for Name provided for Person 1 in Household	17.6%	15.1%
Number of Adults (18 years and older) in Household	83.9%	85.3%
Number of Children (Under Age 18) in Household	21.8%	24.6%
Does Householder Rent or Own the Household	21.8%	24.6%
Age of Person 2 in Household	75.0%	74.8%
Flag for Phone Number provided of Household	70.7%	83.8%
Flag for Presence of Any Person Age 18 to 24 in Household	17.6%	15.1%
Flag for Presence of Any Person Age 25 to 34 in Household	82.6%	15.1%
Flag for Presence of Any Person Age 35 to 64 in Household	82.6%	15.1%
Flag for Presence of Any Person Age ≥65 in Household	17.6%	15.1%

* Sample sizes indicate counts of addresses in the block groups sampled for the Housing and Children Study and addresses in the National Survey of Family Growth roster data considered in the address-level eligibility prediction model.

Table 3 Distribution of Predicted Address-Level Eligibility Probability from Each Imputation for Addresses in Sampled Block Groups, Housing and Children Study

	Pilot Study					Main Study									
						Dallas					Cleveland				
Imputation	N	Mean	SD	Min	Max	n	Mean	SD	Min	Max	n	Mean	SD	Min	Max
1	10,304	0.246	0.148	0.013	0.774	41,536	0.417	0.268	0.004	0.987	26,000	0.259	0.209	0.004	0.984
2	10,304	0.242	0.144	0.013	0.736	41,536	0.423	0.272	0.004	0.992	26,000	0.258	0.204	0.004	0.985
3	10,303	0.246	0.147	0.013	1.000	41,526	0.420	0.271	0.003	0.985	25,999	0.258	0.207	0.003	0.984
4	10,304	0.244	0.143	0.008	0.757	41,534	0.420	0.270	0.002	0.985	25,501	0.253	0.204	0.004	0.987
5	10,304	0.242	0.143	0.011	0.727	41,536	0.424	0.271	0.004	0.986	25,999	0.261	0.208	0.005	0.988
6	10,303	0.247	0.143	0.009	1.000	41,536	0.419	0.267	0.002	0.987	26,000	0.257	0.206	0.004	0.987
7	10,304	0.247	0.143	0.000	0.760	41,536	0.419	0.269	0.002	0.987	26,000	0.260	0.211	0.003	0.985
8	10,304	0.248	0.144	0.013	0.771	41,536	0.419	0.271	0.003	0.987	25,999	0.255	0.205	0.003	0.987
9	10,303	0.246	0.144	0.000	0.766	41,536	0.415	0.268	0.003	0.984	26,000	0.258	0.207	0.004	0.986
10	10,303	0.247	0.147	0.012	0.766	41,536	0.419	0.270	0.004	0.987	26,000	0.260	0.209	0.000	0.988
Average	*10,304*	*0.245*	*0.145*	*0.009*	*0.806*	*41,536*	*0.420*	*0.265*	*0.004*	*0.983*	26,000	0.258	0.202	0.004	0.986

Note. The sample size may differ across imputation. This occurred when the imputation produced fewer categories of the MSG ethnicity variable for the National Survey of Family Growth addresses than for the Housing and Children Study addresses.

Results

Accuracy

Table 4.A provides results of screening interviews by BG and income strata along with the observed and predicted eligibility of sample addresses. The comparison between predicted and observed eligibility provides information about the accuracy of our predictions. Overall, out of 684 sample addresses, 284 completed the screener; and among them, 78 were eligible for H&C. This resulted in a 27.5% eligibility rate, which is 10 percentage points higher than the national eligibility rates of 17-18% estimated from NSFG and ACS. The predicted eligibility rate of 25.8% mapped onto the eligibility observed in the field, 27.5%. When examining the eligibility by BG, there was a substantial variation in its prediction accuracy across BGs measured by the difference between observed and predicted eligibility rates. Although the small number of BGs considered in the pilot study limited a thorough investigation, BGs in the high-income stratum appeared to be subject to a lower level of variability in prediction accuracy than BGs in the lower-income stratum.

Table 4 Block Group Level Screener Results by Income Strata, Housing and Children Study

A. Pilot Study

Block Group		Counts of Addresses			Eligibility		
Number	Income Strata	Sampled	Interviewed	Eligible	Observed	Predicted*	Pred−Obs
1	Low	53	27	9	33.3%	31.2%	-2.1%
2	Low	41	12	3	25.0%	16.4%	-8.6%
3	Low	42	19	9	47.4%	28.9%	-18.5%
4	Low	60	26	9	34.6%	29.8%	-4.8%
5	Low	56	25	7	28.0%	34.5%	6.5%
6	Low	51	18	6	33.3%	29.5%	-3.8%
7	Low	52	22	2	9.1%	22.1%	13.0%
8	Low	33	19	5	26.3%	43.4%	17.1%
Subtotal: Low-Income		*388*	*168*	*50*	*29.8%*	*29.3%*	*-0.5%*
9	Middle	42	23	7	30.4%	32.6%	2.2%
10	Middle	32	11	4	36.4%	29.0%	-7.3%
11	Middle	42	17	6	35.3%	26.1%	-9.2%
12	Middle	39	13	4	30.8%	27.9%	-2.8%
13	Middle	45	14	1	7.1%	12.3%	5.1%
Subtotal: Middle-Income		*200*	*78*	*22*	*28.2%*	*25.2%*	*-3.0%*
14	High	64	23	3	13.0%	9.2%	-3.8%
15	High	32	15	3	20.0%	20.6%	0.6%
Subtotal: High-Income		*96*	*38*	*6*	*15.8%*	*13.0%*	*-2.8%*
Grand Total		*684*	*284*	*78*	*27.5%*	*25.8%*	*-1.7%*

*Average eligibility predicted for sample addresses in a given block group in the secondary sampling stage.

B. Main Study -- Dallas

Block Group		Counts of Addresses			Eligibility		
Number	Income Strata	Sample	Interviewed	Eligible	Observed	Predicted*	Pred−Obs
Quarter 1							
1	Low	111	63	18	28.6%	42.9%	14.3%
2	Low	110	50	7	14.0%	9.9%	-4.1%
3	Low	103	45	9	20.0%	20.7%	0.7%
4	Low	81	31	10	32.3%	22.8%	-9.5%
5	Low	166	80	27	33.8%	30.9%	-2.9%
6	Low	78	47	16	34.0%	23.6%	-10.5%
7	Middle	81	36	4	11.1%	28.5%	17.4%
8	Middle	94	35	9	25.7%	68.3%	42.6%
9	Middle	87	49	9	18.4%	22.4%	4.0%
10	Middle	97	22	7	31.8%	92.1%	60.3%
11	High	90	47	3	6.4%	61.7%	55.3%
12	High	98	45	14	31.1%	39.5%	8.4%
Subtotal: Quarter 1		*1,196*	*550*	*133*	*24.2%*	*38.3%*	*14.1%*
Quarter 2							
1	Low	122	43	6	14.0%	25.8%	11.8%
2	Low	161	72	13	18.1%	42.6%	24.6%
3	Low	135	73	24	32.9%	21.0%	-11.9%
4	Low	140	51	15	29.4%	67.8%	38.4%
5	Low	70	22	12	54.5%	29.8%	-24.8%
6	Low	132	58	15	25.9%	12.5%	-13.3%
7	Middle	105	14	0	0.0%	80.2%	80.2%
8	Middle	156	66	17	25.8%	10.4%	-15.4%
9	Middle	106	79	14	17.7%	30.8%	13.1%
10	Middle	122	55	13	23.6%	11.4%	-12.2%
11	High	99	45	5	11.1%	11.4%	0.3%
12	High	99	45	5	11.1%	26.6%	15.5%
Subtotal: Quarter 2		*1,447*	*623*	*139*	*22.3%*	*30.8%*	*8.5%*
Quarter 3							
1	Low	87	54	15	27.8%	36.0%	8.2%
2	Low	93	29	6	20.7%	31.3%	10.6%
3	Low	99	64	32	50.0%	82.0%	32.0%
4	Low	93	49	4	8.2%	24.6%	16.4%
5	Low	95	58	20	34.5%	57.7%	23.2%
6	Low	102	53	12	22.6%	27.2%	4.5%
7	Middle	168	115	37	32.2%	35.2%	3.1%
8	Middle	100	40	11	27.5%	40.8%	13.3%
9	Middle	96	68	19	27.9%	52.7%	24.8%
10	Middle	111	50	7	14.0%	79.8%	65.8%

Block Group		Counts of Addresses			Eligibility		
Number	Income Strata	Sample	Interviewed	Eligible	Observed	Predicted*	Pred−Obs
11	High	93	55	13	23.6%	16.5%	-7.2%
12	High	82	24	2	8.3%	6.5%	-1.8%
Subtotal: Quarter 3		*1,219*	*659*	*178*	*27.0%*	*41.6%*	*14.6%*
Quarter 4							
1	Low	143	69	7	10.1%	24.7%	14.6%
2	Low	243	145	31	21.4%	30.4%	9.1%
3	Low	136	77	16	20.8%	16.9%	-3.9%
4	Low	115	52	16	30.8%	21.1%	-9.7%
5	Low	122	60	16	26.7%	81.8%	55.1%
6	Low	151	76	15	19.7%	32.1%	12.3%
7	Middle	87	58	7	12.1%	11.9%	-0.2%
8	Middle	87	41	2	4.9%	13.0%	8.1%
9	Middle	116	64	19	29.7%	38.7%	9.0%
10	Middle	130	61	17	27.9%	49.0%	21.1%
11	High	124	66	15	22.7%	28.8%	6.1%
12	High	92	53	9	17.0%	50.6%	33.6%
Subtotal: Quarter 4		*1,546*	*822*	*170*	*20.7%*	*33.4%*	*12.8%*
Reserve							
1	Low	65	32	8	25.0%	38.3%	13.3%
2	Low	81	33	17	51.5%	81.5%	29.9%
3	Low	59	38	9	23.7%	75.5%	51.9%
4	Middle	104	37	8	21.6%	35.4%	13.7%
5	Middle	70	29	11	37.9%	28.1%	-9.9%
6	High	57	19	4	21.1%	47.0%	25.9%
Subtotal: Reserve		*436*	*188*	*57*	*30.3%*	*50.2%*	*19.8%*
Subtotal: Low-Income		*3,093*	*1,524*	*396*	*26.0%*	*36.1%*	*10.1%*
Subtotal: Middle-Income		*1,917*	*919*	*211*	*23.0%*	*40.1%*	*17.1%*
Subtotal: High-Income		*834*	*399*	*70*	*17.5%*	*31.4%*	*13.8%*
Grand Total		*5,844*	*2,842*	*677*	*23.8%*	*36.7%*	*12.9%*

*Average eligibility predicted for sample addresses in a given block group in the secondary sampling stage.

C. Main Study -- Cleveland

Block Group		Counts of Addresses			Eligibility		
Number	Income Strata	Sample	Interviewed	Eligible	Observed	Predicted*	Pred−Obs
Quarter 1							
1	Low	171	79	27	34.2%	23.8%	-10.3%
2	Low	151	56	11	19.6%	19.7%	0.1%
3	Low	186	48	8	16.7%	14.3%	-2.3%
4	Low	179	66	38	57.6%	61.4%	3.8%
5	Low	226	73	19	26.0%	13.0%	-13.1%
6	Low	160	28	5	17.9%	16.5%	-1.4%
7	Middle	171	74	16	21.6%	17.4%	-4.2%
8	Middle	168	33	8	24.2%	15.3%	-8.9%
9	Middle	168	50	4	8.0%	14.5%	6.5%
10	Middle	171	72	20	27.8%	16.0%	-11.8%
11	High	166	91	14	15.4%	15.2%	-0.2%
12	High	169	90	9	10.0%	13.3%	3.3%
Subtotal: Quarter 1		*2,086*	*760*	*179*	*23.6%*	*20.0%*	*-3.5%*
Quarter 2							
1	Low	180	85	23	27.1%	22.6%	-4.5%
2	Low	130	35	14	40.0%	61.6%	21.6%
3	Low	165	42	9	21.4%	13.9%	-7.6%
4	Low	100	32	12	37.5%	19.1%	-18.4%
5	Low	142	12	0	0.0%	20.3%	20.3%
6	Low	148	29	0	0.0%	38.3%	38.3%
7	Middle	150	50	12	24.0%	20.2%	-3.8%
8	Middle	137	58	13	22.4%	17.7%	-4.8%
9	Middle	152	41	8	19.5%	21.4%	1.9%
10	Middle	132	34	9	26.5%	17.4%	-9.1%
11	High	142	56	10	17.9%	13.4%	-4.5%
12	High	132	10	1	10.0%	19.1%	9.1%
Subtotal: Quarter 2		*1,710*	*484*	*111*	*22.9%*	*23.5%*	*0.6%*
Quarter 3							
1	Low	147	46	17	37.0%	20.8%	-16.2%
2	Low	135	17	4	23.5%	52.7%	29.2%
3	Low	147	47	12	25.5%	21.7%	-3.9%
4	Low	243	74	14	18.9%	27.0%	8.1%
5	Low	140	67	32	47.8%	36.4%	-11.4%
6	Low	135	23	2	8.7%	12.2%	3.5%
7	Middle	135	43	8	18.6%	17.3%	-1.3%
8	Middle	166	41	10	24.4%	15.8%	-8.6%
9	Middle	212	35	6	17.1%	11.6%	-5.6%
10	Middle	143	28	13	46.4%	28.4%	-18.1%

Block Group		Counts of Addresses			Eligibility		
Number	Income Strata	Sample	Interviewed	Eligible	Observed	Predicted*	Pred−Obs
11	High	135	41	5	12.2%	16.9%	4.7%
12	High	132	58	9	15.5%	10.8%	-4.7%
Subtotal: Quarter 3		*1,870*	*520*	*132*	*25.4%*	*22.4%*	*-3.0%*
Quarter 4							
1	Low	141	43	6	14.0%	16.4%	2.4%
2	Low	156	58	11	19.0%	18.1%	-0.9%
3	Low	132	62	30	48.4%	57.5%	9.1%
4	Low	104	52	30	57.7%	91.4%	33.7%
5	Low	131	54	13	24.1%	15.0%	-9.1%
6	Low	184	66	17	25.8%	16.2%	-9.6%
7	Middle	120	51	7	13.7%	20.0%	6.2%
8	Middle	172	51	7	13.7%	13.9%	0.2%
9	Middle	141	30	1	3.3%	8.8%	5.5%
10	Middle	145	77	15	19.5%	15.1%	-4.4%
11	High	128	47	7	14.9%	18.6%	3.7%
12	High	130	43	5	11.6%	17.2%	5.6%
Subtotal: Quarter 4		*1,684*	*634*	*149*	*23.5%*	*23.8%*	*0.3%*
Reserve							
1	Low	167	64	11	17.2%	19.0%	1.8%
2	Low	150	47	9	19.1%	10.1%	-9.0%
3	Low	147	23	6	26.1%	7.3%	-18.8%
4	Middle	167	105	21	20.0%	49.3%	29.3%
5	Middle	131	35	3	8.6%	13.5%	5.0%
6	High	141	30	4	13.3%	46.6%	33.2%
Subtotal: Reserve		*903*	*304*	*54*	*17.8%*	*24.7%*	*7.0%*
Subtotal: Low Income		*4,197*	*1,328*	*380*	*28.6%*	*26.4%*	*-2.2%*
Subtotal: Middle Income		*2,781*	*908*	*181*	*19.9%*	*18.5%*	*-1.4%*
Subtotal: High Income		*1,275*	*466*	*64*	*13.7%*	*18.9%*	*5.2%*
Grand Total		*8,253*	*2,702*	*625*	*23.1%*	*22.6%*	*-0.6%*

* Average eligibility predicted for sample addresses in a given block group in the secondary sampling stage.

Sampling Efficiency

Sampling efficiency was examined by comparing sample sizes under the current design and under simple random sampling (SRS) of addresses within BGs. The current design yielded 78 eligible cases from 684 addresses with a screener cooperation rate of 41.5% and an eligibility rate of 27.5%. Under SRS, the eligibility rate would be similar to the national rate (around 17.5%.) To yield 78 cases, SRS would have required 1,074 addresses (=78 / (41.5% cooperation rate x 17.5% eligibility rate)), which is an increase of almost 400 sample addresses.

Main Study

The main study targeted households with at least one child aged between 3 and 10 years old in Dallas, TX and Cleveland, OH. Given the results from the pilot study, an identical sample design was employed in the main survey with more streamlined and updated external data.

Frame

The frame included 998 BGs from the city of Dallas proper as done in the pilot study and 850 BGs from within the city of Cleveland proper, covering 70.5% and 85.3% of the ZIP codes where the voucher applicants resided in the respective locations.

Sample Design

A stratified two-stage PPeS design, identical to the pilot study, was implemented with more up-to-date auxiliary data. Specifically, the ACS 2011-2015 5-Year SF, the NSFG roster data from 2011 to 2017 and the MSG data purchased in 2017 were used in the main study. In particular, the NSFG data included 68,180 addresses from 2,007 BGs. Note that Census planning data was not considered in the main study design, due to a large overlap in its information with ACS SF. Data collection was planned for a year with the fieldwork implemented via 4 replicates. Hence, the sample drawn at the beginning of the project was released sequentially by replicate.

Primary Stage Sampling

The eligibility rate of addresses aggregated from all 2,007 BGs from NSFG was regressed on BG-level variables in ACS SF. The grouped logit model included these 84 components extracted from PCA of 236 variables in ACS SF (see Supplementary Table 3 at https://goo.gl/KtRcfD) and 188 two-way interactions of some components as predictors. This model showed an improved fit compared to the pilot study (area under the ROC curve: 0.67; Hosmer–Lemeshow goodness-of-fit test: χ^2=2.65, *df*=8, *p*=.955).

With the updated ACS data, the income stratification changed. For Dallas, BGs with >51.0% households with annual income less than $35,000 were classified as the low-income stratum; those with 27.0%-51.0% into the middle-income stratum; and those with <27.0% into the high-income stratum. For Cleveland, 62.1% and 38.7% were the respective income cut-off points. Overall, 54 BGs were selected for each site using PPeS for a 3:2:1 ratio of low-, middle- and high-income strata BGs, where 48 BGs were randomly split into 4 replicates and the remaining 6 BGs were set aside as reserve sample.

Secondary Stage Sampling

The address-level eligibility model included 68,180 addresses from the NSFG roster data (41,536 in Dallas and 26,000 in Cleveland). Address-level eligibility was modelled using address-level MSG variables as well as BG-level ACS SF data, where the missingness of the MSG data was handled through sequential multiple imputation and the dimensionality of the ACS data was reduced through PCA. The distribution of predicted eligibility across the 10 imputations is shown in Table 3. The predicted eligibility was similar across imputations and, on average, higher for Dallas (approximately 0.42) than Cleveland (approximately 0.26). The average predicted eligibility from the 10 imputations was used as the MoS.

Income-based stratification used the household income variable in MSG. Unlike the pilot study, the income tertiles calculated *within each BG* were used. This means that the stratification did not use “hard boundaries” but varied by BG. Within each BG, one third of addresses were assigned to low-, middle- and high-income strata. Considering the target ratio of 3:2:1 for these income strata as well as predicted eligibility rates of addresses, 5,844 addresses from Dallas and 8,258 addresses from Cleveland were sampled for data collection, which ran from May 2017 to September 2018.

Results

Accuracy

The results of the screener fieldwork are in Tables 4.B and 4.C. Among the 2,842 households in Dallas that completed the screener, 677 were eligible. This overall eligibility rate of 23.8% was lower by 12.9 percentage points than the predicted eligibility rate of 36.7%. Although the over-prediction of eligibility was persistent across all replicates and across income strata, the observed eligibility rate was still higher than the national average of 17-18%. For Cleveland, the overall eligibility rate was 23.1%, closely matching the predicted eligibility of 22.6% and higher than the national average eligibility. With the exception of BG 11 of Dallas in Quarter 1, the variability in accuracy was smaller for the addresses in the high-income stratum.

Sampling Efficiency and Cost Considerations

Our design yielded 677 eligible cases with a screener cooperation rate of 48.6% and an eligibility rate of 23.8% for Dallas. To yield the same number of eligible households under SRS, the design would have required screening 7,954 addresses (= 677 / (48.6% cooperation rate x 17.5% eligibility rate)), an increase of a little over 2,100 sample addresses. For Cleveland, with a yield of 625 eligible cases, a screener cooperation rate of 32.7%, and an eligibility rate of 23.1% under the current design, SRS would have required 10,909 addresses (=625 / (32.7% cooperation rate x 17.5% eligibility rate)), an increase of over 2,600 sampled addresses. Our design offered a net reduction in required sample size of 27% (5,844 under our design vs. 7,954 under SRS) for Dallas and 24% (8,253 under our design vs. 10,909 under SRS) for Cleveland.

In order to assess the cost savings through improvement in screening efficiency, we fitted a simple cost model with interviewer as the unit of analysis as follows:

$$T_i = \beta_0 + \beta_1 S_i + \beta_2 I_i + \varepsilon_i$$

Where T_i is the total billed hours by interviewer i; S_i is the number of completed screeners by interviewer i; and I_i is the number of completed interviews by interviewer i. Therefore, coefficients β_1 and β_2 are, respectively, the interviewer hours per completed screener and per completed main interview. Using the data from 60 interviewers for the main study, the estimated model (R^2 = 0,913) suggested about 1.9 hours (SE: 0.4) per completed screener and about 10.8 hours (SE: 1.2) per completed interview.

To estimate the cost savings, we consider a counter factual that assumes the same cooperation rate for the screening interview and yields the same number of eligible households (677 in Dallas and 625 in Cleveland) with the national eligibility rate of 17.5%. The standard approach would have required completing screener interviews with 3,869 households (=677/17.5%) in Dallas and 3,571 households (=625/17.5%) in Cleveland, as opposed to 2,842 and 2,702 completed screeners in the respective areas under our design given in Tables 4.B and 4.C. This equates to a 25% reduction in required screener completion. This also means that, with 1.9 interviewer hours estimated per completed screener, our design saved nearly 3,600 interviewer hours. This ignores the additional costs of sampling a larger number of households to reach the required eligible households using the standard approach.

Discussion

Our goal in this study was to improve sampling efficiency and thereby reduce the data collection costs of the H&C study. The screening for eligible members of the target population from the larger sampling population frame contributes greatly to the cost of surveys of uncommon and hard-to-reach populations. For implementing measurements about child development and parent-child interactions, H&C required a face-to-face mode.

Survey research organizations can leverage information from previous studies combined with commercial databases to develop model-assisted sampling designs that may improve sampling efficiency and reduce costs. This case study illustrates a methodology that can be used to leverage information from imperfect sources through imputation and modeling. We note that practical limitations exist for using commercial databases directly for sampling. However, when reflecting on our proposed approach that used imputation and the modeling of study eligibility, it is feasible to address the well-documented availability and accuracy issues of commercial data. It is important to note that, for studies designed to oversample addresses/areas with characteristics associated with lower availability or accuracy of the commercial data (e.g., lower income), the prediction accuracy may be lower as shown in the case of over-prediction of eligibility in Dallas (Table 4.B) than for studies without such oversampling requirements.

Efficiency can also be gained by performing model-based analysis when commercial data is available on all households in the selected geographies and the ACS data is available on all geographies used as sampling units. Alternatively, a Bayesian prediction model can be used to synthesize the entire population through simulations and then construct inferences from the simulated populations, offering a gain in inference efficiency. Whatever the method used, we believe that our case study demonstrates that these methods have great potential for leveraging commercial data to improve efficiency in sampling and inference.

References

Barron, M., Davern, M., Montgomery, R., Tao, X., Wolter, K. M., Zeng, W., … Black, C. (2015). Using Auxiliary Sample Frame Information for Optimum Sampling of Rare Populations. *Journal of Official Statistics, 31*(4), 545–557. https://doi.org/10.1515/JOS-2015-0034

Buskirk, T. D., Malarek, D., & Bareham, J. S. (2014). From Flagging a Sample to Framing It: Exploring Vendor Data That Can Be Appended to ABS Samples. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 111–124.

English, N., Kennel, T., Buskirk, T., & Harter, R. (2019). The construction, maintenance, and enhancement of address-based sampling frames. *Journal of Survey Statistics and Methodology, 7*(1), 66–92. https://doi.org/10.1093/jssam/smy003

Harter, R., Battaglia, M. P., Buskirk, T. D., Dillman, D. A., English, N., Fahimi, M., ... Zukerberg, A. L. (2016). *Address-based Sampling*. Retrieved from https://www.aapor.org/Education-Resources/Reports/Address-based-Sampling.aspx

Kalton, G., Kali, J., & Sigman, R. (2014). Handling Frame Problems When Address-Based Sampling Is Used for In-Person Household Surveys. *Journal of Survey Statistics and Methodology, 2*(3), 283–304. https://doi.org/10.1093/jssam/smu013

Pasek, J., Jang, S. M., Cobb, C. L., Dennis, J. M., & Disogra, C. (2014). Can marketing data aid survey research? Examining accuracy and completeness in consumer-file data. *Public Opinion Quarterly, 78*(4), 889–916. https://doi.org/10.1093/poq/nfu043

Raghunathan, T., Berglund, P., & Solenberger, P. (2018). *Multiple Imputation in Practice: With Examples Using IVEware*. https://doi.org/10.1198/000313001317098266

Rastogi, S., & O'Hara, A. (2012). *2010 Census Match Study Report* (No. 247). Retrieved from https://www.census.gov/content/dam/Census/library/publications/2012/dec/2010_cpex_247.pdf

Roth, S. B., Han, D., & Montaquila, J. M. (2013). The ABS Frame: Quality and Considerations. *Survey Practice, 6*(4), 1–6. https://doi.org/10.29115/SP-2013-0021

Roth, S., Caporaso, A., & DeMatteis, J. (2018). Variables Appended to ABS Frames: Has Data Quality Improved? *Paper Presented at the Annual Conference of American Association for Public Opinion Research, Denver, CO.*

Smith, T. W. (2011). The Report of the International Workshop on Using Multi-level Data from Sample Frames, Auxiliary Databases, Paradata and Related Sources to Detect and Adjust for Nonresponse Bias in Surveys*. *International Journal of Public Opinion Research, 23*(3), 389–402. https://doi.org/10.1093/ijpor/edr035

Valliant, R., Hubbard, F., Lee, S., & Chang, C. (2014). Efficient Use of Commercial Lists in U.S. Household Sampling. *Journal of Survey Statistics and Methodology, 2*(2), 182–209. https://doi.org/10.1093/jssam/smu006

West, B. T. (2013). An examination of the quality and utility of interviewer observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 211–225. https://doi.org/10.1111/j.1467-985X.2012.01038.x

West, B. T., Wagner, J., Hubbard, F., & Gu, H. (2015). The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth. *Journal of Survey Statistics and Methodology, 3*(2), 240–264. https://doi.org/10.1093/jssam/smv004

Appendix 1

List of Acronyms

Acronym	Definition
ACS	U.S. American Community Survey
BG	U.S. Census Block Group
DV	Dependent Variable
H&C	Housing & Children Study
HRS	Health and Retirement Study
HUD	United States Department of Housing and Urban Development
IV	Independent Variable
MoS	Measure of Size
MSG	Marketing Systems Group
NSFG	National Survey of Family Growth
PCA	Principle Component Analysis
PHA	Public Housing Authority
PPeS	Probability Proportionate to Estimated Size
PUMS	U.S. Census Public Microdata Sample
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SE	Standard Error
SF	Summary File
SRS	Simple Random Sample

methods, data, analyses | Vol. 18(2), 2024, pp. 165-184 DOI: 10.12758/mda.2024.02

How to Reduce Item Nonresponse in Face-to-Face Surveys? A Review and Evidence from the European Social Survey

Malte Grönemann

University of Mannheim, Graduate School of Economic and Social Sciences

Abstract

I review the literature on item nonresponse in surveys. Based on this review, I extend the satisficing model with respondents' privacy concerns to incorporate all relevant aspects of the response process for item nonresponse. I review proposed strategies to reduce item nonresponse and test selected strategies. Results suggest that boosting respondents' use of showcards and interviewing in the respondents' primary language might be promising ways to reduce item nonresponse. Other people present during the interview have only a small association with the number of refusals. Matching the age and gender of respondents and interviewers appears not to be a worthwhile strategy.

Keywords: item nonresponse, survey design, survey error, face-to-face surveys, fixed-effects negative binomial regression

Missing data pose a problem to the analysis of survey data. They decrease the effective sample size and can introduce bias to estimates if the causes for missingness are related to the item or respondent characteristics (de Leeuw, Hox, & Huisman, 2003). Although missing data are rare in most single items, they can add up to a considerable loss of observations in multivariate analyses. Item nonresponse can also be seen as an indicator of overall data quality since it can result from *satisficing* (Krosnick, 1991). Satisficing means that the respondent is giving a satisfactory answer instead of the best one. Due to these harmful effects of missing data, one objective of survey researchers is to keep their prevalence as low as possible. Therefore, it is important to understand the processes that can lead to missing data.

I review the existing literature on item nonresponse and extend the satisficing model based on this review to include privacy concerns resulting in an encompassing model of item nonresponse: the probability of item nonresponse depends on the task difficulty of the item(s) divided by the product of ability and motivation of the respondent or the respondents' privacy concerns. This means that higher difficulty results in more item nonresponse and higher ability and motivation in less item nonresponse. Difficulty, ability, and motivation are separate from privacy concerns, e.g. due to item sensitivity or general mistrust. Privacy concerns are relevant when deciding whether to disclose information and do not influence the cognitive burden of retrieving the answer.

I then turn to practical strategies that could be used to decrease item nonresponse by reviewing proposed strategies. In the empirical part of the paper, I compare the effects of selected strategies. Promoting the use of showcards and translating questionnaires appear to be most promising. Both of those reduce the cognitive burden of respondents. Matching respondents' and interviewers' gender and age does not reduce item nonresponse. They could have influenced item nonresponse if respondents are more willing to share private information with interviewers similar to themselves. The results are not causal though. How-

Acknowledgements

I want to thank Achim Koch, Johanna Mehltretter and especially Tobias Gummer for helpful comments in multiple stages of this project. I want to thank Marcel Kappes and Thomas Gautschi for methodological discussions. The manuscript also improved a lot thanks to the generous reviewers and editors of MDA.

This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences.

All code and further information can be found in the supplementary material here:
https://osf.io/m83gy/ (DOI: 10.17605/OSF.IO/M83GY).

Direct correspondence to

Malte Grönemann, University of Mannheim, Graduate School of Economic and Social Sciences, 68131 Mannheim, Germany
E-mail: malte.groenemann@uni-mannheim.de

ever, the results indicate that strategies aiming at a lower cognitive burden for respondents are our best guess to improve data quality.

This introduction follows a short theoretical discussion presenting a theoretical model for the probability of item nonresponse based on satisficing. Structured by this model, I review the literature on strategies how to reduce item nonresponse and test a selection using the *European Social Survey*. The rest of the paper is devoted to this test of strategies, describing data and methods and presenting results. Finally, I summarise and discuss my review and results.

A Model of Item Nonresponse

Whether respondents answer a survey question and which answer they give is always a cognitive process taking place at the very moment of the interview. The *survey response process* (Tourangeau, Rips, & Rasinski, 2000) involves multiple steps on behalf of the respondent. They need to comprehend the question, retrieve information from memory, eventually judge this information, map them onto the response options, and perhaps edit the response due to sensitivity or social desirability. Respondents will most likely take these steps in order but they can jump back and forth, e.g. if they need to form an opinion on the spot. But in all of these steps, item nonresponse can be introduced (de Leeuw, Hox, & Huisman, 2003).

The two types of item nonresponse, ”Don’t know” and refusal, might be related to different steps of the survey response process though. Refusals are likely introduced in the editing step when respondents do not want to answer a question although they could. They may find certain information to be too sensitive or they may not feel comfortable sharing it with the interviewer due to a lack of trust. DK is likely as an answer when the respondent cannot give a substantive answer. Either the respondent cannot answer because they do not know about the content of the question or are unable to remember an event (Beatty & Herrmann, 2002; Turner, Sturgis, & Martin, 2015). In this situation, DK is a valid response and does not constitute a problem for data quality. On the other hand, they might not see value in putting in the effort to give an optimal response and *satisfice* (Krosnick, 1991). Satisficing refers to various shortcuts (heuristics) that survey respondents can take when answering questions. One of these shortcuts is item nonresponse. The data collected in this case are of lower quality.

However, Shoemaker, Eichholz and Skewes (2002) have shown that higher mental effort is related to more refusals as well. And conversely, it is plausible that DK is used as a more polite way to refuse. In the following sections, I will therefore not distinguish between refusal and DK even though they may have varying strengths of predictors (Silber et al., 2021). Similarly, I will not consider the unproblematic case of DK as a genuine answer although differentiating between the two meanings of DK might be relevant for substantive analyses.

In the continuation of this section, I will discuss theoretical concepts that impact the likelihood of item nonresponse in the cognitive process of response formation. Later, I will combine these concepts into a theoretical model based on satisficing (Krosnick, 1991).

When it comes to item nonresponse, the key concept is the *ability* of the respondents to carry out cognitive tasks. Differences in cognitive abilities are the main explanation for differences in item nonresponse across education, age, and health (Colsher & Wallace, 1989; Pickery & Loosveldt, 1998; de Leeuw, Hox, & Huisman, 2003; Messer, Edwards, & Dillman, 2012; Silber et al., 2021). Ethnic minorities tend to have a higher rate of item nonresponse likely caused by lower literacy and worse command of the majority language (Kupek, 1998; Pickery & Loosveldt, 1998). Meitinger and Johnson (2020) conclude that item nonresponse reflects broader social inequalities in abilities and access to information. The ability hypothesis is directly supported by correlations between item nonresponse and measures of intelligence (Hedengren and Stratmann, 2012).

The second relevant concept is *task difficulty*. When questions are more difficult or unclear, they tend to have higher rates of nonresponse (Holbrook, Cho, & Johnson, 2006; Messer, Edwards, & Dillman, 2012; Holbrook et al., 2016; Olson, Smyth, & Ganshert, 2019). Demographic questions are usually easier for respondents to remember, resulting in lower rates of item nonresponse compared to attitudinal and behavioral questions, which may require respondents to formulate an answer on the spot (Olson, Smyth, & Ganshert, 2019; Silber et al., 2021).

Even if people can complete a task, they may not want to do it unless they feel that the effort is worthwhile. They need to have the *motivation* to provide an optimal response. That explains why people who are more interested in the topic of a survey are less likely to leave items unanswered (Koch & Blohm, 2009; Silber et al., 2021). Item nonresponse is linked to conscientiousness measures as well (Hedengren & Stratmann, 2012).

The editing process can also be influenced by motivation. For instance, respondents and interviewers may choose not to answer screening and filtering questions on purpose to lessen the survey workload (Tourangeau, Kreuter, & Eckman, 2015). This statement pertains only to data collections where the respondents know or can guess which questions serve as filters though.

When editing an answer, respondents may have concerns about their *privacy*[1]. Will the interviewer judge me if I answer truthfully? Can I trust that my data will be kept secure and confidential? This is a particular problem for questions perceived as intrusive (Tourangeau & Yan, 2007) like questions on income (Yan,

[1] I use this label to encompass overall privacy concerns related to the survey, such as data processing and anonymity, as well as the desire to avoid answering specific sensitive questions

Curtin, & Jans, 2010) and sexual behaviour (Kupek, 1998), which often show particularly high levels of item nonresponse. However, when it comes to attitude questions about controversial political issues such as immigration, there tends to be more item nonresponse as well (Piekut, 2021). Item nonresponse is indeed frequently used as a measure of question sensitivity (Tourangeau & Yan, 2007). Respondents will likely have such privacy concerns immediately when they hear a sensitive question and jump from the comprehension stage to the editing stage in the survey response process (Tourangeau, Rips, & Rasinski, 2000). They probably refuse to answer before an honest answer has been formed. Increased item nonresponse is associated with reluctance and skepticism towards surveys and science, general privacy concerns, and mistrust (Silber et al., 2021).

Based on the reviewed literature, I have identified four concepts that affect the probability of item nonresponse in surveys: cognitive ability, task difficulty, motivation, and privacy concerns. However, as Krosnick (1991) already hypothesised, these concepts are interrelated in their effect on item nonresponse. Very easy questions can be answered by less able respondents and very hard questions might even cause the most able to struggle. The resulting fraction of difficulty by ability represents the relative mental effort to answer a question. And a highly motivated respondent answers even difficult questions. Krosnick (1991, p. 225) formalised the probability of satisficing.

Item nonresponse is such a satisficing strategy. Additionally, higher privacy concerns lead to more item nonresponse as well. Since this relates to another step in the survey response process, namely editing rather than comprehension, retrieval or judging, I postulate it to be independent from the other concepts.

For a complete theoretical model of harmful item nonresponse, privacy concerns therefore need to be added to the model by Krosnick (1991). As these concepts are (partially) interrelated and have a nonlinear relationship to item nonresponse, it is useful to formalize and summarize their relationship as follows:

$$P_{INR} = f\left(max\left(\frac{difficulty}{ability \, * \, motivation}\,,\; privacy\right)\right)$$

The probability of an ingenuine nonsubstantive answer on behalf of the respondent is a function of the task difficulty divided by the ability and motivation of the respondent or the respondents' privacy concerns, whichever is higher.

Please note that this theoretical model is not able to and not intended to predict the probability of item nonresponse in a given item, as highlighted by the fact that it is an undefined function. Therefore, the individual concepts do not require measurement. The maximum function emphasizes that there are two independent mechanisms, and only the dominant one will impact item nonresponse at a time. This theoretical model specifically addresses item nonresponse for a single item but its meaning is adaptable to every level of a survey.

How to Reduce Item Nonresponse

With these four concepts in mind, we can develop strategies to reduce item nonresponse and ensure better data quality in our surveys. Some of the following strategies may seem obvious and are already established standards in survey design not only because of their potential relationship to item nonresponse but to ensure the quality of substantial answers as well. Others might reduce item nonresponse but they could have negative consequences for other parts of total survey error, the combined effect of all error sources in a survey (Groves & Lyberg, 2010). They require a trade-off before implementation.

I have structured this review of strategies to reduce item nonresponse by the respective concepts they target.

Task Difficulty

The level of difficulty of a task is largely determined by how the questions are designed and what type of answer is expected. To make tasks easier, it is recommended to ask short, straightforward questions that avoid any confusion or unclear concepts. Asking respondents to complete multiple tasks at once should also be avoided. For a more thorough guide on how to design questions and questionnaires, see e.g. Smyth (2016). The difficulty of a task is related to the type of question as well. Questions that are open-ended or allow for multiple options and ordering of categories are more likely to result in higher nonresponse rates than closed single choice items (Schuman & Presser, 1979b; Holbrook, Cho, & Johnson, 2006; Holbrook et al., 2016; Olson, Smyth, & Ganshert, 2019; Silber et al., 2021). To make it easier for respondents, visual aids like images or showcards can be used. Showcards eliminate the need for respondents to recall all response categories while answering a question. However, there is limited research on how showcards affect item nonresponse. According to a study by Holbrook, Johnson et al. (2016), using showcards in survey questions led to more unanswered items. However, this may have been because showcards were only used for more challenging questions. In the European Social Survey (ESS), showcards do not appear to impact the distribution of meaningful responses in survey experiments, as noted by (Jäckle, Roberts, & Lynn, 2010), although they did not investigate item nonresponse.

How question design affects levels of item nonresponse is very well understood and differences between questions constitute the largest part of the variance in item nonresponse (Olson, Smyth, & Ganshert, 2019). This highlights the importance of the single question for overall data quality.

It is important to design the entire questionnaire as simply as possible, not just the individual questions. Questionnaires that include changes in response scales, routing, and filtering tend to result in higher rates of nonresponse (Messer, Edwards, & Dillman, 2012). However, routing and filtering should not increase difficulty in computer-assisted modes. Grouping questions by topic could reduce the required mental effort and item nonresponse but it also increases the likelihood of non-differentiation between items (Krosnick, 1991). Explicitly offering DK and refusal options can increase their use, as it makes respondents more aware of the possibility of an ”easy way out” (Schuman & Presser, 1979a; Beatty & Herrmann, 2002).

To reduce task difficulty for members of language minorities, the questionnaire can be translated so that respondents can take the interview in the language they are most proficient in. But translating questionnaires can be costly and may affect the comparability of cases. For a review on comparability in cross-cultural surveys, see e.g. Behr and Shishido (2016).

To enhance the quality of survey design, identify any errors, and ensure that respondents can complete the required tasks, it is recommended to thoroughly review the questionnaire and its implementation for data collection. Common methods for doing so include conducting reviews and pilot studies.

Ability

While we cannot alter the general cognitive ability of our respondents, we can influence their ability to answer survey questions at the time of participation. To ensure a productive interview, it is important to choose an environment that encourages focus and clear communication. If possible, opt for quiet and not distracting locations at appropriate times. Having other people present during an interview can be distracting, but there is no conclusive evidence to support this claim (Kupek, 1998; Tu & Liao, 2007; Silber et al., 2021). Respondents may become fatigued during lengthy interviews (Holbrook et al., 2016; Olson, Smyth, & Ganshert, 2019).

Motivation

Motivation could decrease throughout the interview as well. While web surveys have used different page layouts and progress bars to combat this issue, the effectiveness of these methods is uncertain (Peytchev et al., 2006; Yan et al., 2011; Sarraf & Tukibayeva, 2014). The research on cooperation enhancement, such as through incentives, has mainly focused on unit nonresponse. However, some of these methods could also be effective in increasing item nonresponse. After all, unit and item nonresponse are linked: respondents that initially refused to participate have higher levels of item nonresponse (Yan & Curtin, 2010; Fricker & Tourangeau, 2010).

Privacy Concerns

Survey researchers should address privacy concerns to encourage respondents to answer by ensuring the security and anonymity of their data. It is important to communicate why the data is collected, how it will be processed, and how privacy is protected. This is not only ethically advisable but also often a legal requirement.

When conducting face-to-face surveys, the trust between the respondent and interviewer is influenced by their relationship. Scholars have hypothesized that respondents are more likely to trust interviewers who they perceive to be similar to themselves. To test this hypothesis, studies have been conducted to examine the impact of matching characteristics between the respondent and interviewer. Vercruyssen, Wuyts and Loosveldt (2017) observe less item nonresponse when interviewers and respondents are matched in age. Additionally, matching gender reduces item nonresponse for males but increases it for females. Piekut (2021) found female interviewers experienced higher rates of nonresponse but there was no significant correlation between the gender of the interviewer and the gender of the respondent. Silber et al. (2021) found that education matching has no effect while Tu and Liao (2007) find age and education matching to be potent predictors of item nonresponse. A test that could be interesting to conduct is whether pairing interviewers and respondents who share the same immigration status and/or ethnicity would make a difference. Immigrants tend to have higher levels of item nonresponse, language barriers, and I could imagine that some of them may be mistrustful towards interviewers due to racist experiences and a fear of discrimination.

Strategies to be Tested

So far, this article has reviewed and summarized the literature on item nonresponse in surveys. I have suggested a theoretical model, an extension of the satisficing model by Krosnick (1991), as a conceptual summary that can inform our survey design and I have reviewed strategies to reduce item nonresponse and categorized them accordingly. In the remainder of the article, I am going to test a few selected strategies to reduce item nonresponse derived from the theoretical model and the literature review. All of these strategies could change at least one of the four concepts from the theoretical model and therefore could have a causal connection to item nonresponse. Whether these strategies actually do change the associated concepts and how strongly their effect translates into changes in item nonresponse will be central to my empirical analysis.

1. During an interview, respondents may experience a decrease in concentration and motivation to answer questions as time goes on. As a result, item

nonresponse may become more common the longer the interview lasts. To maintain high data quality and reduce item nonresponse, it might be advisable to keep questionnaires as short as possible.

2. It is likely that interferences and the presence of others during an interview can cause item nonresponse, as they may distract the respondent and make them hesitant to answer certain questions in front of people they know. As a result, it might be beneficial that interviews are conducted without other people present, if feasible.
3. To make answering easier for respondents, showcards could be provided so they do not have to remember the response scale. Showcards would then lower the required cognitive effort and reduce item nonresponse.
4. Respondents who primarily speak a different language at home may face difficulties in the response process. To ensure data quality from these respondents, one option is to translate the questionnaire, although this can be costly and may present comparability problems.
5. According to previous studies, people may feel more comfortable answering questions from interviewers who share similar social characteristics, such as gender and age. If true, survey agencies could assign interviewers based on demographic information if it is available in the sampling frame.

Table 1 summarises the selected strategies that I am going to test in my empirical analysis. The second column shows which concepts play a role in the hypothesized mechanism linking the respective strategy to item nonresponse. The third column gives the expected direction of the relationship between strategy and item nonresponse, e.g. the longer the questionnaire, the more item nonresponse. These are also the expected signs of the coefficients if the strategies work as imagined.

Table 1 Reduction Strategies to be Tested

Strategy	Mechanism	Expectation
Length of the Questionnaire	Ability, Motivation	positive
Interference of the Interview	Ability, Privacy	positive
Use of Showcards	Difficulty	negative
Interview not primary Language	Difficulty	positive
Gender Matching	Privacy	negative
Interviewer more than 10 years older	Privacy	positive
Respondent more than 10 years older	Privacy	positive

Data and Methods

Data

To test the effectiveness of some potential strategies to reduce item nonresponse, I use the *European Social Survey (ESS)* Round 9 collected between August 2018 and January 2020 (ESS ERIC, 2019). The ESS is a biannual face-to-face trend survey on attitudes and beliefs towards social and political topics in Europe established in 2001. In each country and round, the ESS draws a new random sample of the residential population of 15 years and older aiming for a minimum response rate of 70%. Most countries use computer-assisted personal interviews for data collection and the questionnaire is designed to take about one hour. The data release 3.1 includes data from 49,519 respondents from 29 countries. For more information on the data, see the supplementary material or visit europeansocialsurvey.org.

Dependent Variables

The three dependent variables are the sum of DK, the sum of refusals, and the total sum of item nonresponse for every respondent. Non-responses are only counted for variables that are presented to all respondents and not affected by filtering questions. Respondents are not given the option to respond with DK or refusal, but interviewers are instructed to record them explicitly and without further probing. It is up to the interviewer to interpret a non-response as either a refusal or DK.

Although my argument focuses on item nonresponse which is problematic for data quality opposed to DK as a genuine answer, I have not separated the two meanings in the analysis for two reasons. Firstly, distinguishing between these two meanings is often very challenging, and it requires a deep understanding of the specific question, which is not feasible for this general analysis. Secondly, an additional mechanism that generates item nonresponse may increase the overall variation in the dependent variables but if it is uncorrelated to the other mechanisms, no bias in estimates is to be expected. I do not know how the possibility of genuine DK answers could interfere with the other mechanisms. I therefore assume that they are uncorrelated.

Control Variables

As my empirical analysis is concerned with strategies that potentially could be used in survey design and implementation to reduce problematic item nonresponse and therefore to increase data quality, it aims at *causal inference* (Angrist & Pischke, 2009): Do we expect a difference in item nonresponse if a strategy was implemented compared to the counterfactual when it was not implemented? Or in other words, does the implementation of a strategy *cause* a net decrease in item nonresponse on average?

To identify the average treatment effects of the selected strategies with cross-sectional survey data, I need to control for potential sources of bias in the effects of the strategies, other unrelated influences can be omitted. Such selection of controls always requires a sufficiently complete theory. In this case, the selection of controls can be based on the theoretical model outlined earlier.

In my analysis, I need to control for respondents' ability as it is likely related to respondents' understanding of survey procedures like showcard use. Ability also needs to be controlled to estimate the effect of language differences as immigrant and minority groups in Europe typically differ in education compared to the majority groups. Ability is also related to the respondents' age and could therefore bias the effect of matching interviewers' and respondents' characteristics.

As the use of showcards is evaluated by the interviewer after the interview, the test of the effectiveness of showcards has an endogeneity problem. The overall impression the interviewer has of the respondent might influence the perception of showcard use. I, therefore, control for the interviewer's general impression of the interview.

I am not aware of any mechanisms that could lead to biased estimates for the effects of interferences and other people being present during the interview as well as whether respondent and interviewer have matching gender. In summary, necessary controls are therefore respondents' ability and specifically age and the interviewers' overall assessment of the interview. Based on the theory outlined above, I do not expect that the inclusion of any of these control variables or the other strategies is likely to distort the effect of another variable of interest. Therefore and to be able to compare relative effect size, I am going to test all effects in a single regression model.

However, identifying causal effects in regression modeling requires the conditional independence assumption (Angrist & Pischke, 2009, 52ff) that all sources of bias are sufficiently controlled for. This is a strong assumption as it requires not only a sufficient theory (and the sufficiency of a theory is improvable) but also the operationalization, measurement, and functional form of the statistical model needs to be correct. This is never the case in social research (Martin, 2018). Even though I have carefully selected the controls based on the presented theoretical model, I can only use proxies for the concepts I need to control for. I will therefore not speak of causal effects but of (conditional) associations as the point estimates can still be slightly off. Nonetheless, the regression estimates should reveal which strategies work and which are the most promising for implementation. Future experimental research could investigate the most promising strategies more thoroughly.

Independent Variables of Interest

I calculate the number of questions the respondent was asked by subtracting the number of items coded as not applicable from the total number of questions. Whether the interview was conducted in the respondent's primary language is a dummy variable generated from the metadata in which language the interview was conducted and the respondent's answer to the question of which language they primarily speak at home. Matching social characteristics are also dummy variables and generated from demographic information from the main questionnaire and the interviewer questionnaire. The interviewer questionnaire is a short questionnaire the interviewer fills out after completing the interview. For matching ages, I constructed two dummies whether the respondent is more than ten years older or younger. The reference category is whether the age difference is ten years maximum. I went for a cutoff difference of ten years to have a meaningful and visually perceivable difference in age and enough observations in all categories. The interviewer questionnaire also asks whether other people were present during the interview or not. And interviewers rate the respondents' use of showcards on a three-point scale: respondent used all the applicable showcards, respondent used only some applicable showcards, respondent refused/was unable to use the showcards at all. I treat this latter variable as metric with higher values indicating more frequent use of showcards.

To control for ability, I use education (operationalized by the ISCED scale), age, and squared age of the respondent as proxies. To dampen the endogeneity problem of interviewers' assessment of showcard use, I include the interviewers' assessment of how well the respondent understood the questions, to what extent the respondent answered to the best of their ability, and how often they asked for clarifications. They serve as proxies for ability as well.

Statistical Model

Since the dependent variables are count data and show the typically skewed distribution of count data, I analyze the data using a negative binomial regression with interviewer fixed effects (Allison & Waterman, 2002). The interviewer fixed effects are used to control for mean differences in interviewer behavior regarding accepting and recording item nonresponse. At the same time, they absorb variation between countries. Standard errors are clustered by the interviewer following recommended practice to prevent heteroscedasticity (Cameron & Trivedi, 2013, 358f). Since the population of interest are interviews, no weighting is applied. Missing data are deleted listwise.

The analyses are carried out in *R* (R Core Team, 2023) using the *Tidyverse* (Wickham et al., 2019) for data handling and graphics and the *fixest* package (Bergé, 2018) to estimate the regressions.

More information on the data, variables, summary statistics, all of the code used for preparation and analysis, and discussions on missing values and model specification are available in the supplemental material.

Count data models with fixed effects are quite debated (Wooldridge, 1999; Allison & Waterman, 2002; Cameron & Trivedi, 2013). For a thorough discussion of model choice, see the supplementary material as well.

Results

Figure 1 shows the incidence rate ratios (exponentiated coefficients) of the coefficients of interest. The coefficients to evaluate the tested strategies and of the control variables as well as standard errors and coefficients of model fit can be found in Table 2.

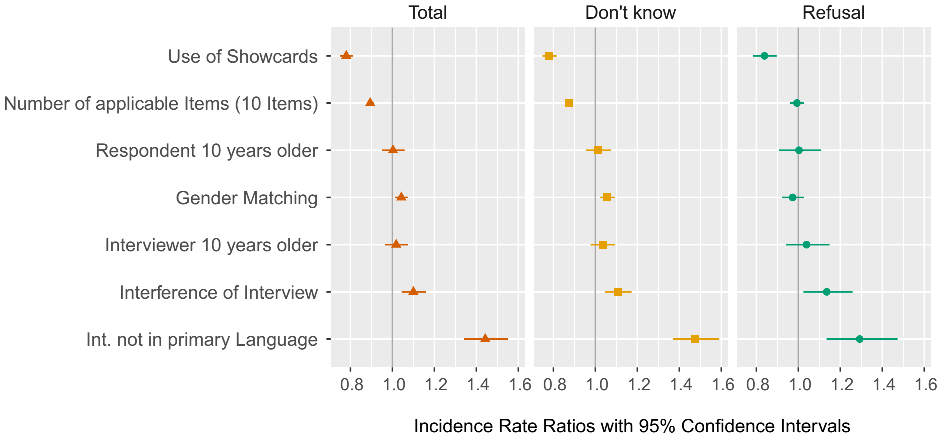


Figure 1 Coefficients of Interest

Table 2 Regression Results

Dependent Variables	Don’t know	Refusal	Total
Model	(1)	(2)	(3)
Variables			
Number of applicable Items (10 Items)	-0.133***	-0.007	-0.112***
	(0.010)	(0.017)	(0.009)
Interference of Interview	0.102***	0.127*	0.095***
	(0.029)	(0.053)	(0.027)
Use of Showcards	-0.247***	-0.176***	-0.249***
	(0.022)	(0.034)	(0.020)
Int. not in primary Language	0.389***	0.256***	0.366***
	(0.038)	(0.067)	(0.037)
Gender Matching	0.055**	-0.028	0.042**
	(0.017)	(0.027)	(0.015)
Respondent 10 years older	0.012	0.003	0.003
	(0.030)	(0.051)	(0.028)
Interviewer 10 years older	0.033	0.038	0.018
	(0.029)	(0.051)	(0.027)
Education (ISCED)	-0.115***	0.068***	-0.085***
	(0.005)	(0.009)	(0.005)
Age	-0.034***	0.009	-0.029***
	(0.003)	(0.005)	(0.003)
Age squared	0.0004***	-4.07×10^{-5}	0.0003***
	(2.54×10^{-5})	(4.79×10^{-5})	(2.41×10^{-5})
Understood Questions	-0.348***	-0.052	-0.298***
	(0.017)	(0.027)	(0.016)
Answered to best Ability	-0.029	-0.180***	-0.057***
	(0.015)	(0.023)	(0.014)
Amount of Clarifications	0.254***	0.598***	0.330***
	(0.011)	(0.021)	(0.011)
Fixed-effects			
Interviewer	Yes	Yes	Yes
Fit statistics			
Observations	43,745	36,745	44,000
Pseudo R^2	0.12036	0.18128	0.12289
Within Pseudo R^2	0.05886	0.08776	0.05999
BIC	199,301.2	92,395.0	214,873.6
Over-dispersion	0.89873	0.75842	1.0379

Clustered (Interviewer) standard-errors in parentheses
Signif. Codes: ***: 0.001, **: 0.01, *: 0.05

The most promising ways to reduce item nonresponse seem to be boosting the use of showcards and translating questionnaires. With more frequent showcard use as indicated by the interviewer, the amount of DK reduces by about a quarter and the amount of refusal by about 18% on average. And compared to interviews conducted in the language the respondent primarily speaks at home, interviews conducted in a language different from the respondents' primary language show on average 42% more DKs and 26% more refusals.

As a general observation for all variables, the effects on the total number of item nonresponse closely mirror the effect on DK. This is not surprising since there are many more DKs than refusals. The effects on the number of refusals are typically weaker than the effect on DK but still present. This supports the idea that refusals and DKs are not perfectly separate in their meaning but not identical as well. For the variables presented so far, stronger effects on DK make substantial sense as well since they are all based on respondents' ability or difficulty of the task.

Matching respondents' gender has a small positive effect on the number of DKs. This is contrary to expectations, which suggested that matching the socio-demographics of interviewers and respondents leads to a more trusting interview situation and reduces item nonresponse. The effect on refusals is not significant but should be pronounced since this strategy is partly based on the privacy mechanism. Matching by age has no significant effect. Matching respondents and interviewers seems not to be a promising strategy to reduce item nonresponse.

Other people present during the interview raised the number of refusals by 14% in line with the reasoning that respondents do not want to answer some questions in the presence of others they know. The effect on DK is not significant. Other people present might therefore influence item nonresponse more via privacy than a distraction. However, due to the relatively small number of refusals, the effect on the total item nonresponse is not significant.

Contrary to expectation, the number of applicable items has a significantly negative effect on DK (and total item nonresponse). A respondent that has been asked 10 questions more has a 12% lower average number of DK answers.

Discussion

I have reviewed the literature on item nonresponse and extended the cognitive *satisficing* model (Krosnick, 1991) with concerns about privacy to encompass all aspects that can interfere with the response process in survey interviews. Organizing our knowledge into such theoretical models highlights the interrelations between theoretical constructs which is necessary to reduce total error and is not achievable with piece-meal empirical studies. Based on this new model, I have reviewed possibilities to reduce item nonresponse, particularly in face-to-face surveys.

In an empirical analysis using data from the *European Social Survey Round 9*, I found that boosting the respondents' use of showcards and conducting the interview in the respondents' primary language might be promising ways to reduce item nonresponse in face-to-face surveys. These strategies reduce the cognitive effort on behalf of respondents. Other people present during the interview are moderately associated with more refusals. Respondents are probably unwilling to disclose private information in front of people they know. However, my hypothesis that respondents might trust interviewers more and share more information if the interviewer and respondent are socially similar has received no support: matching the socio-demographic characteristics of interviewers and respondents seems not a worthwhile strategy. And surprisingly, longer questionnaires were associated with less item nonresponse. However, this might be related to a problem of operationalization. Most questions that might not apply to respondents are demographics asked at the end. But this would explain no association, but I observe a negative effect for which I do not have an explanation.

Although I carefully selected control variables, I cannot rule out violations of the *conditional independence assumption* which is necessary to identify causal effects with regression analysis. Most variables are influenced by respondents' ability (e.g. to understand survey procedures) as is item nonresponse. Respondents' ability is notoriously hard to measure in surveys and proxies like education and age that I have used as controls are not perfect. A second threat to the results is the endogeneity of some variables of interest, in particular, showcard use and others present during the interviews. They are measured in the interviewer questionnaire after the interview and are likely biased by the interviewers' overall assessment of the interview, including the amount of item nonresponse. I tried to control for that using other variables from the interviewer questionnaire. A third limitation of this analysis concerns the external generalizability of the results. We know that specific types of questions are more prone to item nonresponse, for example, opinions and sensitive questions. The results obtained here reflect the effects on item nonresponse especially on matters of opinion as this is the primary object of study of the ESS. While opinion surveys constitute a large share of surveys and the results should be generalizable to them, other types of surveys might have slightly different challenges concerning item nonresponse.

I nonetheless see value in this analysis for two reasons. First, item nonresponse is relatively rare in single items and therefore difficult to study using survey experiments. Second, and more importantly, the empirical analysis aims to compare multiple potential strategies in their strength of effect (which is not possible using experiments). While it is difficult to assess the true causal effect of the strategies that do make a difference, the strategies that have no effects even in this biased analysis will likely not be successful. This analysis can provide a meaningful starting place for more rigorous tests of the most promising strategies and nonetheless inform survey design choices.

Although I analyze data from a face-to-face survey, I think it is important to anticipate some of the results and especially the implications of the theoretical model for the current shift to self-administrating modes of data collection. The results of my analysis highlight that the most promising strategies to decrease item nonresponse are tools that decrease task difficulty, like showcards and translating questionnaires. In self-administered modes, designing easy-to-use and clear questionnaires and page layouts will be important for item nonresponse. For paper-based modes, this will limit the options for routing and filtering. Specifically, offering refusal and DK as response options is an important design choice. Ethically, respondents need to have the option not to respond. On one hand, this will likely increase item nonresponse. On the other hand, forcing an answer will generate low-quality responses. In self-administered modes, we have less control over the interview situation, for example, whether other people are present. My analysis has shown that the latter is associated with more refusals. The absence of an interviewer reduces social desirability. No social desirability is often considered an advantage as respondents do not need to disclose information to a stranger. But no interviewer could also be a disadvantage as there might be a lower hurdle to satisfice. But the strategies based on the idea that respondents trust socially similar interviewers indicated that the presence of the interviewer might be less important for general levels of item nonresponse. Finally, self-administered modes can be conducted in multiple languages easily because we do not need interviewers that speak a minority language.

References

Allison, P. D., & Waterman, R. P. (2002). Fixed-Effects Negative Binomial Regression Models. *Sociological Methodology, 32*(1), 247–265. https: //doi.org/10.1111/1467-9531.00117

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. OCLC: ocn231586808.

Beatty, P., & Herrmann, D. (2002). To Answer or Not to Anwer: Decision Processes Related to Survey Item Nonresponse. In R. M. Groves (Ed.), *Survey Nonresponse*. Wiley.

Behr, D., & Shishido, K. (2016). The Translation of Measurement Instruments for Cross-Cultural Surveys. In *The SAGE Handbook of Survey Methodology* (pp. 269–287). SAGE Publications Ltd. https://doi.org/10.4135/9781473957893.n19

Bergé, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: The R package FENmlm. *CREA Discussion Papers, 13*.

Cameron, A. C., & Trivedi, P. (2013). *Regression Analysis of Count Data* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139013567

Colsher, P. L., & Wallace, R. B. (1989). Data Quality and Age: Health and Psychobehavioral Correlates of Item Nonresponse and Inconsistent Responses. *Journal of Gerontology, 44*(2), 45–52. https://doi.org/10.1093/geronj/44.2.P45

de Leeuw, E. D., Hox, J., & Huisman, M. (2003). Prevention and Treatment of Item Nonresponse. *Journal of Official Statistics, 19*(3), 153–176.

ESS ERIC. (2019). European Social Survey (ESS), Round 9 - 2018. https://doi.org/10.21338/NSD-ESS9-2018

Fricker, S., & Tourangeau, R. (2010). Examining the Relationship Between Nonresponse Propensity and Data Quality in Two National Household Surveys. *Public Opinion Quarterly, 74*(5), 934–955. https://doi.org/10.1093/poq/nfq064

Groves, R. M., & Lyberg, L. E. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly, 74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Hedengren, D., & Stratmann, T. (2012). The Dog that Didn't Bark: What Item Nonresponse Shows about Cognitive and Non-Cognitive Ability. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2194373

Holbrook, A. L., Cho, Y. I., & Johnson, T. P. (2006). The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties. *Public Opinion Quarterly, 70*(4), 565–595. https://doi.org/10.1093/poq/nfl027

Holbrook, A. L., Johnson, T. P., Cho, Y. I., Shavitt, S., Chavez, N., & Weiner, S. (2016). Do Interviewer Errors Help Explain the Impact of Question Characteristics on Respondent Difficulties? *Survey Practice, 9*(2), 1– 11. https://doi.org/10.29115/SP-2016-0009

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review, 78*(1), 3–20. https://doi.org/10.1111/j.1751-5823.2010.00102.x

Koch, A., & Blohm, M. (2009). Item Nonresponse in the European Social Survey. *ASK: Research & Methods, 18*(1), 45–65.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Kupek, E. (1998). Determinants of Item Nonresponse in a Large National Sex Survey. *Archives of Sexual Behavior, 27*(6), 581–594.

Martin, J. L. (2018). *Thinking through statistics*. The University of Chicago Press.

Meitinger, K. M., & Johnson, T. P. (2020). Power, Culture and Item Nonresponse in Social Surveys. In P. S. Brenner (Ed.), *Understanding Survey Methodology* (pp. 169–191, Vol. 4). Springer. https://doi.org/10.1007/978-3-030-47256-6 8

Messer, B. L., Edwards, M. L., & Dillman, D. A. (2012). Determinants of Item Nonresponse to Web and Mail Respondents in Three AddressBased Mixed-Mode Surveys of the General Public. *Survey Practice, 5*(2), 1–9. https://doi.org/10.29115/SP-2012-0012

Olson, K., Smyth, J. D., & Ganshert, A. (2019). The Effects of Respondent and Question Characteristics on Respondent Answering Behaviors in Telephone Interviews. *Journal of Survey Statistics and Methodology, 7*(2), 275–308. https://doi.org/10.1093/jssam/smy006

Peytchev, A., Couper, M. P., McCabe, S. E., & Crawford, S. D. (2006). Web Survey Design: Paging versus Scrolling. *Public Opinion Quarterly, 70*(4), 596–607. https://doi.org/10.1093/poq/nfl028

Pickery, J., & Loosveldt, G. (1998). The Impact of Respondent and Interviewer Characteristics on the Number of "No Opinion" Answers. *Quality & Quantity, 32,* 31–45.

Piekut, A. (2021). Survey nonresponse in attitudes towards immigration in Europe. *Journal of Ethnic and Migration Studies, 47*(5), 1136–1161.
https://doi.org/10.1080/1369183X.2019.1661773

R Core Team. (2023). R: A language and environment for statistical computing.

Sarraf, S., & Tukibayeva, M. (2014). Survey Page Length and Progress Indicators: What Are Their Relationships to Item Nonresponse? *New Directions for Institutional Research, 2014*(161), 83–97. https://doi.org/10.1002/ir.20069

Schuman, H., & Presser, S. (1979a). The Assessment of ”No Opinion” in Attitude Surveys. *Sociological Methodology, 10*, 241. https://doi.org/10.2307/270773

Schuman, H., & Presser, S. (1979b). The Open and Closed Question. *American Sociological Review, 44*(5), 692–712. https://doi.org/10.2307/2094521

Shoemaker, P. J., Eichholz, M., & Skewes, E. A. (2002). Item Nonresponse: Distinguishing between don't Know and Refuse. *International Journal of Public Opinion Research, 14*(2), 193–201. https://doi.org/10.1093/ijpor/14.2.193

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),* 1–18.
https://doi.org/10.1111/rssa.12703

Smyth, J. D. (2016). Designing Questions and Questionnaires. In *The SAGE Handbook of Survey Methodology* (pp. 218–235). SAGE Publications Ltd.
https://doi.org/10.4135/9781473957893.n16

Tourangeau, R., Kreuter, F., & Eckman, S. (2015). Motivated Misreporting: Shaping Answers to reduce Survey Burden. In U. Engel (Ed.)*, Survey measurements: Techniques, data quality and sources of error.* Campus Verlag.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859–883. https://doi.org/10.1037/00332909.133.5.859

Tu, S.-H., & Liao, P.-S. (2007). Social Distance, Respondent Cooperation and Item Nonresponse in Sex Survey. *Quality & Quantity, 41*(2), 177–199.
https://doi.org/10.1007/s11135-007-9088-0

Turner, G., Sturgis, P., & Martin, D. (2015). Can Response Latencies Be Used to Detect Survey Satisficing on Cognitively Demanding Questions? *Journal of Survey Statistics and Methodology, 3*(1), 89–108. https://doi.org/10.1093/jssam/smu022

Vercruyssen, A., Wuyts, C., & Loosveldt, G. (2017). The effect of sociodemographic (mis) match between interviewers and respondents on unit and item nonresponse in Belgium. *Social Science Research*, *67*, 229– 238.
https://doi.org/10.1016/j.ssresearch.2017.02.007

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., Francois, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686.
https://doi.org/10.21105/joss.01686

Wooldridge, J. M. (1999). Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics, 90*(1), 77–97.
https://doi.org/10.1016/S0304-4076(98)00033-5

Yan, T., Conrad, F. G., Tourangeau, R., & Couper, M. P. (2011). Should I Stay or Should I go: The Effects of Progress Feedback, Promised Task Duration, and Length of Questionnaire on Completing Web Surveys. *International Journal of Public Opinion Research, 23*(2), 131–147. https://doi.org/10.1093/ijpor/edq046

Yan, T., & Curtin, R. (2010). The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective. *International Journal of Public Opinion Research, 22*(4), 535–551. https://doi.org/10.1093/ijpor/edq037

Yan, T., Curtin, R., & Jans, M. (2010). Trends in Income Nonresponse Over Two Decades. *Journal of Official Statistics, 26*(1), 145–164.

methods, data, analyses | Vol. 18(2), 2024, pp. 185-212 DOI: 10.12758/mda.2023.05

Adjusting to the Survey: How Interviewer Experience Relates to Interview Duration

André Pirralha, Christian Haag &
Jutta von Maurice

Center for Study Management (CSM), Leibniz Institute for Educational Trajectories (LIfBi)

Abstract

Interview duration has been shown to become shorter as fieldwork progresses. This has been attributed to a learning effect interviewers go through as they gather experience. In this study, we expand on this knowledge by focusing on how two kinds of interviewer experience relate to interview duration in telephone surveys. Using data from the German National Educational Panel Study (NEPS), we employ multilevel models, accounting for the clustering of respondents within interviewers. The results strengthen previous findings associating within-survey interviewer experience with decreasing interview duration. On the other hand, countering previous work, we find evidence that interview duration also decreases with overall interviewer experience. The results add to our knowledge concerning the effect of interviewer experience in the telephone survey mode. The effects are robust to several model specifications and to different interviewer, respondent, and interview characteristics. We conclude with a discussion about how to manage interviewer experience during training and fieldwork.

Keywords: Interview Duration; Interviewer experience; CATI surveys; panel data; paradata

Interviewers are important actors in the collection of standardized data in survey studies. For example, they set the pace of an interview, elicit respondent cooperation, or help and warn respondents of the cognitive effort they should put into their responses (Ackermann-Piek & Massing, 2014; Olson, Smyth, Dykema, et al., 2020b; West & Blom, 2017). Interview duration is a widely used and easy to measure indicator to monitor and evaluate fieldwork as it allows to detect deviations from the standardized interview protocol and because of its impact in determining survey costs in interviewer-administered studies (Jin et al., 2019; Lepkowski et al., 2010; Vandenplas et al., 2019).

Ever since the seminal work of Olson and Peytchev (2007), it is well-known that the time spent on administering a survey interview differs extensively not only between interviewers but also throughout the fieldwork phase, with a clear tendency for interviews to become shorter (Böhme & Stöhr, 2014; Kirchner & Olson, 2017; Kosyakova et al., 2021; Loosveldt & Beullens, 2013b). This pattern has been associated with a learning effect interviewers take on when conducting interviews within the same study. There is also evidence that overall interviewer experience is important, with more and less experienced interviewers varying in how they follow survey protocols (Fowler & Mangione, 1990; Kirchner & Olson, 2017). While there is extended empirical evidence showing that interviewers in face-to-face survey modes (Computer-Assisted Personal Interview – CAPI) tend to become faster as fieldwork progresses (Kosyakova et al., 2021; Loosveldt & Beullens, 2013b; Vandenplas et al., 2018), there are only very few attempts directed to uncover whether this same effect also holds in telephone surveys (Computer-Assisted Telephone Interview – CATI). In particular, the most relevant publications focusing on CATI surveys are Kirchner and Olson (2017) as well as Olson and Smyth (2015, 2020), which support the finding that interview duration decreases throughout fieldwork and that within-survey interviewer experience is associated with shorter interviews. Nonetheless, these papers use data and paradata from the same study and for showing the robustness of these results other empirical analyses should be sought.

In this paper, we focus on the effect of interviewer experience on interview duration. We analyze how interviewer experience is associated with interview duration in a probability-based telephone survey of an educational panel study, with a sample of parents of children in school-age in Germany. Our contribution to the literature is an empirical examination of the relation between two aspects of interviewer experience and interview duration in a telephone survey, while accounting for an extended set of interviewer, respondent, and interview charac-

Direct correspondence to
André Pirralha, Leibniz-Institut für Bildungsverläufe, Zentrum für Studienmanagement,
Wilhelmsplatz 3, 96047 Bamberg
E-mail: andre.pirralha@lifbi.de

teristics. We add to previous work by exploring telephone interview data from a large-scale panel survey, consisting of a substantial number of both respondents and randomly assigned interviewers with high variability of characteristics at both sides. We employ a multilevel modeling strategy aiming to disentangle how different aspects of interviewer experience are associated with interview duration and to account for the clustering of respondents and interviewers. Finally, we discuss some of the implications of our results for interviewer training and fieldwork management.

The Role of Interviewer Experience for Interview Duration

Interviewer effects are expected to differ between survey modes to some extent because “the mode or device for the interaction changes the nature of the interaction between interviewers and respondents” (Olson, Smyth, Dykema, et al., 2020b, p. 6). This should be the case with face-to-face and telephone interviews due to particular aspects of field control and payment scheme. In particular, CATI interviewers are subjected to more direct quality controls, as interviewing takes place in a centralized location and the survey agency actively supervises the outcome and the interviewers’ actions (Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V [ADM] et al., 2021; Kosyakova et al., 2021; Stiegler & Biedinger, 2015).

Even if interviewer effects may be less pronounced than in CAPI surveys, research has shown that around 25% of the variance in interview duration in CATI surveys is due to interviewers (Kirchner & Olson, 2017; Olson & Smyth, 2015). What is less clear, though, is how this can be explained. Previous studies argue that, as the fieldwork advances, interviewers progressively collect task-related experience and/or implement prior interviewing experience, thus speeding up and reducing interview time (Kirchner & Olson, 2017; Kosyakova et al., 2021). The research discusses a variety of possibilities that can explain this finding: learning effects can lead to an increase in interviewers’ reading speed; a reduction of misreading, corrections, or use of filler words; reduced time that has to be invested in reading additional interviewer instructions in computer-based modes; the reduction of (unnecessary) side communications; more efficient prompting behavior and subsequently quicker and/or clearer responses of the interviewee (Ackermann-Piek & Massing, 2014; Kosyakova et al., 2021; Olson, Smyth, Dykema, et al., 2020a).

A less benevolent view argues that shorter interviews throughout fieldwork are the result of behaviors that are grossly inconsistent with the interview protocol’s standardized practices, such as shortening of introduction texts, deviations from instructions concerning how often answer schemes are to be read, or accepting interviewees' nonresponse too quickly (for instance in case of items that are sensitive or might lead to longer discussions). In a more extreme form, interviewers might even skip items or avoid entire loops in the instrument by influencing answers in filter questions, they might phrase items liberally, or even suggest answers (Kosyakova et al., 2021; Olson & Smyth, 2020). This 'misbehavior' can be coupled with the interviewer's payment scheme and the main interest of finishing the interview as quickly as possible (Vandenplas et al., 2018).

Regardless of what specific behavior accounts for this effect, previous research has shown that interviewer experience impacts interview duration, albeit to a different degree and depending on the mode of data collection (Vandenplas et al., 2019). Olson and Smyth (2015) demonstrate that telephone interviews become shorter as the fieldwork progresses. Olson and Peytchev (2007), in addition to finding no difference between CAPI and CATI modes of data collection, also observe that inexperienced interviewers speed up faster which might be related to a more pronounced learning effect. Loosveldt and Beullens (2013b, p. 1429) report that "interviewers strongly determine the interview speed" and they do so to a greater degree than the respondents. Substantial evidence from the literature shows that within-survey interviewer experience is positively associated with declining interview length (Kirchner & Olson, 2017; Kosyakova et al., 2021; Loosveldt & Beullens, 2013a; Olson & Smyth, 2015; Vandenplas et al., 2019). Kirchner and Olson (2017) show that the behavior of telephone interviewers, which is derived from experience and resulting in decreasing interview duration over the field phase, remains of influence regardless of changes in the composition of the sample. All this previous research underscores that interviewer experience remains an important factor in understanding interview duration, with other aspects potentially adding nuance to this relationship. In light of this well-established understanding, this study seeks to revisit and build upon these findings, providing both a confirmation and potential new insights especially for CATI mode. Interviewer experience is usually defined through two specific facets of survey interviewing: within-survey interviewer experience and overall interviewer experience (Kirchner & Olson, 2017; Kosyakova et al., 2021). Within-survey interviewer experience relates to the experience gathered by the interviewer in one specific survey or wave. On the other hand, overall interviewer experience is defined as being independent of a given study or wave and the result of the total time working as a survey interviewer. While probably both facets of interviewer experience operate through the mechanism of learning, they might differ in actual interviewer behavior. Whereas within-survey interviewer experience could be more closely linked to the growing knowledge concerning the specific instrument currently in the field, overall interviewer experience acknowledges the possibility that some interviewer behaviors can originate from professional knowledge or experience unrelated to the current study. Thus, for within-survey interviewer experience, we expect the interview duration to decrease as the interviewer gathers more experience by conducting more interviews in the same survey.

On the other hand, overall experienced interviewers have a greater level of knowledge and routine about how to conduct interviews, as well as on the protocols of the specific survey institute, thus they operate differently with the specificities of the study currently in the field (Kirchner & Olson, 2017; Kosyakova et al., 2021). Based on this discussion, we expect that more overall interviewer experience will lead, on average, to shorter interview duration when compared to less experienced interviewers. We formulate this expectation even though overall interviewer experience was not previously found to affect interview duration in telephone survey interviews (Kirchner & Olson, 2017; Olson & Peytchev, 2007).

Additionally, the possibility that both types of interviewer experience interact should also be considered. It can well be that interviewers with less overall experience will have a more pronounced learning curve as they gather more within-survey interviewer experience compared to more experienced interviewers which start working on a new survey study already with shorter interview durations (Kirchner & Olson, 2017; Kosyakova et al., 2021).

In sum, the differentiation between overall and within-survey interviewer experience may be crucial, and thus call for different adjustments in training, feedback, and supervision.

The Role of Other Factors on Interview Duration

There are other influences on interview duration besides interviewer experience (Kosyakova et al., 2021; Loosveldt & Beullens, 2013b; Vandenplas et al., 2018). An alternative explanation for decreasing interview duration over the field phase is based on the changing composition of the sample over time (Kirchner & Olson, 2017). As the fieldwork progresses, harder-to-reach respondents become more common and the lower cooperativeness of the remaining sample could be what makes the interview duration shorter. These respondents could also have a greater tendency toward satisficing behaviors, leading to shorter interviews (Krosnick, 1991). On the other hand, harder-to-reach respondents could struggle with the answers and this would lead to longer interviews (Jin et al., 2019).

Further respondent characteristics that have been discussed as potentially accounting for interview duration, albeit with inconsistent evidence, are age, education, employment status, and family and work time demands (Loosveldt & Beullens, 2013a; Timbrook et al., 2018; Vandenplas et al., 2018).

These personal characteristics (particularly sex, age, and education), when extended to the interviewer level, have also been considered as explanatory factors for interview duration, both on their own and paired with respondent characteristics (Kirchner & Olson, 2017; Kosyakova et al., 2021; Olson & Peytchev, 2007). Similarly, there is also some inconsistency in the findings regarding their effect, as some research fails to find significant results (Sturgis et al., 2021), while other findings show that older and male interviewers are associated with longer interviews (Timbrook et al., 2018).

Finally, some specific characteristics of the interview itself can impact interview duration and therefore should also be controlled. For example, using a mobile phone is associated with significantly longer interviews when compared to landline connections, because mobile communication tends to be more prone to interruptions and takes longer on average (Timbrook et al., 2018). The type of telephone connection could also confound the association between interviewer experience and interview duration as the interview situation might vary substantially. Additional interview characteristics can have similar impacts such as conducting interviews on a weekday vs. the weekend; having interviews conducted at first call; the time of day; or the number of contact attempts until a successful interview (Kirchner & Olson, 2017; West & Blom, 2017).

Data & Methods

Data

This paper uses data from the National Educational Panel Study (NEPS; see Blossfeld & Roßbach, 2019), Starting Cohort 4 – Grade 9 (doi:10.5157/NEPS:SC4:11.0.0; NEPS Network, 2020). The NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in cooperation with a nationwide network and it is a multi-cohort longitudinal survey designed to find out more about how education is acquired in Germany and how competencies develop over time. Following a multi-informant perspective, the study is not restricted to student data but also includes data from relevant context persons such as parents, teachers, and school heads. For our analyses, we use the parents CATI interviews of wave 1 of a sample of students in grade 9, recruited from 545 randomly selected regular schools in Germany as well as 103 schools for students with special educational needs. The interview is directed to the parent primarily responsible for students' school aspects. From a total child sample of 16,425 cases, 11,097 (68%) parents gave their permission to be contacted for the parent interviews. Going along with the progression of the data collection within schools (and obtaining parents' permission) the addresses for the parent interviews were handed over to the responsible fieldwork agency in three tranches. From that parent sample, 9,180 (83%) CATI sessions were completed during fieldwork between January and July 2011 (for a description of the fieldwork for the parent interview see Aust et al., 2012). Additionally, some parents have more than one child involved in the study and some interviews were conducted in Turkish and Russian. As we are interested in keeping the interview duration as comparable as possible, we exclude incomplete interviews and interviews with parents with more than one child in the NEPS from the sample. Furthermore, as there is evidence of language accounting for shorter interviews (Vandenplas et al., 2018), we kept only those that were conducted in German. After excluding cases who asked for data deletion, the final sample consists of N=8,622 parent interviews (AAPOR Response Rate 1: 0.824; American Association for Public Opinion Research, 2016). The parent interviews were conducted by 180 interviewers. While children that participate in the NEPS do receive incentives, the parents only receive an advance letter with detailed accompanying information by regular mail.

Dependent Variable: Interview Duration

The dependent variable is interview duration. Within this paper, it is operationalized as core interview duration that measures the time in minutes passed from the start until the end of all content-related questions asked to the respondent. The contact module with information concerning the study is excluded as well as the verification of the status of the respondent as the child's legal guardian and main education contact parent. Also excluded are final questions concerning updates of contact data. The interview duration is calculated by using the time stamps which indicate the transition between questionnaire modules and ranges from 12.3 minutes minimum to 100.7 maximum, with a mean of 31.3 minutes and a standard deviation of 8.8. Following previous research (Garbarski et al., 2020; Olson & Smyth, 2015), we apply two transformations to the dependent variable: First, the response times with values below the 1st and higher than the 99th percentile were trimmed and replaced by those percentile values; and second, the variable was log-transformed to correct for skewness.

Explanatory Variables

We model two types of interviewer experience: within-survey interviewer experience and overall interviewer experience. We operationalize within-survey interviewer experience as the count of successful interviews per interviewer in the chronological order (time and date) as registered in the CATI software timestamps. In terms of indicators for interviewer workload we register a range of interviews from 1 to 253, with a mean of 47.9, a standard deviation of 42.5, and a median of 38.5. The within-survey interviewer experience variable was also log-transformed to account for a possible nonlinear learning effect (Kirchner & Olson, 2017; Kosyakova et al., 2021).

Overall interviewer experience was operationalized as the number of years each interviewer has worked for the contracted fieldwork agency. Overall interviewer experience with the given fieldwork agency is aggregated into the categories of below 2 years of employment, 2 to 3, 4 to 5, or more than 5 years of experience. This is notoriously different from previous studies where interviewer experience was operationalized as a dichotomous variable distinguishing between no previous experience and at least 1 year of experience (Kirchner & Olson, 2017; Kosyakova et al., 2021). Furthermore, we include three additional distinct sets of control variables related to the interviewer, respondent, and interview characteristics, detailed in Table 1.

Table 1 Variables included in the analysis

Block	Variables
Main explanatory variables	Within-survey interviewer experience Overall interviewer experience
Interviewer-level controls	Gender Age Education
Respondent-level controls	Gender Age Education Employment status Net equivalent income Household size Type of child’s school
Interview-level controls	Number of contact attempts before a successful interview Interview at the first call Days since advance letter Telephone connection Time of day Day of the week Item nonresponse (%) Number of questions

Gender and age of interviewer and respondent are included as dichotomous variables, while education is operationalized as a three-level categorical variable consisting of lower, intermediate, and higher education. Also at the respondent level, employment status is included as dichotomous variable while income is modeled as net equivalent income (OECD, 2013) distinguishing between three income groups (risk of poverty, average income, high income) using the official national median income threshold for the year 2011 of 1,416 € (Statistische Ämter, 2021). Household size distinguishes between one to three persons, four persons, and more than four persons in the household; the type of school the child attends is divided between "*Gymnasium*" (the school that leads to a university entrance certificate) and other German educational possibilities. Finally, under the interview characteristics block, we further introduce the number of contact attempts before a successful interview; whether the interview was conducted at the first realized telephone contact; the number of days between the posting of the advance letter requesting the parents' participation (controlling for three tranches handed over from the school field) and the interview; the type of telephone connection (landline, mobile, undefined); the time of day and whether the interview was conducted during the week or at the weekend; the percentage of item nonresponse; and the total number of questions answered. As discussed in the previous section, the effect of several of these variables on interview duration is inconsistent throughout the literature. Consequently, we do not elaborate further on our theoretical expectations. The descriptives for all variables under study are given in Table A.1 in the appendix.

Method

The main interest in this paper is to study how interviewer experience is associated with interview duration. Given that each interviewer conducted several interviews, we follow a multilevel modeling strategy where the first level corresponds to respondents and the higher level to interviewers, under a two-level hierarchical linear model with random intercepts framework. The models are estimated using the *R* (R Core Team, 2021) environment and we fit several two-level hierarchical linear models with random intercepts using the package *lme4* (Bates et al., 2015). The model is formulated in the following way:

$$log(Interview\ Duration)_{i,j} = \beta_0 + \beta_1 Var1_i + \beta_2 Var2_j + u_j + \varepsilon_{i,j} \quad (1)$$

In this equation, the subscript formalizes the clustered nature of the data where respondents (i) are nested within interviewers (j). The different explanatory variables are represented by $\boldsymbol{\beta}$, where $\boldsymbol{\beta_1}$ ($\mathbf{Var1_i}$) denotes regression coefficients for respondent-level variables, such as respondent age or gender, and $\boldsymbol{\beta_2}$ ($\boldsymbol{Var2_j}$) for the interviewer-level variables, such as the interviewer experience indicators. The parameter $\boldsymbol{\beta_0}$ reflects the fixed overall effect and $\mathbf{u_j}$ the interviewer random-effects component. Finally, we assume that the individual unobserved heterogeneity is uncorrelated with the explanatory variables, following a normal distribution, and the residual error term is represented by $\boldsymbol{\varepsilon_{i,j}}$.

The model is estimated in a stepwise approach starting with the unconditional or empty model (Model 0), which shows how much of the variance of interview duration is explained by the higher level (interviewer). The introduction of within-survey interviewer experience and overall interviewer experience follows in the next step in Model 1. Each thematic block of control variables is then introduced sequentially (interviewer-level, respondent-level, and interview-level characteristics), respectively Model 2, Model 3, and Model 4. The last model (Model 5), adds the interaction term between within-survey and overall interviewer experience, introduced to test whether within-survey interviewer experience has a differential impact concerning each of the overall interviewer experience categories (<2 years, 2–3 years, 4–5 years, >5 years). In each of the steps, we will look closely at the intraclass correlation coefficient (ICC) as a measure indicating the variance due to the interviewer.

There was a total of less than 1% missing data, with no missing observations on the dependent variable (Figure S.1, online supplementary material). Most affected by item nonresponse is the income variable. Aiming to minimize bias due to missing data, we used the package *missForest* for multiple imputation (Stekhoven, 2022). The *missForest* is a nonparametric method of imputation in which the algorithm used is an iterative process that assigns initial values to the missing data, fits a random forest for each variable based on the observed values predicting new imputed observations until convergence (Stekhoven & Bühlmann, 2012). Also as a robustness check, Table S.3 (online supplementary material) replicated the main model without the imputation procedure, considering only the cases with complete information. The results are very close to the main model, indicating therefore that the imputation process is unlikely to be driving our main results.

Results

How is interview duration related to interviewer experience? Before the multilevel model results, Figure 1 shows a descriptive analysis comparing the mean interview duration as the fieldwork progressed, by overall interviewer experience and the number of interviews.

Altogether, the mean interview duration for more overall experienced interviewers (>5 years) is 30.5 minutes and 32.9 minutes for less experienced interviewers (<2 years). Interviewers with two to three years of experience (2–3 years) have a mean of 32.1 minutes while in the remaining overall experience category (4–5 years) the average is 31.6 minutes. These differences indicate that more experienced interviewers are generally faster than the less experienced ones; a tendency also present in Figure 1, which also considers, in chronological order, the increasing number of interviews by a given interviewer within that specific survey. It should be noted, though, that in the highest within-survey interviewer experience category ("+200"), only the more overall experienced interviewers are included because there are no cases in the lower experience categories. Nevertheless, these initial descriptive findings indicate that, in disagreement with previous research on telephone surveys (Kirchner & Olson, 2017), both types of interviewer experience are likely associated with a tendency towards shorter interview durations. Next, we take the analysis forward and consider other factors that can affect the relationship of the experience variables and interview duration in the NEPS parent interviews.

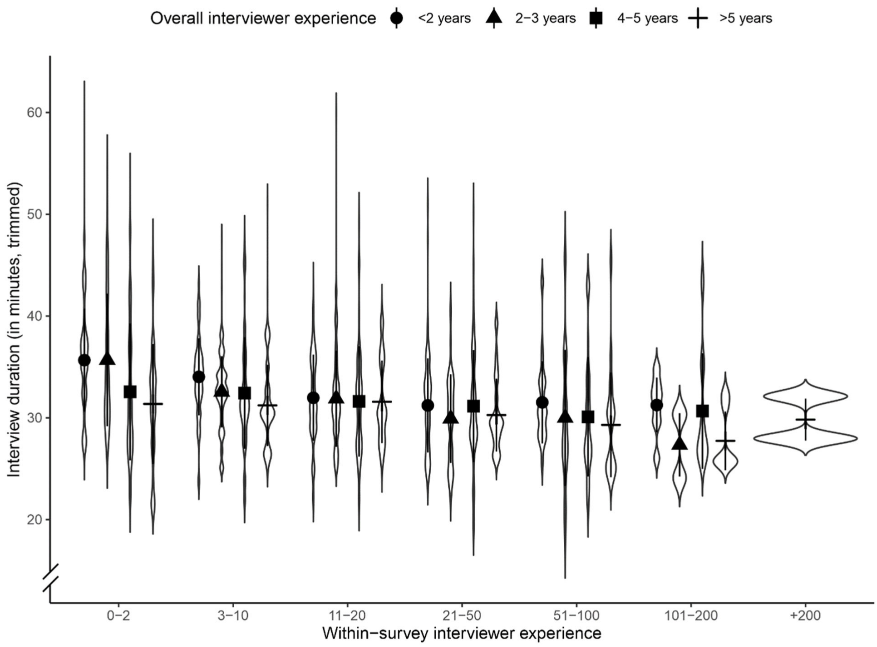


Figure 1 Average interview duration distribution by within-survey interviewer experience intervals and mean overall interviewer experience

Table 2 presents the coefficients and respective standard errors, variance components, and model comparison statistics for all the estimated multilevel regression models for interview duration. For reasons of clarity, we only present the estimates for the main explanatory variables. The complete table, including the estimates for the control variables, can be consulted in Table A.2 in the appendix.

Table 2 Hierarchical linear random intercept models: estimated coefficients, standard errors, and variance components (reduced to main explanatory variables and interaction)

	Model 0	Model 1	Model 2	Model 3	Model 4	Model 5
	Null	+	+	+	+	+
		Experience	Interviewer	Respondent	Interview	Interaction
Main explanatory variables						
Within-survey interviewer experience (log)		-0.029***	-0.030***	-0.031***	-0.043***	-0.046***
		(0.002)	(0.002)	(0.002)	(0.002)	(0.004)
Overall interviewer experience (ref. < 2 years)						
2–3 years		-0.029	-0.047	-0.045	-0.030	-0.003
		(0.027)	(0.026)	(0.026)	(0.022)	(0.027)
4–5 years		-0.023	-0.053	-0.052*	-0.032	-0.049
		(0.027)	(0.026)*	(0.026)	(0.022)	(0.027)
> 5 years		-0.031	-0.083***	-0.081**	-0.070**	-0.122***
		(0.032)	(0.032)	(0.032)	(0.027)	(0.032)
Interaction						
Within-survey * 2–3 years experience						-0.011
						(0.006)
Within-survey * 4–5 years experience						0.006
						(0.005)
Within-survey* > 5 years experience						0.018**
						(0.005)

Table 2 (continued)

	Model 0 Null	Model 1 + Experience	Model 2 + Interviewer	Model 3 + Respondent	Model 4 + Interview	Model 5 + Interaction
Intercept	3.466*** (0.011)	3.564*** (0.021)	3.539*** (0.031)	3.575*** (0.032)	3.467*** (0.029)	3.477*** (0.030)
Residual variance (interviewer)	0.020	0.017	0.015	0.014	0.010	0.010
Residual variance (respondent)	0.047	0.046	0.046	0.043	0.032	0.032
ICC	0.294	0.264	0.238	0.245	0.240	0.241
Marginal R^2	-	0.020	0.050	0.103	0.319	0.318
Conditional R	0.294	0.276	0.275	0.323	0.482	0.482
LogLikelihood	703.09	770.51	783.57	1,091.53	2,377.67	2,388.44
Pr (>Chisq)		0.000***	0.000***	0.000***	0.000***	0.000***
Sample size (respondents/interviewers)	8,622/180					

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Effects at the interviewer-level, respondent-level, and interview-level controls are omitted; for complete results see appendix, Table A.2.

The ICC for the null model shows that interviewers account for 29% of the variance of interview duration. While this proportion is higher than typically observed for substantive variables in telephone surveys, it aligns closely with previous studies examining face-to-face interviews (e.g. Olson & Peytchev, 2007; Loosveldt & Beullens, 2013a; Kosyakova et al., 2021). This indicates that the interviewer grouping variable significantly affects the mean interview duration. As we add more variables to the model, the null model is taken as the baseline.

The next step is to introduce the interviewer experience variables: within-survey and overall interviewer experience (Model 1). Introducing these variables results in an improvement of model fit, as indicated by the likelihood χ^2 test, and the ICC is reduced to 26%. The effect of within-survey interviewer experience itself is negative and significant, thus giving support to the argument that study-specific experience explains the reduction of interview duration. On the other hand, Model 1 shows no significant effect of overall interviewer experience on interview duration. However, this result changes as more explanatory variables are included in the model.

The next steps introduce the respective blocks of control variables: Model 2—when including the interviewer-level controls—shows that while within-survey interviewer experience still has a negative significant effect on interview duration, the impact of overall interviewer experience categories is negative and statistically significant (at 5% level) on the highest experience categories (4–5 years and >5 years). The remaining blocks of controls for respondent and interview characteristics are sequentially introduced in Model 3 and Model 4. Every time, model fit improved significantly, but the ICC is only reduced slightly to 24%.

Figure 2 shows the predicted conditional effect of within-survey interviewer experience by overall interviewer experience on interview duration. Less overall experienced interviewers start with higher interview durations and this difference holds throughout the fieldwork.

This indicates that interviewers with more than 2 years of experience on average start the fieldwork with a shorter interview duration and are consistently faster over the whole field phase. Furthermore, Model 4 shows that within-survey interviewer experience has a statistically significant negative effect on interview duration, even after including all control variables in the model. On the other hand, Model 4 also hints to a more nuanced interpretation regarding the effect of overall interviewer experience. In Table 2, we can see that after all blocks of controls are introduced, only the most overall experience category (>5 years) is still statistically associated with shorter interviews. While in other experience categories the significant negative effect on duration is explained away, only in the more experienced category of interviewers we see the persistence of the negative effect on the dependent variable. As for within-survey interviewer experience, the tendency of interviewers reducing interview duration as they conduct more interviewers is clear.

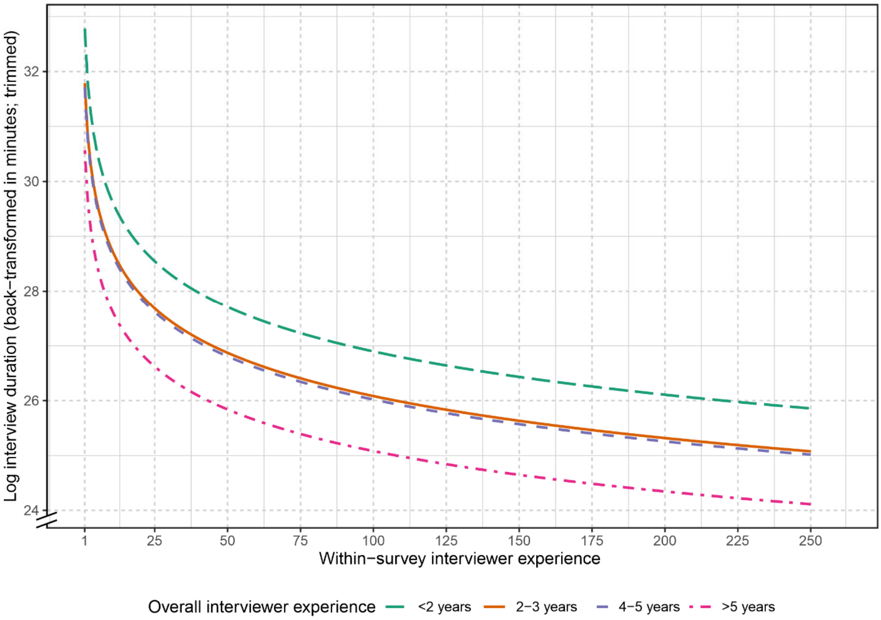


Figure 2 Predicted conditional interview duration by within-survey interviewer experience and overall interviewer experience (based on Model 4)

We checked the effect of the control variables (see Model 4 with all controls in Table A.2 in the appendix): Older interviewers conduct longer interviews but there is no effect of gender and education on the interviewers' side. On the respondent level, those under risk of poverty in terms of net equivalent income have also significantly longer interviews than parents in the intermediate category of net equivalent income. Female respondents and households with more or fewer than four members also take significantly less time to be interviewed; respondent age and educational level other than intermediate go along with longer interviews. Moreover, being unemployed is not a significant predictor of interview duration. In contrast, the child enrolled in a "*Gymnasium*" is a significant predictor of a shorter interview. As the NEPS is dedicated to studying the German educational system, this result is particularly relevant as it is consistent across models and even after controlling for the number of questions asked (Model 4). Regarding the effects of interview characteristics, we see that a higher number of contact attempts before a successful interview is associated with longer interviews which could be due to the characteristics of harder-to-reach interviewees. Furthermore, respondents with higher rates of item nonresponse also have longer interviews. Another interesting result is that interviews conducted at the first realized contact are faster. It seems that if the respondent agrees to answer the survey immediately, the time used to complete the questionnaire is significantly shorter. Having the interview on the weekend is not significantly associated with interview duration, whereas interviews in the afternoon are shorter than those in the morning. In contrast, the number of days since the sending of the advance letter to the respondents, with the request to participate in the survey, is significant albeit with a very close to zero effect on interview duration. Also, as discussed in the literature, we find that respondents who use a mobile phone connection take significantly longer than respondents who use a landline. Finally, the number of questions is positively associated with interview duration.

If shorter interview durations are attributed to within-survey interviewer experience, estimating an interaction between within-survey and overall interviewer experience lets us examine if conducting more interviews within one survey affects interview duration differently for interviewers with more or less overall professional experience. Including the interaction (Model 5) between both interviewer experience measures improves model fit significantly, as shown in Table 2. Figure 3 shows the predicted conditional effect on interview duration by different levels of the main explanatory variables within-survey interviewer experience and overall interviewer experience. While there are no significant differences across overall experience categories in the first interviews, this eventually changes. When examining Model 5 (Figure 3), which allows the within-survey coefficient to differ across overall interviewer experience–groups, we observe nuances suggesting varied learning trajectories across experience brackets. Figure 3 reveals that from the start, the “2–3 years” overall interviewer experience–group has a steeper decrease in interview duration than the “>5 years” experience group. By the 50th interview, the two durations intersect. This observation aligns with Table 2, where the “<2 years”, “2–3 years”, and “4–5 years” groups show a sharper decline in duration compared to the “>5 years” group. Keeping all other variables constant, by the 50th interview the interviewers belonging to the overall experience category of “2–3 years” and “>5 years” take on average approximately 26.2 minutes to conduct an interview. This means that by the fiftieth interview, on average, the “2–3 years” overall experience group take less 6.6 minutes than their first interview while the “>5 years” overall experience group take 3.0 minutes less than their first interview.

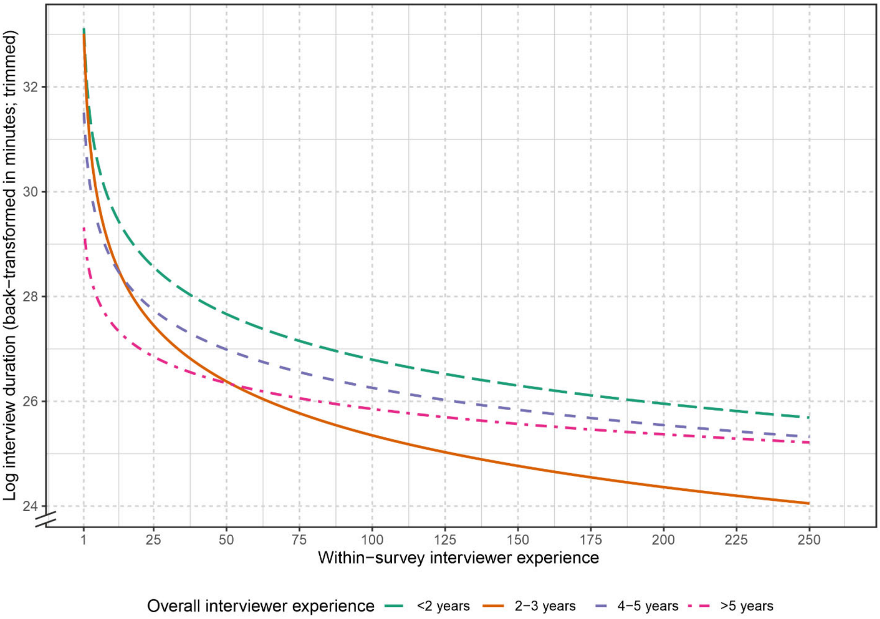


Figure 3 Predicted conditional interview duration by within-survey interviewer experience and overall interviewer experience (based on Model 5)

The robustness of all these results was tested by estimating alternative model specifications. The results of these estimations can be found in the online supplementary material. First, our main model was replicated without trimming the dependent variable at the 1st and 99th percentile, without log transforming the dependent variable and, finally, without imputation of the missing variables. The results are very similar and can be found respectively in Table S.1, Table S.2 and Table S.3 (online supplementary material). Second, as discussed by previous research, an alternative explanation to survey experience driving the tendency for interviews to become shorter as the fieldwork develops is related to socio-demographic changes within the respondent sample (Kirchner & Olson, 2017). Throughout the field period, the characteristics of the respondents may change, making it more likely for interviews to take longer because of older and less educated respondents, for example. Even though we control for several respondent characteristics, in order to rule out a “compositional aspect” effect over the field period (Kirchner & Olson, 2017, p. 86), we divide the respondents of the first (and by sample size largest) tranche into three different samples of early, middle, and late respondents and compared socio-demographic characteristics. This effort showed some differences between early, middle, and late responders namely in terms of respondent age, employment status, household size, and the type of school attended by the child (Table S.4, online supplementary material). However, our main results have shown that the survey experience effect is robust even with respondent-level socio-demographic characteristics in the model. Following this line of thought, Table S.5 (online supplementary material) repeats the main model for the first tranche of 5,975 respondents only. The estimated coefficients for the first tranche are very similar to the main model results. Thus, it is not likely that the effect of interviewer experience on interview duration is being driven by the socio-demographic composition of the sample.

Overall, the proportion of variance explained by the interviewers varies between 29% in the null model (Model 0) and 24% for the complete model (Model 4). The introduction of within-survey and overall interviewer experience, as well as further control variables, did impact the ICC, but only reducing it by 5 percentage points. Given that several other potential confounders are included in the model, this indicates that interviewers have a large impact on how long the survey interview lasts.

Discussion

This paper aimed to investigate how interviewer experience impacts interview duration in a CATI-based large-scale panel study. First, as in previous research on face-to-face interviewing, our results show that interviewers are an important source of variation for interview duration also in telephone surveys, expanding the available empirical evidence to other modes of data collection. The variance explained by the interviewer level in this study is large and slightly higher when compared to other telephone surveys (Kirchner & Olson, 2017). Second, our results give further support to the findings that within-survey interviewer experience impacts interview duration (Kirchner & Olson, 2017; Kosyakova et al., 2021). This effect is stable and robust to the introduction of control variables for interviewer, respondent, and interview characteristics. On the other hand, contrasting previous findings using CATI survey data, we find that overall interviewer experience does have a significant negative impact on interview duration but only for the more experienced interviewers. This indicates that the effect of overall interviewer experience on time duration is not really continuous as interviewers gain experience. Instead, it appears more likely that interviewers working for more than 5 years (>5 years) in the survey fieldwork agency conduct interviews faster.

Third, the effect of within-survey experience on interview duration differs between categories of overall interviewer experience. While the difference becomes evident as interviews progress during fieldwork, interviewers with up to five years of experience tend to speed up at a faster rate than those with an experience of more than five years. It is particularly telling that the same effect was not found for the more inexperienced interviewers. For interviewers, it appears to be necessary to have some previous experience and knowledge to change their conduct in order to achieve shorter interview duration.

Furthermore, we also find some of the controls with important effects. Namely, the demographics of the interviewer, characteristics of the respondent, their socio-economic conditions, and the child's school situation as well as several interview characteristics impact interview duration. Most notably, parents whose child attends a "*Gymnasium*" have a shorter interview duration compared to children from other school types. A possible explanation for this result could be that the main aim of the parent interview is to talk about their children and, in the German context, "*Gymnasium*" children have a somewhat more streamlined and easier to explain educational trajectory. On side of the interview characteristics, the picture is a bit more mixed. Conforming with previous findings, a higher item non-response rate is associated with a longer interview duration, suggesting that interviewers might invest additional time to evoke an answer from the respondent—and not quickly accept a nonresponse and jump to the next question. This positive but also partly counter-intuitive result has also been found in other studies (Kirchner & Olson, 2017).

Our analyses go along with a set of methodological limitations: (1) Whereas the measurement of within-survey interviewer experience is automatically recorded within the interviews and available in a fine-grained manner, the result concerning overall interviewer experience should be seen with caution as the inexperienced category is below 2 years of experience. It can be argued that this interviewer overall experience category is a measure too blunt to distinguish between experienced and not experienced interviewers. (2) Another limitation of this study is that we were not able to include the characteristics of the questions as item-based timestamps are not available. Previous research has shown that response times are also related to question type, question length, response format, presence of instructions, or the labeling of the response scale (Garbarski et al., 2020; Olson, Smyth, & Kirchner, 2020). (3) Our study can also be considered limited due to the lack of information regarding interviewer behavior and the interaction between the respondent and the interviewer. While we uncovered some patterns about how interviewers and respondents interact, we are still some distance away from unveiling the actual dynamics in each interview. More measures of this adaptive relationship between interviewers and respondents are necessary to link more closely how both of these agents' behaviors differentially impact interview duration. On a final note, while we look at the percentage of item missings, interview duration is an indirect measure and any further steps should include additional indicators of interviewer performance and data quality.

Nonetheless, the results of this paper can be used for optimizing interviewer training and supervision as well as for more adequate cost-forecasting within large-scale panel studies.

An early transition to shorter interviews due to learning effects and routine with the instrument would be desirable concerning survey costs. This could also impact the forecast of cost-aspects as CATI interview time in the NEPS is billed by the minute. Extension of practical sessions could be introduced in interviewer training and this way reduce interview duration earlier, even though that might not always be desirable. Nevertheless, more detailed research is needed to distinguish whether the decline in interview duration is a general learning effect or due to special interviewer behaviors, such as deviating from the standard protocols or reducing unnecessary conversations during the first interviews.

References

Ackermann-Piek, D., & Massing, N. (2014). Interviewer Behavior and Interviewer Characteristics in PIAAC Germany. *Methods, Data, Analyses, 8*(2), 199–222. https://doi.org/10.12758/MDA.2014.008

American Association for Public Opinion Research. (2016). *Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys*. American Association for Public Opinion Research (AAPOR). https://aapor.org/wp-content/uploads/2022/11/Standard-Definitions20169theditionfinal.pdf

Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V., Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e. V., BVM Berufsverband Deutscher Markt- und Sozialforscher e. V., & Deutsche Gesellschaft für Online-Forschung e. V. (2021). *Richtlinie für telefonische Befragungen*. https://www.adm-ev.de/wp-content/uploads/2021/01/RL-Telefon-neu-2021.pdf

Aust, F., Hess, D., & Prussog-Wagner, A. (2012). *Methodenbericht. NEPS. Startkohorte 4 (Elternbefragung). Haupterhebung Frühjahr 2011 B34*. Bonn. infas Institut für angewandte Sozialwissenschaft. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/Methodenbericht_B34.pdf

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Blossfeld, H.-P., & Roßbach, H. G. (Eds.). (2019). *Education as a Lifelong Process. The German National Educational Panel Study (NEPS)* (Second Edition). Springer VS.

Böhme, M., & Stöhr, T. (2014). Household Interview Duration Analysis in CAPI Survey Management. *Field Methods, 26*(4), 390–405. https://doi.org/10.1177/1525822X14528450

Fowler, F. J., Jr., & Mangione, T. W. (1990). *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. SAGE.

Garbarski, D., Dykema, J., Schaeffer, N. C., & Edwards, D. F. (2020). Response Times as an Indicator of Data Quality: Associations with Question, Interviewer, and Respondent Characteristics in a Health Survey of Diverse Respondents. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, & B. T. West (Eds.), *Interviewer Effects from a Total Survey Error Perspective* (pp. 253–264). CRC Press.

Jin, J., Vandenplas, C., & Loosveldt, G. (2019). The Evaluation of Statistical Process Control Methods to Monitor Interview Duration During Survey Data Collection. *SAGE Open*, *9*(2), 1-14. https://doi.org/10.1177/2158244019854652

Kirchner, A., & Olson, K. (2017). Examining Changes of Interview Length over the Course of the Field Period. *Journal of Survey Statistics and Methodology*, *5*(1), 84–108. https://doi.org/10.1093/jssam/smw031

Kosyakova, Y., Olbrich, L., Sakshaug, J. W., & Schwanhäuser, S. (2021). Positive learning or deviant interviewing? Mechanisms of experience on interviewer behavior. *Journal of Survey Statistics and Methodology*, 1–27. https://doi.org/10.1093/jssam/smab003

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Lepkowski, J., Axinn, W., Kirgis, N., West, B. T., Kruger Ndiaye, S., Mosher, W., & Groves, R. (2010). *Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection* (NSFG Survey Methodology 10-012). Population Studies Center, Institute for Social Research, University of Michigan.

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. https://doi.org/10.18148/srm/2013.v7i2.5086

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. https://doi.org/10.1016/j.ssresearch.2013.06.005

NEPS Network. (2020). *Scientific Use File of Starting Cohort Grade 9 (Version 11.0.0)* [Computer software]. Leibniz Institute for Educational Trajectories (LIfBi). Bamberg, Germany. https://doi.org/10.5157/NEPS:SC4:11.0.0

OECD. (2013). *OECD Framework of statistics on the distribution of household income, consumption and wealth*. OECD Publishing. https://doi.org/10.1787/9789264194830-en

Olson, K., & Peytchev, A. (2007). Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes. *Public Opinion Quarterly*, *71*(2), 273–286. https://doi.org/10.1093/poq/nfm007

Olson, K., & Smyth, J. D. (2015). The Effect of CATI Questions, Respondents, and Interviewers on Response Time. *Journal of Survey Statistics and Methodology*, *3*(3), 361–396. https://doi.org/10.1093/jssam/smv021

Olson, K., & Smyth, J. D. (2020). What Do Interviewers Learn? Changes in Interview Length and Interviewer Behaviors over the Field Period. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, & B. T. West (Eds.), *Interviewer Effects from a Total Survey Error Perspective* (pp. 279–290). CRC Press.

Olson, K., Smyth, J. D., Dykema, J., Holbrook, A. L., Kreuter, F., & West, B. T. (Eds.). (2020a). *Interviewer Effects from a Total Survey Error Perspective*. CRC Press.

Olson, K., Smyth, J. D., Dykema, J., Holbrook, A. L., Kreuter, F., & West, B. T. (2020b). The Past, Present, and Future of Research on Interviewer Effects. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, & B. T. West (Eds.), *Interviewer Effects from a Total Survey Error Perspective* (pp. 3–15). CRC Press.

Olson, K., Smyth, J. D., & Kirchner, A. (2020). The Effect of Question Characteristics on Question Reading Behaviors in Telephone Surveys. *Journal of Survey Statistics and Methodology*, *8*(4), 636–666. https://doi.org/10.1093/jssam/smz031

R Core Team. (2021). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Statistische Ämter. (2021). *Armutsgefährdung und Einkommensverteilung (MZ-Kern).* Statistische Ämter des Bundes und des Landes. Gemeinsames Statistikportal. https://www.statistikportal.de/de/sbe/ergebnisse/einkommensarmut-und-verteilung

Stekhoven, D. J. (2022). *missForest: Nonparametric Missing Value Imputation using Random Forest* (Version R package version 1.5) [Computer software].

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Stiegler, A., & Biedinger, N. (2015). *Interviewer Qualifikation und Training* (GESIS Survey Guidelines). GESIS - Leibniz-Institut für Sozialwissenschaften. https://www.gesis.org/gesis-survey-guidelines/operations/interviewertraining-und-effekte/interviewertraining https://doi.org/10.15465/GESIS-SG_013

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. https://doi.org/10.1093/jssam/smaa009

Timbrook, J., Olson, K., & Smyth, J. D. (2018). Why Do Cell Phone Interviews Last Longer? A Behavior Coding Perspective. *Public Opinion Quarterly, 82*(3), 553–582. https://doi.org/10.1093/poq/nfy022

Vandenplas, C., Beullens, K., & Loosveldt, G. (2019). Linking interview speed and interviewer effects on target variables in face-to-face surveys. *Survey Research Methods, 13*(3). https://doi.org/10.18148/SRM/2019.V13I3.7321

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*(4), 516–538. https://doi.org/10.1093/jssam/smx034

West, B. T., & Blom, A. G. (2017). Explaining Interviewer Effects: A Research Synthesis. *Journal of Survey Statistics and Methodology, 5*(2), 175–211. https://doi.org/10.1093/jssam/smw024

Appendix

Table A.1 Descriptive statistics

Variable	N	Mean (SD) / Proportion %	Min	Max
Dependent variable				
Interview duration	8,622	31.27 (8.41)	16.5	60.8
Main explanatory variables				
Within-survey interviewer experience	8,622	43.15 (39.74)	1	253
Overall interviewer experience	180			
< 2 years	49	27.2%		
2–3 years	50	27.8%		
3–4 years	53	29.4%		
> 5 years	28	15.6%		
Interviewer-level controls				
Gender	180			
Male	92	51.1%		
Female	88	48.8%		
Age	180			
< 30	62	34.4%		
30–49	62	34.4%		
50–65	45	25%		
> 65	11	6.1%		
Education	175			
Lower	38	21.1%		
Intermediate	32	17.8%		
Higher	105	58.3%		
Respondent-level controls				
Gender	8,622			
Male	1,451	16.8%		
Female	7,171	83.2%		
Age	8,622	45.68 (5.16)	25	92
Education	8,608			
Lower	762	8.9%		
Intermediate	4,936	57.3%		
Higher	2,910	33.8%		
Employment status	8,615			
Employed	7,265	84.3%		
Unemployed	1,350	15.7%		

Table A.1 (continued)

Variable	N	Mean (SD) / Proportion %	Min	Max
Net equivalent income	7,128			
Risk of poverty	2,153	30.2%		
Average income	4,316	64.8%		
High income	359	5.0%		
Household size	8,620			
1–3 persons	2,593	30.1%		
4 persons	3,772	43.8%		
> 4 persons	2,255	26.2%		
Type of child’s school	8,622			
Other school	5,196	60.3%		
Gymnasium	3,426	39.7%		
Interview-level controls				
Number of contact attempts	8,622	5.84 (8.01)	1	100
Interview at the first call	8,622			
First call	1,453	16.9%		
Not first call	7,169	83.1%		
Days since advance letter	8,622	47.89 (24.32)	7	165
Telephone connection	8,622			
Landline	5,813	67.4%		
Mobile phone	435	5.0%		
Undefined	2,374	27.5%		
Time of day	8,622			
Morning	1,956	22.7%		
Afternoon	4,151	48.1%		
Evening	2,515	29.2%		
Day of the week	8,622			
Weekday	6,885	79.9%		
Weekend	1,737	20.1%		
Item nonresponse	8,622	0.77 (1.02)	0	11.5
Number of questions	8,622	254.17 (30.72)	165	327

Table A.2 Hierarchical linear random intercept models: estimated coefficients, standard errors, and variance components (including controls)

	Model 0	Model 1	Model 2	Model 3	Model 4	Model 5
	Null	+ Experience	+ Interviewer	+ Respondent	+ Interview	+ Interaction
Main explanatory variables						
Within-survey interviewer experience (log)		-0.029*** (0.002)	-0.030*** (0.002)	-0.031*** (0.002)	-0.043*** (0.002)	-0.046*** (0.004)
Overall interviewer experience [ref. < 2 years]						
2–3 years		-0.029 (0.027)	-0.047 (0.026)	-0.045 (0.026)	-0.030 (0.022)	-0.003 (0.027)
4–5 years		-0.023 (0.027)	-0.053 (0.026)*	-0.052* (0.026)	-0.032 (0.022)	-0.049 (0.027)
5 years		-0.031 (0.032)	-0.083*** (0.032)	-0.081** (0.032)	-0.070** (0.027)	-0.122*** (0.032)
Interviewer-level controls						
Gender [female]			-0.009 (0.019)	-0.014 (0.019)	-0.015 (0.016)	-0.017 (0.016)
Age [ref. < 30 years old]						
30–49			0.087*** (0.024)	0.080*** (0.023)	0.064*** (0.020)	0.067*** (0.020)
50–65			0.120*** (0.026)	0.120*** (0.026)	0.106*** (0.022)	0.105*** (0.022)
> 65			0.157*** (0.045)	0.153*** (0.044)	0.147*** (0.038)	0.149*** (0.038)
Education [ref. intermediate]						
Lower education			-0.0214 (0.031)	-0.014 (0.030)	0.006 (0.026)	0.002 (0.026)
Higher education			-0.018 (0.026)	-0.012 (0.026)	-0.009 (0.022)	-0.010 (0.022)

Table A.2 (continued)

	Model 0 Null	Model 1 + Experience	Model 2 + Interviewer	Model 3 + Respondent	Model 4 + Interview	Model 5 + Interaction
Respondent-level controls						
Gender [female]				-0.045*** (0.006)	-0.016** (0.005)	-0.016** (0.005)
Age (centered)				0.001** (0.000)	0.002*** (0.000)	0.002*** (0.000)
Education [ref. intermediate]						
Lower education				0.141*** (0.008)	0.118*** (0.007)	0.118*** (0.007)
Higher education				0.010* (0.005)	0.126** (0.004)	0.126** (0.004)
Employment status [employed]				0.005 (0.006)	0.000 (0.005)	0.000 (0.005)
Net equivalent income [ref. average]						
Risk of poverty				0.014* (0.005)	0.049*** (0.005)	0.049*** (0.005)
High income				0.017 (0.011)	-0.006 (0.010)	-0.006 (0.010)
Household size [ref. 4 persons]						
1–3 persons				-0.028*** (0.005)	-0.010* (0.004)	-0.010* (0.004)
> 4 persons				0.022*** (0.005)	-0.010* (0.005)	-0.010* (0.005)
Type of child’s school [“*Gymnasium*”]				-0.051*** (0.005)	-0.024*** (0.004)	-0.024*** (0.004)
Interview-level controls						

Table A.2 (continued)

	Model 0 Null	Model 1 + Experience	Model 2 + Interviewer	Model 3 + Respondent	Model 4 + Interview	Model 5 + Interaction
Number of contact attempts (centered)					0.008***	0.008***
					(0.000)	(0.000)
Interview at the first call [not first call]					0.066***	0.066***
					(0.005)	(0.005)
Days since advance letter (centered)					-0.000***	-0.000***
					(0.000)	(0.000)
Telephone connection [ref. landline]						
Mobile phone					0.045***	0.044***
					(0.009)	(0.009)
Undefined					0.023***	0.023***
					(0.004)	(0.004)
Time of day [ref. morning]						
Afternoon					-0.026***	-0.025***
					(0.005)	(0.005)
Evening					-0.006	-0.005
					(0.006)	(0.006)
Day of the week [weekday]					-0.001	-0.002
					(0.005)	(0.005)
Item nonresponse (%)					0.029***	0.029***
					(0.002)	(0.002)
Number of questions (centered)					0.027***	0.027***
					(0.000)	(0.000)
Within-survey * overall interviewer experience						
Within-survey * 2–3 years						-0.011
						(0.006)

Table A.2 (continued)

	Model 0 Null	Model 1 + Experience	Model 2 + Interviewer	Model 3 + Respondent	Model 4 + Interview	Model 5 + Interaction
Within-survey * 4–5 years						0.006 (0.005)
Within-survey * > 5 years						0.018** (0.006)
Intercept	3.466*** (0.011)	3.564*** (0.021)	3.539*** (0.031)	3.575*** (0.032)	3.467*** (0.029)	3.477*** (0.030)
Residual variance (interviewer)	0.020	0.017	0.015	0.014	0.010	0.010
Residual variance (respondent)	0.047	0.046	0.046	0.043	0.032	0.032
ICC	0.294	0.264	0.238	0.245	0.240	0.241
Marginal R^2	-	0.020	0.050	0.103	0.319	0.318
Conditional R	0.294	0.276	0.275	0.323	0.482	0.482
LogLikelihood	703.09	770.51	783.57	1,091.53	2,377.67	2,388.44
Pr (>Chisq)		0.000***	0.000***	0.000***	0.000***	0.000***
Sample size (respondents/interviewers)	8,622/180					

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

methods, data, analyses | Vol. 18(2), 2024, pp. 213-248 DOI: 10.12758/mda.2024.04

Measurement Invariance and Quality of Attitudes Towards Immigration in the European Social Survey

Amelie Nickel[1] & Wiebke Weber[2]

[1] *Bielefeld University, IKG, Germany*

[2] *LMU Munich, Germany & RECSM, Universitat Pompeu Fabra, Spain*

Abstract

The number of studies assessing measurement invariance of the European Social Survey's (ESS) immigration scale increased in recent years. However, the comparability of findings is limited due to the lack of consistency in the analytic strategies and methods employed across these studies. The present study aims to address this issue by employing a consistent approach: a multigroup confirmatory factor analysis (MGCFA), to test for measurement invariance of attitudes towards immigration in each of the first nine rounds of the ESS. Moreover, we estimate the measurement quality by computing the reliability coefficient Omega in each country in each round of the ESS.

Our results reveal that metric invariance holds for all countries but one (Finland) in all rounds, indicating that covariances and regression coefficients can be compared meaningfully. While scalar invariance only holds for different subgroups of countries within each round, partial invariance is fulfilled in all countries, meaning that at least one indicator is equal for all countries allowing for latent mean comparisons. Furthermore, assessing the measurement quality, we find the attitudes towards immigration index similarly good across the different countries and rounds.

Keywords: Measurement invariance, attitudes towards immigration, European Social Survey, MGCFA, Measurement quality

Undoubtedly, the topic of migration will largely shape the national and international political agenda of the 21st century. In election and public debates, the question of how to deal with migration is one of the most pressing concern, effectively capitalized on by the political right. The so-called ‘refugee crisis’ in 2015 has deepened cleavages within the European Union, providing an opportunity for populist radical right parties to advocate for more restrictive policies and shift the overall political discourse to the right (Mudde, 2007, 2020).

Given its ongoing social and political relevance, understanding and analyzing attitudes toward immigration has emerged as one of the most extensively studied aspects of the social sciences (Bohman, 2015; Borgonovi & Pokropek, 2019; Quillian, 1995; Scheepers et al., 2002; Weldon, 2006). This has resulted in extensive literature from various disciplines, such as sociology, psychology, political science, and economics. So far, empirical studies have mainly focused on the individual level, but with the increasing availability of large cross-national datasets, the amount of international comparative research is rising (Meuleman & Billiet, 2012).

Measuring psychological constructs such as values or attitudes across countries raises methodological questions on the comparability of measurements that are often insufficiently or not at all addressed by researchers (Davidov & Meuleman, 2012; Meitinger et al., 2020; Roots et al., 2016). As question wording and items can have different meanings in different countries depending on the linguistic and cultural background, it is essential to verify and ensure that the used measurements are comparable across the observed countries (Roots et al., 2016). According to Meuleman et al. (2022, p.3), the basic idea behind so-called measurement invariance testing (also referred to as measurement equivalence) of multi-item instruments in cross-cultural research is that “when we compare any measurement across groups, that comparison should reflect true differences rather than measurement differences.”

The lack of testing measurement comparability is increasingly criticized in the literature as it may lead to misinterpretation of findings (Meuleman & Billiet, 2012). However, due to improved and new statistical techniques, measurement invariance testing has become more accepted in applied social science research over the last decade (Davidov, Muthen, & Schmidt, 2018; Leitgöb et al., 2023).

Acknowledgements

We extend our sincere gratitude to Carlos Poses for his contribution to the analysis of the paper, and to the anonymous reviewers who provided valuable feedback, which has significantly enhanced the quality of our work.

Direct correspondence to

Amelie Nickel, Bielefeld University, IKG, Germany
E-mail: amelie.nickel@uni-bielefeld.de

When testing measurement invariance we can at first decide between two traditions in measurement theory: Item-response theory (IRT) or structural equation modeling (SEM) (Bauer et al., 2006; Putnick & Bornstein, 2016; Tsaousis et al., 2020)[1].

IRT examines the relationship between an individual's latent trait (e.g., an attitude) and their response to a specific item. In the IRT tradition, measurement invariance is assessed through the lens of differential item functioning (DIF; Holland & Wainer, 2015) across groups, which determines whether item behavior measures equivalent levels of the latent trait across members of different groups (Tsaousis et al., 2020).

We focus on SEM approaches, namely confirmatory factor analysis (CFA) and multi-group confirmatory factor analysis (MGCFA), in which the relations between observed variables and latent construct(s) are tested for measurement invariance between groups (Vandenberg & Lance, 2000). CFA is a statistical technique used to test the invariance of measurement model parameters within subpopulations, while MGCFA is an extended version of CFA that allows invariance testing across multiple groups. MGCFA (Jöreskog, 1971; Millsap, 2011) is most widely used for testing measurement invariance. However, the scientific community is inconsistent about the correct methods as well as the usefulness of measurement invariance testing in general, as a recent debate in *Sociological Methods & Research* shows (Fischer et al., 2022; Meuleman et al., 2022; Welzel et al., 2021; Welzel et al., 2022). In the concluding section of our paper, we outline the advantages and main limitations of MGCFA and highlight some recently developed alternative methods.

In our study, we aim to contribute to the field by employing multigroup confirmatory factor analysis with local fit testing to assess measurement invariance of attitudes towards immigration as measured in each round of the ESS (ESS R1 (2002) to ESS R9 (2018))[2].

In order to make meaningful cross-country comparisons, it is essential to check not only that the measures are comparable across countries, but also that the quality of the measures is comparable. Testing measurement quality is an imperative to correct for measurement errors (Pirralha & Weber, 2020; Poses et al., 2021; Saris & Revilla, 2016). For meaningful comparisons (e.g. correlations), it is crucial that the size of the measurement errors is similar between the groups (e.g. countries) being compared. In general, "the lower the quality of measurement, the more careful researchers need to be in their conclusions [...],

[1] For recent efforts to combine the two approaches, see Raju et al. (2002); Reise et al. (1993); Stark et al. (2006); Widaman and Grimm (2014) quoted from Putnick and Bornstein (2016).

[2] It was not possible to include the most recent round from the European Social Survey (round 10, conducted in 2020) due to the timeframe of the study. In addition, the data collection for round 10 of the ESS took place during the COVID-19 pandemic, which implied unique circumstances such as online interviews and self-completion of questionnaires.

since higher levels of measurement errors are more likely to disturb the results" (Pirralha & Weber, 2020; Poses et al., 2021, p. 245). Therefore, we also estimate the measurement quality by calculating the reliability coefficient Omega (Hayes & Coutts, 2020) of the sum score of attitudes towards immigration for each country in each round of the ESS.

We acknowledge the growing number of studies that have assessed the measurement invariance of the ESS immigration scale in recent years. However, the comparability of findings across these studies is limited due to the lack of consistency in the analytical strategies and methods used. In our study, we aim to enhance comparability and provide more reliable insights by adopting a constituent approach. In addition, we aim to contribute to the field by providing an accessible and reader-friendly introduction to MGCFA as a method for testing measurement invariance, which may enhance its practical application in the context of migration research.

This paper sets out by introducing the European Social Survey (ESS) as a data source for studying attitudes towards immigration. We then provide a comprehensive introduction to MGCFA and present an overview of previous research testing the comparability of attitudes towards immigration in the ESS. The following section outlines the present study – sample, model testing, and analytic strategy. Finally, the results of measurement invariance testing, latent means comparison, and measurement quality assessment are presented and discussed.

Attitudes Towards Immigration in the European Social Survey

The European Social Survey is a biannual cross-national survey aimed to track Europeans' attitudes, beliefs, and behaviors on different topics. Implemented in most European countries, the ESS is a cross-sectional, probability-based sample in which all individuals, residents in private households over the age of 15, are eligible.

Since its first round in 2002, the European Social Survey (ESS) has continuously surveyed attitudes towards immigration in several European countries and is thus one of the most widely used surveys for cross-national research on attitudes towards immigration (Roots et al., 2016). Across each round, it includes several items to assess attitudes towards immigration in its main questionnaire. Besides, in rounds 1 and 7, the ESS conducted a more comprehensive immigration module that specifically focused on various dimensions of attitudes towards immigration (Heath et al., 2016).

In this paper, we focus on three items measuring the concept attitudes towards immigration, included in the core module, and displayed in Table 1.

Table 1 Items used for measuring attitudes towards immigration (ATI)

Question wording	Item name	Item number	Response scale
Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries?	*imbgeco*	B41	0 (Bad for the economy) – 10 (Good for the economy)
And, using this card, would you say that [country]'s cultural life is generally undermined or enriched by people coming to live here from other countries?	*imueclt*	B42	0 (Cultural life undermined) – 10 (Cultural life enriched)
Is [country] made a worse or a better place to live by people coming to live here from other countries?	*imwbcnt*	B43	0 (Worse place to live) – 10 (Better place to live)

Invariance Testing with Multigroup Confirmatory Factor Analysis (MGCFA)

Measurement invariance tests rely on a latent variable approach. As a confirmatory factor analysis model, multigroup confirmatory factor analysis (MGCFA) techniques assume that the responses people provide to different items (observed responses) are caused by their position on an unobserved construct or factor (latent variable). Figure 1 represents this model for the latent factor attitudes towards immigration (ATI) that determines the answers to the three items in Table 1.

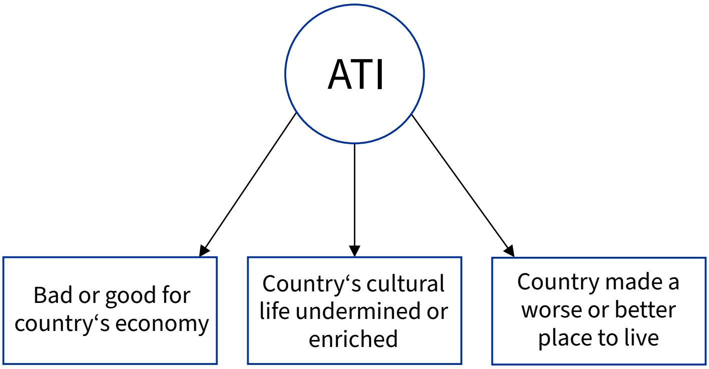

ATI
Bad or good for country‘s economy
Country‘s cultural life undermined or enriched
Country made a worse or better place to live

Figure 1 Measurement model for the latent factor ATI

Equation 1 provides an equation for the same model, where X_i is the observed item answer for the observed variable i, ξ is the latent factor ATI, and ε is the error term, i.e., the item variance unaccounted by the latent factor. τ represents the intercept (the expected value of each observed item when the value of the latent variable is zero), and λ_i is the loading (the expected increase in X_i for each one unit increase in ξ).

$$X_i = \tau_i + \lambda_i \xi + \varepsilon_i \tag{1}$$

The multigroup extension of MGCFA implies that this same measurement model is separately estimated in different groups, indicated with the subscript j, as depicted in Equation 2.

$$X_{ij} = \tau_{ij} + \lambda_{ij} \xi + \varepsilon_{ij} \tag{2}$$

Declaring measurement invariance implies asserting that the factor is measured in the same way across the different groups. Thus, in this framework, measurement invariance means that the parameters of the measurement model τ_i and λ_i are equal across all groups j. There are different possible equalities between parameters that can be satisfied across groups, giving rise to different measurement invariance levels.

First, *configural* invariance means that the general structure of the factor is equal across groups, i.e., that the same items load on the same factor(s). Since the model we estimate is unifactorial, this simply means that the loadings for none of the items are zero in any group and that there are no correlated error terms in only some groups. Whether the value of the parameters is equal across groups is not important at this level. The establishment of configural invariance is interpreted as evidence suggesting that, since the factor can be measured with the same items in all groups, the factor has a similar theoretical content across groups.

Second, *metric* (also known as *loading*) invariance indicates that the factor loadings are equal across groups. This means that a one unit increase in the factor leads to the same change in the observed item responses in all groups. This level of equivalence implies that factor variances and covariances (i.e., the relationship of the factor with other measures) can be compared meaningfully across groups.

Third, *scalar* (also known as *intercept*) invariance means that the item intercepts are equal across groups. This indicates that when the value of the latent variable is zero, the expected mean value of the item responses will be the same across all groups. Importantly, when metric and scalar invariance for an item are established across groups, it means that any given level in the latent variable of interest will lead to the same expected value of the observed item. Therefore, the simultaneous establishment of metric and scalar invariance allows for the comparison of sum scores or observed means.

To understand this last point, it is important to highlight the differences between sum scores, observed means, and latent means. Sum scores are the scores produced by simply summing, for each individual on the sample, the scores of all items that measure the latent factor. Observed means refer to the means of sum scores across all individuals in a given country. Latent means are the means of the latent factor ATI. While observed means are very simple to compute, latent means need to be estimated with specialized software and using structural equation modeling (or other latent variable) techniques. From Equation 1, we can infer the reason why observed means should not be compared in the absence of measurement invariance. If the loadings and intercepts are not equal across groups, the same level in the latent factor will lead to different expected values in the observed items. This implies, for instance, that the same mean level in an observed item across groups may correspond to different mean levels of the latent variable; or that different mean levels of an observed item across groups might correspond to the same mean value of the latent variable (see also Steinmetz, 2013). Said differently, if the loadings and intercepts are not equal across groups, the correspondence between latent means and observed means differ across groups. This means that differences in observed means are not trustworthy indicators of differences in latent means. Thus, observed means should not be compared because they do not necessarily reflect differences in latent means.

The situation of equality of all loadings and intercepts across groups is sometimes called *full invariance*. Full invariance has often been found to be too strict to achieve, especially for the intercepts (Davidov, Muthen, & Schmidt, 2018). This means that comparisons of observed means (or sum scores) across groups cannot be guaranteed, as differences in observed means might or might not reflect true differences in the latent means. A way to overcome this issue is to compare latent means instead, which can be done by establishing *partial invariance*. Partial invariance implies that only some of the parameters of the measurement model are equal across groups. Classic advice has been that latent means can be compared in situations with partial invariance when at least two of the loadings and intercepts are equal (Steenkamp & Baumgartner, 1998). More recent, simulations by Pokropek et al. (2019) have shown that the estimation of latent means is satisfactory in partial measurement invariance models, if items with partial measurement invariance are identified and freed, and that at least one item is invariant across groups.

Therefore, our aim is twofold. First, testing for measurement invariance of the attitudes towards immigration scale across countries and establishing the level of invariance (configural, metric, or scalar) that holds across each group of countries. Second, establishing partial invariance for cases where no full invariance is found, so that latent means can be compared.

Previous Research Testing the Comparability of Attitudes Towards Immigration in the European Social Survey

The number of studies evaluating measurement invariance of the immigration attitudes scale across the ESS countries and over time has increased but remains limited (Table 2). These studies differ in their methodological approaches, analytical strategies, and terminology used: “perceived ethnic threat” (Pirralha & Weber, 2020), “attitudes towards migration” (Borgonovi & Pokropek, 2019), “anti-immigrant attitudes” (Nickel, 2022). In the following, we will use the term “attitudes towards immigration” (ATI). For a detailed overview of the constructs, questions, and response scales used in the studies, see Table A1, Appendix.

Table 2 Studies testing the comparability of attitudes towards immigration (ATI) in ESS[3]

Study	Round	ATI	Countries	Method	Results
Cross-time and cross-national measurement invariance					
Meuleman, Davidov & Billiet 2009	ESS R1 (2002) – ESS R3 (2006)	REJECT	17	Multigroup confirmatory factor analysis (MGCFA)	Full scalar invariance within 17 countries; partial scalar invariance between 17 countries
Borgonovi & Pokropek 2019	ESS R5 (2010) – ESS R8 (2016)	REJECT THREAT	18	Multigroup Bayesian Structural Equation Modeling (MG-BSEM)	Full scalar invariance for each country over time; Metric invariance between countries
Cross-national measurement invariance					
Meuleman & Billiet 2012	ESS R1 (2002)	REJECT CONDITION ECOTHREAT CULTTHREAT	21	MGCFA	REJCET: Partial scalar invariance in all 21 countries; CULTTHREAT: Partial scalar invariance in 11 countries; CONDITION + ECOTHREAT: Partial scalar invariance in 14 countries
Davidov et al. 2015	ESS R1 (2002) – ESS R6 (2012)	REJECT	35	Approximate measurement invariance using Bayesian estimation	Approximate scalar invariance across all countries in each round
Davidov, Cieciuch & Schmidt 2018	ESS R7 (2014)	ALLOWANCE CONDITION RT (Realistic Threat)	15	Approximate measurement invariance using Bayesian estimation; Exact measurement invariance by MGCFA	ALLOWANCE: Approximate scalar invariance in 12 countries; RT: approx. scalar invariance in 13/14 countries; CONDITION: metric invariance in 7 countries
Pirralha & Weber 2020	ESS R3 (2006)	THREAT	19	MGCFA + correction for measurement error	Partial scalar invariance in 19 countries
Nickel 2022	ESS R9 (2018)	THREAT	28	MGCFA	Metric invariance in 29 countries

[3] Without any claim to completeness

Studies testing cross-time and cross-national measurement invariance

Meuleman et al. (2009) first started testing the comparability of the ESS immigration attitudes scale across three time points (ESS R1 (2002); ESS R2 (2004); ESS R3 (2006)). The authors provide technical guidance on how to measure scale invariance by applying multigroup confirmatory factor analysis (MGCFA) and using a top-down strategy: testing the most constrained model (full scalar invariance across time and countries) at first and then incrementally reducing the number of constraints assessing whether the model fit is improving. To measure ATI they construct a latent factor that measures the rejection of further immigration in general (REJECT). In their final model, full scalar invariance holds over time within the 17 countries and partial scalar invariance between the countries, implying that the ESS immigration attitudes can be meaningfully compared across countries and over the three time points.

Borgonovi and Pokropek (2019) published a study examining the measurement invariance, both across countries and across time, of two latent constructs measuring immigration attitudes: generalized threat (THREAT) and opposition to migration (REJECT). They considered four time points ESS R5 (2010) – R8 (2016), including 18 countries that participated in each round of the ESS. To test for partial and approximate measurement invariance they apply sequential methods using the multigroup Bayesian structural equation modeling (MG-BSEM; B. Muthén & Asparouhov, 2012). First, they measure cross-time comparability separately for each country, and second, cross-national comparability for each time point. They establish full scalar invariance over time within each county but only metric invariance across the countries. Indicating that the country means can be compared meaningfully over time for each country but that the different country means cannot be compared to each other within one time point.

Studies testing cross-national measurement invariance

Making use of the first more comprehensive module assessing immigration attitudes conducted in ESS round 1 (2002), Meuleman and Billiet (2012) test for measurement invariance for four latent factors: opposition against new immigration (REJECT); support for imposing conditions to immigration (CONDITION); perceived economic threat (ECOTHREAT); perceived cultural threat (CULTTHREAT). The REJECT scale holds partial scalar invariance (invariance applies at least for two items per construct) for all 21 countries, which allows for cross-national mean comparisons. The other three scales hold partial metric invariance in 18 to 19 countries, guaranteeing the cross-national comparability of regression coefficients and covariances. Partial scalar invariance holds only for 11 (CULTHREAT) to 14 countries (CONDITION, ECOTHREAT), implying that the country means of these three scales can only be meaningfully compared in some of the countries.

Davidov et al. (2015) extend these findings by testing for approximate measurement invariance of the REJECT scale across 35 countries and the first 6 ESS rounds[4]. As the traditional (exact) approach failed to support scalar and even partial scalar measurement invariance, the authors test for approximate measurement invariance using the Bayesian framework (B. Muthén & Asparouhov, 2012; van de Schoot et al., 2013). This procedure "allows variance around the point estimates for the factor loadings and intercepts of the indicators" (Davidov et al., 2015, p. 261), whereas in the exact approach factor loadings and intercepts would be constrained to be exactly equal. Their findings reveal that approximate scalar measurement invariance is established across all countries in each ESS round, guaranteeing comparable country means.

Based on data from the second comprehensive immigration module surveyed in ESS round 7 (2014), Davidov, Cieciuch, and Schmidt (2018) test for approximate measurement invariance of three latent constructs: opposition towards immigration (ALLOWANCE); qualification for entry or exclusion (CONDITION); realistic threat (RT). Their results show that approximate (not exact) scalar invariance for ALLOWANCE (12 countries) and RT (13 to 14 countries) can be found in most of the 15 countries considered. For CONDITION, neither exact nor approximate invariance holds, and metric invariance is established only in 7 countries.

Pirralha and Weber (2020) disentangle the cognitive from the measurement part and correct for measurement errors. They refer to the concept of perceived ethnic threat (similar to THREAT) and find partial scalar invariance which allows comparing the latent means across all 19 countries that participated in the ESS R3 (2006).

Further evidence for metric invariance of anti-immigrant attitudes (similar to THREAT) can be found in Nickel (2022). Using MGCFA for structural modeling, the results show that metric invariance holds for all 29 countries participating in ESS round 9 (2018), indicating that factor loadings are equivalent across these countries.

While the above-mentioned studies use different methods and analytical strategies, making it difficult to compare their results, we follow the same approach here for all nine ESS rounds: multigroup confirmatory factor analysis (MGCFA). Moreover, we also estimate the measurement quality of the sum score attitudes towards immigration to quantify how strong the relationship between the latent variable of interest, attitudes towards immigration, and its observed measure is.

[4] Measurement invariance was tested separately for each ESS round, the authors did not test for over-time comparability.

The Current Study

Sample

We focus on three items measuring the concept attitudes towards immigration, which are included in the core module and shown in Table 1. As these items are repeated in each round of the ESS, our analyses are based on data from round 1 (2002) to round 9 (2018)[5]. In total, we analyze data from 38 countries: Austria, Belgium, Switzerland, Czechia, Denmark, Spain, Finland, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Sweden, Slovenia, Estonia, Iceland, Slovakia, Turkey, Ukraine, Bulgaria, Cyprus, Russia, Croatia, Latvia, Romania, Lithuania, Albania, Kosovo, Montenegro, Serbia, and the United Kingdom. This led to a total sample of 390,276 individuals[6].

Model testing

In order to establish partial measurement invariance, we estimate models with equality constraints on the parameter among groups. We then use local fit testing to determine whether the imposed constraints are supported by the data. Local fit testing focuses on whether, in each group, each specific parameter of the model is misspecified. Concretely, we follow the local fit testing procedure suggested by Saris et al. (2009). This local fit testing procedure is based on a combination of the modification indices (approximating a significance test for the retrieval of one constraint), the expected parameter change when a constraint is relieved, and the power of the test to detect a misspecified parameter of a given effect size. The researcher must set the expected parameter change that they do consider to be a relevant misspecification: misspecifications lower than this size are not considered relevant and are thus ignored. Our criteria for the size of the misspecifications to be detected are 0.1 for the loadings, 0.15 for the intercepts[7],

[5] European Social Survey Round 1 Data, (2002); Round 2 Data, (2004); Round 3 Data, (2006); Round 4 Data, (2008); Round 5 Data, (2010); Round 6 Data, (2012); Round 7 Data, (2014 // 2015); Round 8 Data, (2016); Round 9 Data, (2018)

[6] To test for invariance across countries, we included only those cases where respondents provided answers to all three items. We employed listwise deletion, meaning that any case with missing data for any of the specified variables was excluded from the analysis. The item non-response varies over time and across countries but without a clear pattern. For detailed information on the sample size for each country in each ESS round, see Table A2, Appendix. We acknowledge that there are alternative methods for dealing with missing data, but follow the usual approach adapted by ESS Core Scientific Team (Zavala-Rojas & Saris, 2018; Revilla, 2012; Weber, 2011).

[7] In practice, given the standard deviations of the items in each country, this corresponds to an unstandardized effect size of between 0.3 and 0.5 in all cases, meaning that we aim to detect misspecifications larger than 10% of the total length of the response scale

and 0.2 for correlated error terms (all in standardized metrics), based on suggestions by Saris et al., 2009).

This procedure contrasts with the typical approach of relying on global fit indices (statistics such as Chi-square or fit indices such as the comparative fit index (CFI) or the root mean square error of approximation (RMSEA)) and evaluating models as a whole. We avoid global fit indices for two reasons. First, because of some of their drawbacks reported in the literature (sensitivity to sample size and different model characteristics, unequal sensitivity to different model misspecifications; Groskurth et al., 2021; Saris et al., 2009). Second, because relying on global fit indices does not allow for fine-grained assessments of invariance in the context of measurement invariance. Concretely, the use of global fit indices does not allow for recovering complex patterns of invariance across groups, i.e., among 24 groups, different levels of invariance are likely to be present in different subgroups of countries – but this level of detail cannot be achieved using global fit indices. Moreover, using global fit indices does not allow for identifying invariant items, which is a prerequisite to then free the invariant items and establish partial invariance. This is critical because establishing partial invariance is important when we want to compare latent means under conditions where full invariance is not present.

Analytical Strategy

The analyses were conducted in R (R Core Team, 2021), using the packages lavaan (Rosseel, 2012) and semTools (Jorgensen et al., 2021)[8]. We used ML estimation because the items were 11-point scales and did not present important skewness or kurtosis. The Saris et al. (2009) approach to local fit testing was implemented using the function miPowerFit in semTools. Factor variances were identified using the fixed factor approach, with fixing variances at 1 and means at 0, unless equality constraints in the model allowed to free these specifications (Schroeders & Gnambs, 2020). Our analyses are cross-sectional and not longitu-

(11-points), and occasionally misspecifications larger than a percentage of the scale smaller than 10 %.

[8] As van de Schoot et al. (2012) point out, measurement invariance testing is feasible with various structural equation modeling software programs. Lisrel (Jöreskog & Sörbom, 1996-2001) possesses the capability to handle categorical data, although it demands a proficiency in syntax and matrix algebra. AMOS (Arbuckle, 2007) is recognized for its user-friendly interface, but its capacity to handle categorical data is limited. Currently, Mplus (L. K. Muthén & Muthén, 2012) stands out as the most versatile program for measurement invariance testing, albeit requiring a proficiency in syntax. Additionally, Lavaan (Rosseel, 2012) and OpenMx (Boker et al., 2011), both open-source R packages in ongoing development, provide alternative options for measurement invariance testing, thereby enhancing the array of available tools in this domain.

dinal, i.e., we test the measurement invariance across countries in each round, but not across time for a given country.

For each round, the invariance test proceeds as follows. First, we estimate the configural model and check that no estimates are 0. Second, we estimate the loading invariance model – all loadings constrained to be equal – and test it using miPowerFit. If miPowerFit detects misspecified loadings, we free them and re-estimate the model. Each time, we free only one loading because model misspecifications are often related. We repeat this process until no misspecifications are present according to miPowerFit. Once a model with no misspecifications is reached, we compare the value of the freed loadings. This step is done to evaluate further comparability across subgroups of countries: it might be that some countries are non-invariant with respect to the majority of the groups but invariant among them. When freed loadings deviate in the same direction compared to most groups (e.g., the freed loadings of more than one group are higher than for the rest of the countries), we additionally constrain them to be equal to each other. We then re-estimate and test the model again; in the rare occasion that misspecifications reappear, we also free them one by one.

After establishing the highest possible level of metric invariance, we move on to scalar invariance. The process for scalar invariance is the same as for metric invariance. First, we constrain the intercepts to be equal – although in this step we do not constrain the intercepts for the groups and items for which metric invariance was not established. We then test the model using miPowerFit and free the intercepts one by one. In the end, we compare the value of the freed intercepts and set additional equality constraints among the freed intercepts with similar estimated values.

Measurement Quality

Estimating the measurement quality is essential to correct for measurement errors (Saris & Gallhofer, 2014), but also to understand how much of the concept of interest – attitudes towards immigration – is measured by the created sum score[9]. A perfect relationship would be 1 with no measurement errors present. The measurement quality of the unweighted sum scores (q_s^2) is defined as:

[9] The survey quality predictor (SQP) database, developed by Saris et al (2011), serves as an open source tool for evaluating the quality of individual questions in the ESS (see https://www.europeansocialsurvey.org/methodology/ess-methodology/data-quality-assessment). Saris and Gallhofer (2014) suggest that SQP can also be used to assess the quality of composite scores by utilizing information on the quality of individual questions.
For further insights into how measurement quality can be improved by correcting measurement errors in the ESS, various reports, working papers, and articles are available at https://www.europeansocialsurvey.org/methodology/methodological-research/correction-measurement-error

$$q_s^2 = 1 - \left[\frac{\sigma^2(e_s)}{\sigma^2(s)}\right]$$

$\sigma^2(e_s)$ is the variance of the errors in the sum score and $\sigma^2(s)$ is the variance of the sum score (s). This can be estimated, using the loadings (λ_i) of the final scalar model, as follows:

$$q_s^2 = 1 - \left[\frac{(\Sigma(1 - \lambda_i^2) * \sigma^2(y_i))}{\sigma^2(s)}\right]$$

The measurement quality of the sum score can range from 0 to 1, where we consider a $q^2 < 0.6$ as poor, $0.6 \leq q^2 < 0.7$ as questionable, $0.7 \leq q^2 < 0.8$ as acceptable, $0.8 \leq q^2 < 0.9$ as good, and $q^2 \geq 0.9$ as excellent quality, and 1 as perfect (DeCastellarnau & Revilla, 2017).

Results

Measurement Invariance

Figure 2 shows the results for the invariance of loadings (metric invariance) across countries. Countries illustrated in gray are not comparable, and countries shown without color were not part of the analysis. For countries with the same color, either green or purple, factor variances and covariances can be compared. We followed a specific analytical procedure: Initially, we released the equality constraints for all non-invariant countries, allowing for measurement variations across these countries. Subsequently, we conducted tests of invariance within this group. This process led us to identify a second group of countries, represented in purple, that are comparable to each other.

As can be seen in the maps, metric invariance is generally satisfied for the items in almost all countries in all rounds, except one or two countries in each group. Only the items in one country show a clear pattern of non-invariance in most rounds: Finland. In other countries, occasionally non-invariant items are found: in Italy in R1, in Denmark and France in R2, in Denmark, France and Estonia in R3, in Romania and Slovakia in R4, in Portugal and Slovakia in R5, in Hungary and Portugal in R6, and in Poland in R8. These results imply that for most countries, a one unit increase in the latent factor of interest leads to the same change in the expected value of the responses to the item across countries.

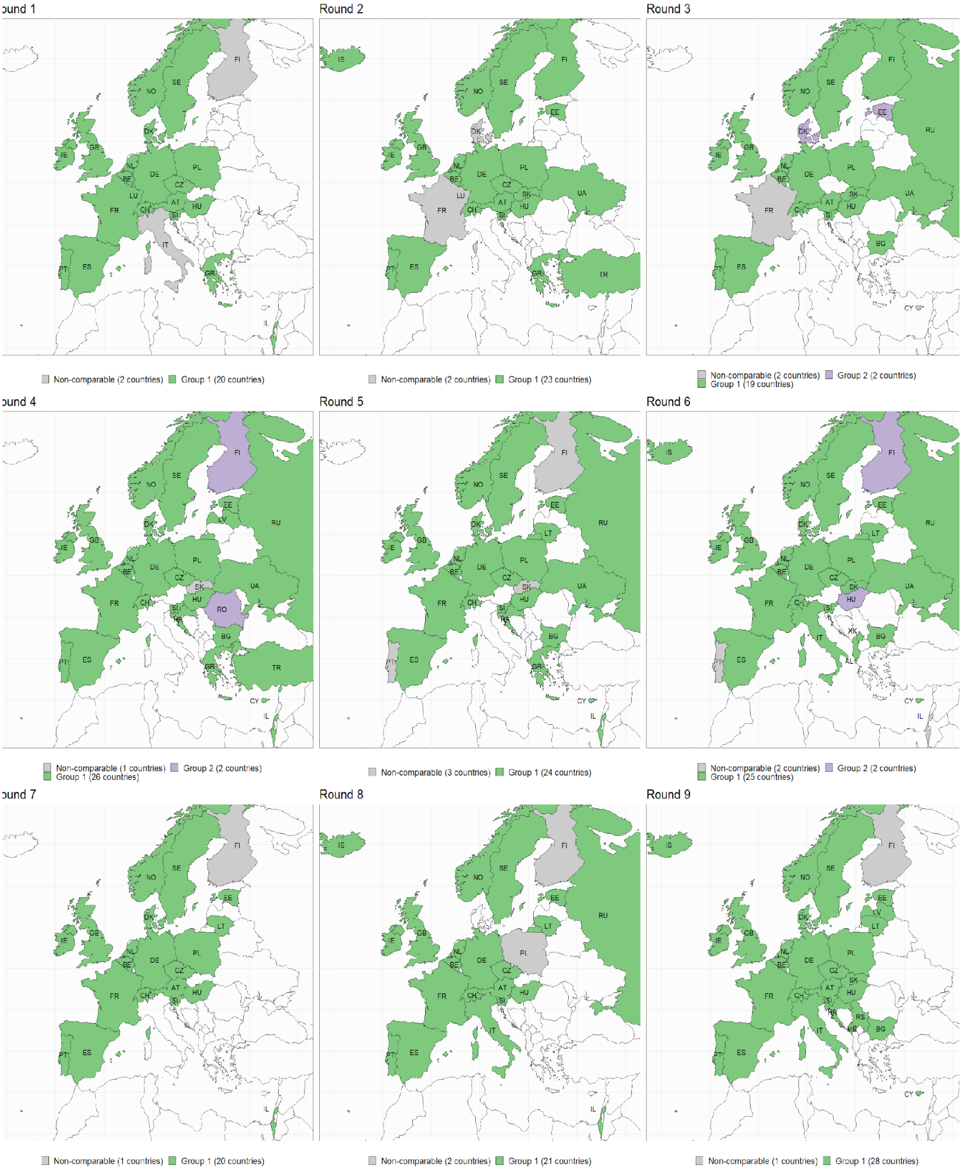
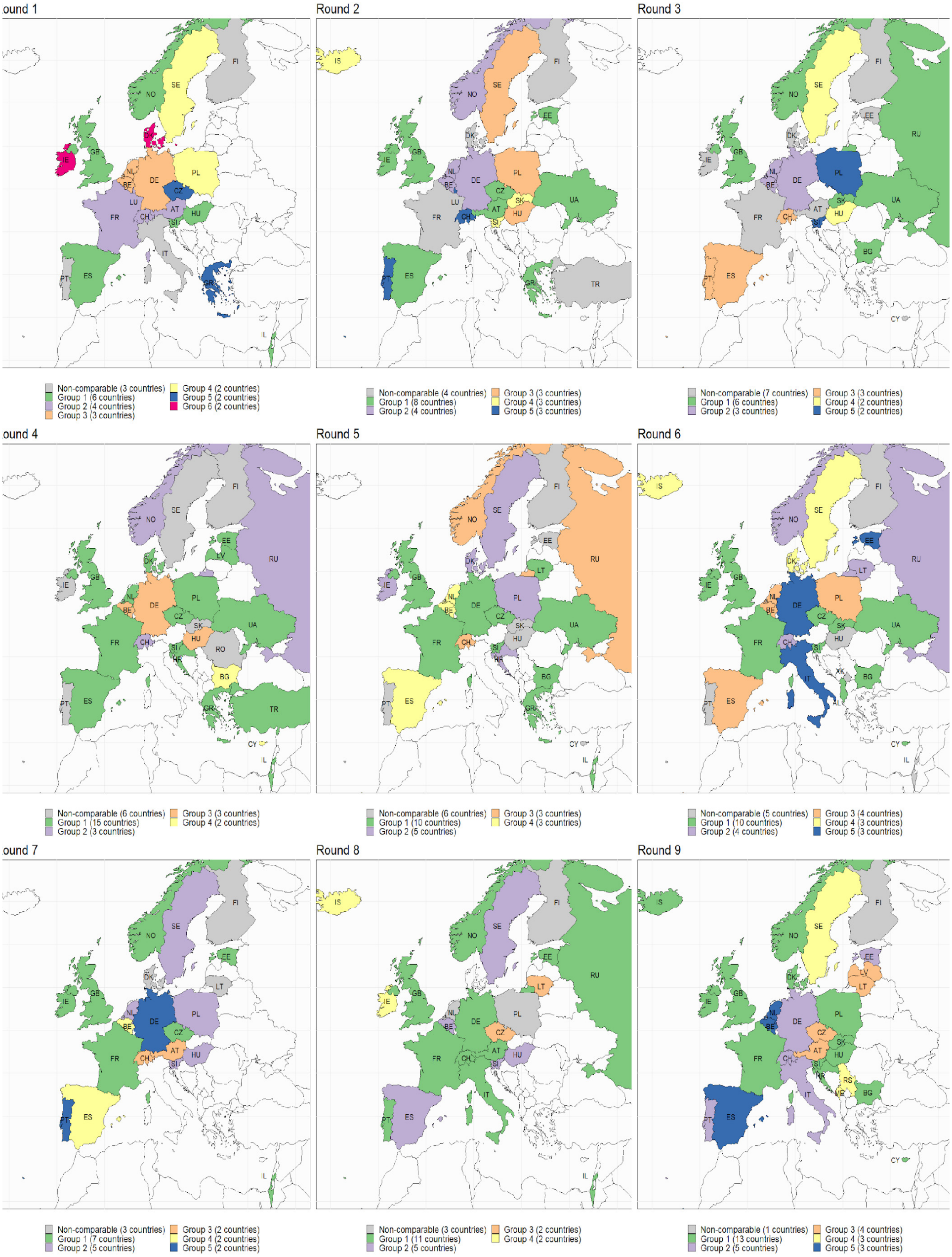


Figure 2 Metric invariance across countries in the nine rounds of the ESS

Figure 3 shows the results for scalar (intercept) invariance. In contrast to metric invariance, scalar invariance is much less widespread. First, there are different subgroups of countries for which scalar invariance holds. However, the larger groups includes a maximum of 52% of the countries (in round 4), and a minimum of 26% of the countries (in round 3). Moreover, between 30% (in round 3) to 3% (in round 9) of the countries in each round do not share intercepts with any other country of that round. This implies that for most of the countries, the same level of the latent factor corresponds to a different mean level in the responses to at least one of the items. In Figure 3, "comparability" refers to the comparability of the observed means across countries. As can be seen, comparisons of observed means across countries are not possible in many cases. In each round, observed means can only be compared across countries that have the same intercepts (shown with the same color), i.e., across a relatively small subset of countries.

Regarding the sources of scalar invariance, the most non-invariant item is the item 'Immigration bad or good for the economy' (*imbgeco*). Across all rounds and countries, 28% of the intercepts had to be freed for this item. This is followed by the item 'Immigration undermines or enriches cultural life' (*imueclt*), for which 24% of the intercepts had to be freed. Lastly, 13% of the intercepts for the item 'Immigration makes countries a worse or better place to live' (*imwbcnt*) had to be freed.

Regarding the countries, the country with the most non-invariant items was Finland (across all three items, 40% of its intercepts had to be freed; most of these corresponded to the item '*imueclt*', which had to be freed in every round). Finland is followed by Sweden and Portugal (37% of the items had to be freed; most of these corresponded to '*imbgeco*' for Sweden and to '*imwbcnt*' for Portugal). In contrast, for the countries with fewer non-invariant items, only one item in one round was found to be non-invariant across all items and rounds. These countries were Israel (representing 6% of all intercepts), Bulgaria (7%), Greece (8%), and Croatia (11%).



ound 1

Round 2

Round 3

ound 4

Round 5

Round 6

ound 7

Round 8

Round 9

Figure 3 Scalar invariance across countries in the nine rounds of the ESS

Comparison of Latent Means

Given that scalar invariance does not hold for most countries, differences in observed means are not reliable indicators of differences in latent means. Thus, in most cases, the observed means should not be compared across all countries. However, since partial invariance is satisfied for all countries (at least one indicator is equal for all countries), latent means can be compared directly. The exact latent means are shown in Table A3 in the Appendix. In all rounds, we chose Germany as a reference point for identification purposes (i.e., its mean is always 0, and the latent mean estimates are relative to those of Germany).

Consistent with previous research, our results again confirm that the Northern European countries – Sweden, Norway, Denmark, Finland – tend to be more positive towards immigration, while Eastern Europe – Czechia, Hungary, Ukraine, Slovakia – and Southern Europe – Italy, Greece, Cyprus, Slovenia – hold the most negative attitudes. Some countries remain in the middle range, representing moderate attitudes towards immigration – Netherlands, Ireland, and to some extent Germany. Over the years, Sweden and Iceland have consistently been the most immigration-friendly countries, while Finland, Norway and Denmark rank at least in the top half.

Measurement Quality

As summarized in Table A4, Appendix, the measurement quality ranges from .68 in Luxembourg in round 1 to .94 in Bulgaria in round 9. This means that between 32% and 6% of the variance in the sum scores is due to measurement error, which should be accounted for in further analyses (Saris & Gallhofer, 2014; Saris & Revilla, 2016). Table 3 shows all measurement quality estimates for each round and country analyzed. The performance of the measure is better in countries such as the United Kingdom, Bulgaria, and Ukraine and worse in countries such as the Netherlands or Switzerland. Overall, the differences between countries are small. Besides, the performance of the measure is worse in round 1 compared to the other rounds, while it is rather similar in the rest of the rounds.

Discussion and Conclusions

In cross-national research, it is not common practice to test for measurement invariance. However, it is becoming more popular due to simplified analytical strategies in widely used statistical software. Assuming measurement equivalence without testing it can cause biased mean comparisons, covariances, and regression coefficients. Thus, it is essential to assess whether metric or scalar invariance holds for the countries and time points considered.

The aim of this study was to test the comparability and quality of the ATI scale within each round of the ESS (ESS R1 (2002) to ESS R9 (2018)). While previous research used different methods and analytic strategies, we applied the same approach in all rounds: multigroup confirmatory factor analysis (MGCFA) with local fit testing. Our results reveal that metric (loading) invariance generally holds for the items in almost all countries in all rounds except Finland. As in Davidov, Cieciuch, and Schmidt (2018), our findings again show a clear pattern of non-invariance for Finland in most rounds. Moreover, the factor loadings (slopes) are the same in most countries, indicating that covariances and regression coefficients can be meaningfully compared across most countries in all ESS rounds from 2002 to 2018.

In contrast, a less positive conclusion must be drawn in the case of scalar (intercept) invariance. It holds only for different subgroups of countries within each round, and the size of these subgroups varies considerably between ESS rounds. While scalar invariance holds for 52 % of the countries in round 4, it holds for only 26 % in round 3. Between 30 % (round 3) to 3 % (round 9) of the countries have different intercepts for at least one of the items. Thus, mean comparisons are only possible for a relatively small subset of countries.

With respect to the sources of scalar invariance, the most non-invariant item is the item ‘Immigration bad or good for the economy’ (*imbgeco*) (see Borgonovi & Pokropek, 2019). Since the question is asked quite generally on the topic of immigration, the answers may strongly depend on whether the respondents – and this is influenced by their cultural and political background – think of immigration in terms of illegal migration or skilled labor migration or whether they think of immigrants as people of the same or different ethnic or religious origin.

However, while scalar invariance does not hold for most countries, partial invariance is fulfilled in all countries, meaning that at least one item is equal for all countries. Therefore, latent means can be compared directly (Pokropek et al., 2019).

Our results are more or less in line with previous research showing that Europe can be classified geographically in terms of attitudes towards immigration: Whereas Northern Europe is generally more supportive of immigration, Eastern and Southern Europe are more opposed to it.

By providing an accessible and reader-friendly introduction to measurement invariance testing using multi-group confirmatory factor analysis, we aimed to support its practical application. Researchers can confidently rely on our findings and compare regression coefficients and latent means of attitudes towards immigration across countries within all ESS rounds from 2002 to 2018.

One major advantage of MGCFA is the assessment of the equivalence of measurements and structural relations across multiple groups (Harrington, 2008). MGCFA is particularly useful for comparing groups when dealing with tests comprising a substantial number of continuous items or subscale scores that are assumed to measure a limited set of underlying factors. It ensures that observed group differences are not attributable to measurement bias or variation in the underlying construct structures (Lubke, 2003).

However, several limitations need to be acknowledged: First, when comparing a large number of groups, or in longitudinal research when comparing many periods or periods far apart in time, the use of the MGCFA approach has an increased likelihood of incorrectly detecting non-invariance (Immekus, 2021; Kim et al., 2017; Leitgöb et al., 2023). To address these challenges, alternatives such as multilevel confirmatory factor analysis (ML CFA), multilevel factor mixture modeling (ML FMM), Bayesian approximate measurement invariance testing (Muthén & Asparouhov, 2013a), and alignment optimization (Asparouhov & Muthén, 2014) are suggested.

Second, the length of the scale affects the effectiveness of fit measures (D'Urso et al., 2022). When using MGCFA for measurement invariance testing of long scales, the commonly used cut-off values for RMSEA and CFI may be insufficient.

Third, the multiple indicators and multiple causes (MIMIC) modeling procedure is a recent addition to the SEM family (Tsaousis et al., 2020). In contrast to MGCFA, the MIMIC approach allows to test for measurement invariance of both categorical and continuous individual difference variables (Barendse et al., 2010) and has smaller sample size requirements than MGCFA (Leitgöb et al., 2023).

In addition, we estimated the measurement quality of the ATI score. However, our findings reveal that although the measurement quality differs across the countries, these differences are relatively small. Moreover, the performance of the measurement is quite similar across the ESS rounds, except for the first time the ESS was conducted. While this appears to give credit to the rigorous methodological approach of the ESS, there are still some measurement errors as the quality is not perfect. This stresses the importance of measurement errors correction (Saris & Revilla, 2016).

Our study is limited to cross-sectional invariance testing, which provides insights into the measurement invariance of attitudes towards immigration at a specific point in time. However, to ensure the comparability of ATI within countries across different rounds, future research is needed to incorporate cross-time invariance testing.

Ongoing and comparative research on attitudes towards immigration remains an essential task for the social sciences. Understanding the dynamics of public opposition to immigration is crucial, as it has been shown to have negative effects on social cohesion, on the lives of immigrants and refugees, and to contribute to the rise of populist radical right parties. To understand, explain, and effectively address this, accurate measurement is essential.

References

Arbuckle, J. L. (2007). *Amos 16.0 User's guide*. Spring House.

Barendse, M. T., Oort, F. J., & Garst, G. J. A. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias; a simulation study. *AStA Advances in Statistical Analysis, 94*(2), 117–127. https://doi.org/10.1007/s10182-010-0126-1

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and Testing Random Indirect Effects and Moderated Mediation in Multilevel Models: New Procedures and Recommendations. *Psychological Methods, 11*((2)), 142–163. https://doi.org/10.1037/1082-989X.11.2.142

Bohman, A. (2015). It's who you Know. Political Influence on Anti-Immigrant Attitudes and the Moderating Role of Intergroup Contact. *Sociological Research Online, 20*(3), 62–78. https://doi.org/10.5153/sro.362

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., Bates, T., Mehta, P., & Fox, J. (2011). Openmx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika, 76*(2), 306–317. https://doi.org/10.1007/s11336-010-9200-6

Borgonovi, F., & Pokropek, A. (2019). Education and Attitudes Toward Migration in a Cross Country Perspective. *Frontiers in Psychology, 10,* 2224. https://doi.org/10.3389/fpsyg.2019.02224

Davidov, E., Cieciuch, J., Meuleman, B., Schmidt, P., Algesheimer, R., & Hausherr, M. (2015). The Comparability of Measurements of Attitudes toward Immigration in the European Social Survey. *Public Opinion Quarterly, 79*(S1), 244–266. https://doi.org/10.1093/poq/nfv008

Davidov, E., Cieciuch, J., & Schmidt, P. (2018). The cross-country measurement comparability in the immigration module of the European Social Survey 2014-15. *Survey Research Methods, 12*(1), 15–27. https://doi.org/10.18148/srm/2018.v12i1.7212

Davidov, E., & Meuleman, B. (2012). Explaining Attitudes Towards Immigration Policies in European Countries: The Role of Human Values. *Journal of Ethnic and Migration Studies, 38*(5), 757–775. https://doi.org/10.1080/1369183X.2012.667985

Davidov, E., Muthen, B., & Schmidt, P. (2018). Measurement Invariance in Cross-National Studies. *Sociological Methods & Research, 47*(4), 631–636. https://doi.org/10.1177/0049124118789708

DeCastellarnau, A., & Revilla, M. (2017). Two Approaches to Evaluate Measurement Quality in Online Surveys: An Application Using the Norwegian Citizen Panel. Advance online publication (415-433 Pages / Survey Research Methods, Vol 11, No 4 (2017)). https://doi.org/10.18148/srm/2017.v11i4.7226

D'Urso, E. D., Roover, K. de, Vermunt, J. K., & Tijmstra, J. (2022). Scale length does matter: Recommendations for measurement invariance testing with categorical factor analysis and item response theory approaches. *Behavior Research Methods, 54*(5), 2114–2145. https://doi.org/10.3758/s13428-021-01690-7

European Social Survey Round 1 Data. (2002). *Data file edition 6.6. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.* https://doi.org/10.21338/NSD-ESS1-2002

European Social Survey Round 2 Data. (2004). *Data file edition 3.6. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.* https://doi.org/10.21338/NSD-ESS2-2004

European Social Survey Round 3 Data. (2006). *Data file edition 3.7. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.* https://doi.org/10.21338/NSD-ESS3-2006

European Social Survey Round 4 Data. (2008). *Data file edition 4.5. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.* https://doi.org/10.21338/NSD-ESS4-2008

European Social Survey Round 5 Data. (2010). *Data file edition 3.4. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.* https://doi.org/10.21338/NSD-ESS5-2010

European Social Survey Round 6 Data. (2012). *Data file edition 2.4. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.* https://doi.org/10.21338/NSD-ESS6-2012

European Social Survey Round 7 Data, & European Social Survey ERIC. (2014 // 2015). *Data file edition 2.2. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC // European Social Survey (ESS), Round 7 - 2014.* https://doi.org/10.21338/NSD-ESS7-2014

European Social Survey Round 8 Data. (2016). *Data file edition 2.2. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.* https://doi.org/10.21338/NSD-ESS8-2016

European Social Survey Round 9 Data. (2018). *Data file edition 3.1. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.* https://doi.org/10.21338/NSD-ESS9-2018

Fischer, R., Karl, J. A., Fontaine, J. R. J., & Poortinga, Y. H. (2022). Evidence of Validity Does not Rule out Systematic Bias: A Commentary on Nomological Noise and Cross-Cultural Invariance. *Sociological Methods & Research*, 004912412210917. https://doi.org/10.1177/00491241221091756

Groskurth, K., Bluemke, M., & Lechner, C. M. (2021). *Why We Need to Abandon Fixed Cutoffs for Goodness-of-Fit Indices: An Extensive Simulation and Possible Solutions.* https://doi.org/10.31234/osf.io/5qag3

Harrington, D. (2008). Use of Confirmatory Factor Analysis with Multiple Groups. In D. Harrington (Ed.), *Confirmatory Factor Analysis* (pp. 78–99). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195339888.003.0005

Hayes, A. F., & Coutts, J. J. (2020). Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But… *Communication Methods and Measures, 14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Heath, A., Schmidt, P., Green, E., Ramos, A., Davidov, E., & Ford, R. (2016). *Attitudes towards Immigration and their Antecedents: Topline Results from Round 7 of the European Social Survey* (ESS Topline Results Series), *7*. https://www.europeansocialsurvey.org/docs/findings/ESS7_toplines_issue_7_immigration.pdf

Holland, P. W., & Wainer, H. (Eds.). (2015). *Differential Item Functioning* (1st edition). London. Routledge.

Immekus, J. C. (2021). Multigroup CFA and alignment approaches for testing measurement invariance and factor score estimation: Illustration with the schoolwork-related anxiety survey across countries and gender. *Methodology, 17*(1), 22–38. https://doi.org/10.5964/meth.2281

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409–426. https://doi.org/10.1007/BF02291366

Jöreskog, K. G., & Sörbom, D. (1996-2001). *LISREL 8: User's reference guide (2nd ed.)*. Scientific Software International.

Jorgensen, T., Pornprasertmant, S., Schoeman, A., & Rosseel, Y. (2021). *semTools: Useful tools for structural equation modeling.* (0.5-5.912) [Computer software].

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 524–544. https://doi.org/10.1080/10705511.2017.1304822

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., Roover, K. de, Jak, S., Meitinger, K., Menold, N., Muthén, B., Rudnev, M., Schmidt, P., & van de Schoot, R. (2023). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research, 110,* 102805. https://doi.org/10.1016/j.ssresearch.2022.102805

Lubke, G. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence, 31*(6), 543–566. https://doi.org/10.1016/S0160-2896(03)00051-5#

Meitinger, K., Davidov, E., Schmidt, P., & Braun, M. (2020). Measurement Invariance: Testing for It and Explaining Why It is Absent. *Survey Research Methods, 14*(4), 345–349. https://doi.org/10.18148/srm/2020.v14i4.7655

Meuleman, B., & Billiet, J. (2012). Measuring Attitudes toward Immigration in Europe: The Cross-cultural Validity of the ESS Immigration Scales. *Ask Research & Methods, 21*(1), 5–29. http://hdl.handle.net/1811/69578

Meuleman, B., Davidov, E., & Billiet, J. (2009). Changing attitudes toward immigration in Europe, 2002-2007: A dynamic group conflict theory approach. *Social Science Research, 38*(2), 352–365. https://doi.org/10.1016/j.ssresearch.2008.09.006

Meuleman, B., Żółtak, T., Pokropek, A., Davidov, E., Muthén, B., Oberski, D. L., Billiet, J., & Schmidt, P. (2022). Why Measurement Invariance is Important in Comparative Research. A Response to Welzel et al. (2021). *Sociological Methods & Research,* 004912412210917. https://doi.org/10.1177/00491241221091755

Millsap, R. E. (2011). *Statistical approaches to measurement invariance,* New York. Routledge.

Mudde, C. (2007). *Populist radical right parties in Europe,* Cambridge, UK, New York. Cambridge University Press.

Mudde, C. (2020). Riding the fourth wave. *IPPR Progressive Review, 26*(4), 296–304. https://doi.org/10.1111/newe.12175

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335. https://doi.org/10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide*. Muthén & Muthén.

Nickel, A. (2022). Institutional Anomie, Market-Based Values and Anti-Immigrant Attitudes: A Multilevel Analysis in 28 European Countries. *International Journal of Conflict and Violence (IJCV), 16*(1), 1–15. https://doi.org/10.11576/ijcv-5126

Pirralha, A., & Weber, W. (2020). Correction for measurement error in invariance testing: An illustration using SQP. *PloS One, 15*(10), e0239421. https://doi.org/10.1371/journal.pone.0239421

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo Simulation Study to Assess The Appropriateness of Traditional and Newer Approaches to Test for Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(5), 724–744. https://doi.org/10.1080/10705511.2018.1561293

Poses, C., Revilla, M., Asensio, M., Schwarz, H., & Weber, W. (2021). Measurement quality of 67 common social sciences questions across countries and languages based on

28 Multitrait-Multimethod experiments implemented in the European Social Survey. Advance online publication (235-256 Pages / Survey Research Methods, Vol 15 No 3 (2021)). https://doi.org/10.18148/srm/2021.v15i3.7816

Putnick, D. L., & Bornstein, M. H. (2016). Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Developmental Review : DR, 41,* 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Quillian, L. (1995). Prejudice as a Response to Perceived Group Threat: Population Composition and Anti-Immigrant and Racial Prejudice in Europe. *American Sociological Review, 60*(4), 586. https://doi.org/10.2307/2096296

R Core Team. (2021). *R: A Language and Environment for Statistical Computing (4.0.4) [R].* R Foundation for Statistical Computing. https://www.R-project.org

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *The Journal of Applied Psychology, 87*(3), 517–529. https://doi.org/10.1037/0021-9010.87.3.517

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552–566. https://doi.org/10.1037/0033-2909.114.3.552

Revilla, M. A. (2012). Measurement invariance and quality of composite scores in a face-to-face and a web survey. Advance online publication (17-28 Pages / Survey Research Methods, Vol 7, No 1 (2013)). https://doi.org/10.18148/srm/2013.v7i1.5098

Roots, A., Masso, A., & Ainsaar, M. (2016). *Measuring Attitudes towards Immigrants: Validation of Immigration Attitude Index Across.* Lausanne: European Social Survey Conference, 13-15th July.

Rosseel, Y. (2012). lavaan : An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2). https://doi.org/10.18637/jss.v048.i02

Saris, W. E., & Gallhofer, I. N. (2014). *Design, Evaluation, and Analysis of Questionnaires for Survey Research.* Wiley. https://doi.org/10.1002/9781118634646

Saris, W. E., & Revilla, M. (2016). Correction for Measurement Errors in Survey Research: Necessary and Possible. *Social Indicators Research, 127*(3), 1005–1020. https://doi.org/10.1007/s11205-015-1002-x

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal, 16*(4), 561–582. https://doi.org/10.1080/10705510903203433

Scheepers, P., Gijberts, M., & Coenders, M. (2002). Ethnic Exclusionism in European Countries. Public Opposition to Civil Rights for Legal Migrants as a Response to Perceived Ethnic Threat. *European Sociological Review, 18*(1), 17–34. https://doi.org/10.1093/esr/18.1.17

Schroeders, U., & Gnambs, T. (2020). Degrees of Freedom in Multigroup Confirmatory Factor Analyses. *European Journal of Psychological Assessment, 36*(1), 105–113. https://doi.org/10.1027/1015-5759/a000500

Seddig, D., Maskileyson, D., & Davidov, E. (2020). The Comparability of Measures in the Ageism Module of the Fourth Round of the European Social Survey, 2008-2009. Advance online publication (351-364 Pages / Survey Research Methods, Vol 14 No 4 (2020): Special Issue: Measurement Equivalence: Testing for It and Explaining Why It is Absent). https://doi.org/10.18148/srm/2020.v14i4.7369

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *The Journal of Applied Psychology, 91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research, 25*(1), 78–107. https://doi.org/10.1086/209528

Steinmetz, H. (2013). Analyzing Observed Composite Differences Across Groups. *Methodology, 9*(1), 1–12. https://doi.org/10.1027/1614-2241/a000049

Tsaousis, I., Sideridis, G. D., & AlGhamdi, H. M. (2020). Measurement Invariance and Differential Item Functioning Across Gender Within a Latent Class Analysis Framework: Evidence From a High-Stakes Test for University Admission in Saudi Arabia. *Frontiers in Psychology, 11*, 622. https://doi.org/10.3389/fpsyg.2020.00622

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4*, 770. https://doi.org/10.3389/fpsyg.2013.00770

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods, 3*(1), 4–70. https://doi.org/10.1177/109442810031002

Weber, W. (2011). Testing for measurement equivalence of individuals' left-right orientation. Advance online publication (1-10 Pages / Survey Research Methods, Vol 5, No 1 (2011)). https://doi.org/10.18148/srm/2011.v5i1.4622

Weldon, S. A. (2006). The Institutional Context of Tolerance for Ethnic Minorities: A Comparative, Multilevel Analysis of Western Europe. *American Journal of Political Science, 50*(2), 331–349. https://www.jstor.org/stable/3694276

Welzel, C., Brunkert, L., Kruse, S., & Inglehart, R. F. (2021). Non-invariance? An Overstated Problem With Misconceived Causes. *Sociological Methods & Research*, 004912412199552. https://doi.org/10.1177/0049124121995521

Welzel, C., Kruse, S., & Brunkert, L. (2022). Against the Mainstream: On the Limitations of Non-Invariance Diagnostics. *Sociological Methods & Research*, 004912412210917. https://doi.org/10.1177/00491241221091754

Widaman, K. F., & Grimm, K. J. (2014). Advanced psychometrics: Confirmatory factor analysis, item response theory, and the study of measurement invariance. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 534–570). Cambridge University Press.

Zavala-Rojas, D., & Saris, W. E. (2018). Measurement Invariance in Multilingual Survey Research: The Role of the Language of the Questionnaire. *Social Indicators Research, 140*(2), 485–510. https://doi.org/10.1007/s11205-017-1787-x

Appendix

Table A1: Study constructs of attitudes towards immigration, questions, and response scales

Questions	Response scale	Studies
ESS Round 1 (2002) – 9 (2018)		
To what extent do you think [country] should allow people … of the same race or ethnic group from most [country] people to come and live here? … of a different race or ethnic group from most [country] people to come and live here? … from the poorer countries outside Europe to come and live here?	1 “allow none” to 4 “allow many”	Meuleman, Davidov, and Billiet 2009: „REJECT“; Davidov et al. 2015: “Attitudes towards migration”; Borgonovi and Pokropek 2019: “Opposition to migration”
Would you say that … it is generally bad or good for [country]’s economy that people come to live here from other countries? … [country]’s cultural life is generally undermined or enriched by people coming to live here from other countries? … [country] is made a worse or a better place to live by people coming to live here from other countries?	0 “Bad / undermined/ worse” to 10 “good/ enriched/ better”	Borgonovi and Pokropek 2019: „Economic Threat“; Pirralha and Weber 2020: “Perceived ethnic threat”; Nickel 2022: “Anti-immigrant attitudes”

Questions	Response scale	Studies
ESS Round 7 (2014) Immigration Module		
To what extent do you think [country] should allow people … of the same race or ethnic group from most [country] people to come and live here? … of a different race or ethnic group from most [country] people to come and live here? … from the poorer countries outside Europe to come and live here? … from the richer countries in Europe? ... from the poorer countries in Europe to come and live here? ... from the richer countries outside Europe to come and live here?	1 “many” to 4 “none”	Meuleman and Billiet 2012: “REJECT”
Please tell me how important you think each of these things should be in deciding whether someone born, brought up, and living outside [country] should be able to come and live here. … have good educational qualifications. … have close family living here. … be able to speak [country]’s official language(s). … have work skills that [country] needs.	0 “extremely unimportant” to 10 “extremely important”	Meuleman and Billiet 2012: “CONDITION”

Questions	Response scale	Studies
People who come to live and work here generally harm the economic prospects of the poor more than the rich.	1 “agree strongly” to 5 “disagree strongly”	Meuleman and Billiet 2012: “ECOTHREAT”
If people who have come to live and work here are unemployed for a long period, they should be made to leave.		
Would you say that people who come to live here generally take jobs away from workers in [country], or generally help to create new jobs?	0 “take jobs away” to 10 “create new jobs”	
Most people who come to live here work and pay taxes. They also use health and welfare services. On balance, do you think people who come here take out more than they put in or put in more than they take out?	0 “generally take out more” to 10 „generally put in more“	
Would you say it is generally bad or good for [country]’s economy that people come to live here from other countries?	0 “bad for the economy” to 10 “good for the economy“	
Would you say that [country]’s cultural life is generally undermined or enriched by people coming to live here from other countries?	0 “cultural life undermined” to 10 “cultural life enriched“	Meuleman and Billiet 2012: “CULTTHREAT”
Please say how much you agree or disagree with each of the following statements. It is better for a country if almost everyone shares the same customs and traditions It is better for a country if there are a variety of different religions	1 “agree strongly” to 5 “disagree strongly”	

Questions	Response scale	Studies
To what extent do you think [country] should allow . . . people from other countries to come and live in [country]? … different race … Jewish … Muslims … Gypsies	1 “many” to 4 “none”	Davidov, Cieciuch, and Schmidt 2018: “ALLOWANCE”
Please tell me how important you think each of these things should be in deciding whether someone born, brought up and living outside [country] should be able to come and live here. … have good educational qualifications. … be able to speak [country]’s official language(s). . . . come from Christian background? . . . be white? … have work skills that [country] needs. . . . be committed to the way of life in [country]?	0 “extremely unimportant” to 10 “extremely important”	Davidov Cieciuch, and Schmidt 2018: “CONDITIONS”;

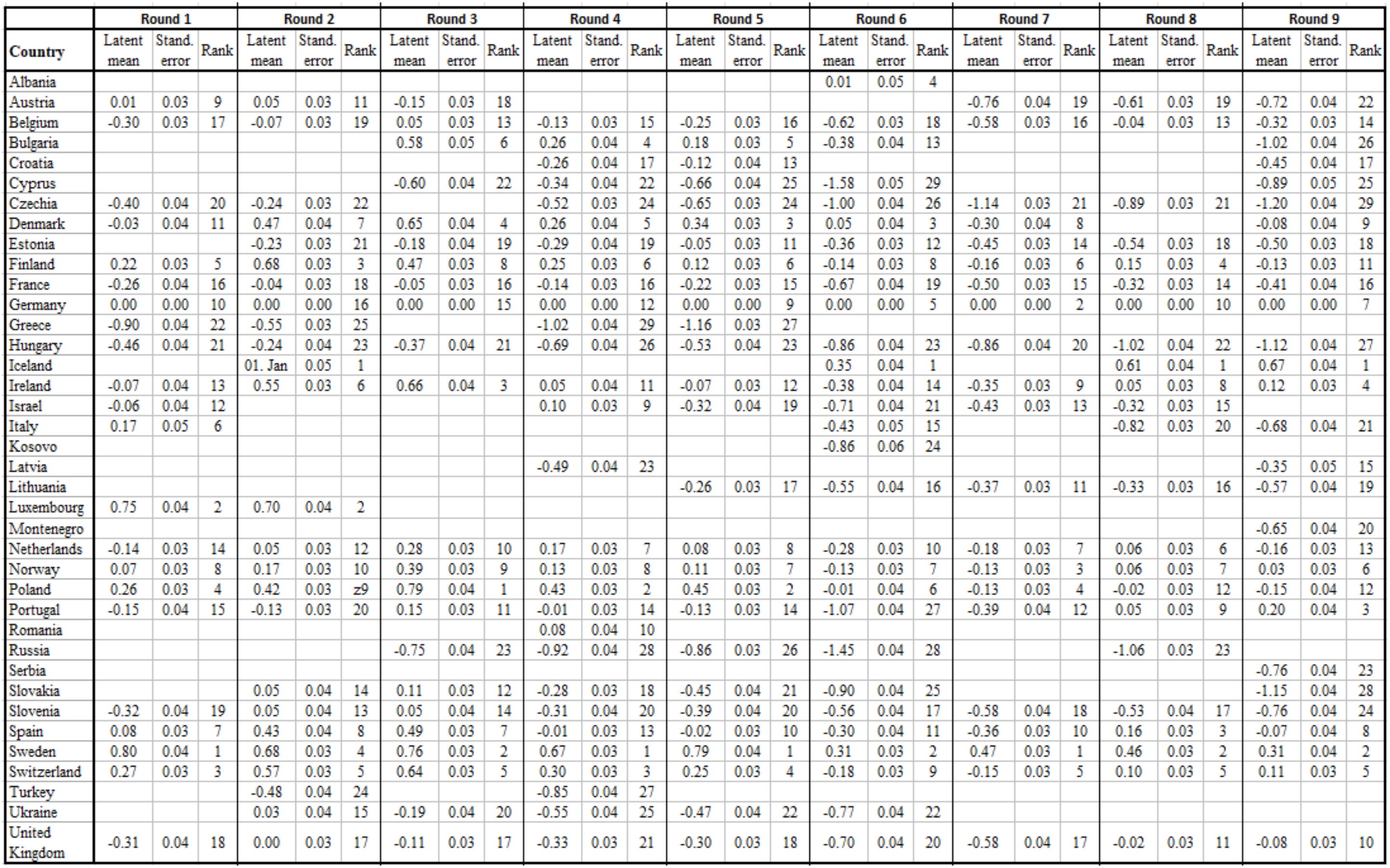
Questions	Response scale	Studies
Would you say that people who come to live here generally take jobs away from workers in [country], or generally help to create new jobs?	0 “take jobs away” to 10 “create new jobs”	Davidov, Cieciuch, and Schmidt 2018: “RT”
Would you say it is generally bad or good for [country]’s economy that people come to live here from other countries?	0 “bad for the economy” to 10 “good for the economy“	
Are [country]’s crime problems made worse or better by people coming to live here from other countries?	0 “crime problems made worse” to 10 “crime problems made better”	
Most people who come to live here work and pay taxes. They also use health and welfare services. On balance, do you think people who come here take out more than they put in or put in more than they take out?	0 “generally take out more” to 10 “generally put in more”	

Table A2: Sample size per country and per ESS round after listwise deletion (N (round 1 – round 9) = 390.276)

	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	Round 9
	2002	2004	2006	2008	2010	2012	2014	2016	2018
Albania						1.086			
Austria	1.941	2.021	2.147				1.664	1.875	2.292
Belgium	1.729	1.716	1.750	1.717	1.680	1.845	1.747	1.750	1.730
Bulgaria			898	1.578	1.806	1.671			1.650
Switzerland	1.893	2.021	1.726	1.698	1.451	1.420	1.480	1.457	1.422
Cyprus			952	1.177	1.020	1.081			740
Czechia	1.051	2.463		1.803	2.143	1.722	1.932	2.135	2.205
Germany	2.697	2.653	2.695	2.614	2.824	2.865	2.965	2.788	2.302
Denmark	1.344	1.383	1.405	1.539	1.507	1.577	1.455		1.511
Estonia		1.615	1.272	1.489	1.615	2.133	1.903	1.946	1.826
Spain	1.431	1.512	1.707	2.320	1.780	1.796	1.715	1.768	1.489
Finland	1.927	1.957	1.850	2.156	1.834	2.152	2.028	1.890	1.717
France	1.453	1.756	1.952	2.008	1.699	1.935	1.868	2.014	1.910
United Kingdom	1.947	1.794	2.297	2.266	2.285	2.158	2.178	1.886	2.139
Greece	2.313	2.280		2.020	2.628				
Croatia				1.304	1.443				1.668
Hungary	1.327	1.261	1.239	1.273	1.325	1.719	1.441	1.381	1.470
Ireland	1.853	2.133	1.682	1.732	2.458	2.534	2.239	2.632	2.141
Israel	2.172			2.230	1.943	2.038	2.215	2.208	
Iceland		538				685		852	828
Italy	1.064					915		2.427	2.566
Lithuania	1.195	1.458							
Luxembourg					1.311	1.733	1.807	1.794	1.541
Latvia				1.750					774

	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	Round 9
	2002	2004	2006	2008	2010	2012	2014	2016	2018
Montenegro									1.142
Netherlands	2.216	1.809	1.800	1.708	1.744	1.763	1.827	1.586	1.582
Norway	1.981	1.721	1.711	1.523	1.516	1.594	1.405	1.505	1.356
Poland	1.715	1.450	1.520	1.390	1.476	1.606	1.348	1.393	1.271
Portugal	1.209	1.804	1.751	1.941	1.877	1.897	1.171	1.190	963
Serbia				1.656					
Romania									1.724
Russia			1.938	2.061	2.195	2.136			
Sweden	1.820	1.809	1.778	1.726	1.413	1.763		2.124	
Slovenia	1.369	1.290	1.325	1.191	1.297	1.149	1.721	1.473	1.485
Slovakia		1.182	1.532	1.533	1.570	1.669	1.092	1.242	1.251
Turkey		1.516		2.077					998
Ukraine		1.439	1.515	1.345	1.435	1.600			
Kosovo						1.054			
Total	37.647	42.581	38.442	50.825	47.275	49.296	37.201	41.316	45.693

Table A3: Latent means, standard errors and rank



	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6			Round 7			Round 8			Round 9		
Country	Latent mean	Stand. error	Rank	Latent mean	Stand. error	Rank	Latent mean	Stand. error	Rank	Latent mean	Stand. error	Rank	Latent mean	Stand. error	Rank	Latent mean	Stand. error	Rank	Latent mean	Stand. error	Rank	Latent mean	Stand. error	Rank	Latent mean	Stand. error	Rank
Albania																0.01	0.05	4									
Austria	0.01	0.03	9	0.05	0.03	11	-0.15	0.03	18										-0.76	0.04	19	-0.61	0.03	19	-0.72	0.04	22
Belgium	-0.30	0.03	17	-0.07	0.03	19	0.05	0.03	13	-0.13	0.03	15	-0.25	0.03	16	-0.62	0.03	18	-0.58	0.03	16	-0.04	0.03	13	-0.32	0.03	14
Bulgaria							0.58	0.05	6	0.26	0.04	4	0.18	0.03	5	-0.38	0.04	13							-1.02	0.04	26
Croatia										-0.26	0.04	17	-0.12	0.04	13										-0.45	0.04	17
Cyprus							-0.60	0.04	22	-0.34	0.04	22	-0.66	0.04	25	-1.58	0.05	29							-0.89	0.05	25
Czechia	-0.40	0.04	20	-0.24	0.03	22				-0.52	0.03	24	-0.65	0.03	24	-1.00	0.04	26	-1.14	0.03	21	-0.89	0.03	21	-1.20	0.04	29
Denmark	-0.03	0.04	11	0.47	0.04	7	0.65	0.04	4	0.26	0.04	5	0.34	0.03	3	0.05	0.04	3	-0.30	0.04	8				-0.08	0.04	9
Estonia				-0.23	0.03	21	-0.18	0.04	19	-0.29	0.04	19	-0.05	0.03	11	-0.36	0.03	12	-0.45	0.03	14	-0.54	0.03	18	-0.50	0.03	18
Finland	0.22	0.03	5	0.68	0.03	3	0.47	0.03	8	0.25	0.03	6	0.12	0.03	6	-0.14	0.03	8	-0.16	0.03	6	0.15	0.03	4	-0.13	0.03	11
France	-0.26	0.04	16	-0.04	0.03	18	-0.05	0.03	16	-0.14	0.03	16	-0.22	0.03	15	-0.67	0.04	19	-0.50	0.03	15	-0.32	0.03	14	-0.41	0.04	16
Germany	0.00	0.00	10	0.00	0.00	16	0.00	0.00	15	0.00	0.00	12	0.00	0.00	9	0.00	0.00	5	0.00	0.00	2	0.00	0.00	10	0.00	0.00	7
Greece	-0.90	0.04	22	-0.55	0.03	25				-1.02	0.04	29	-1.16	0.03	27												
Hungary	-0.46	0.04	21	-0.24	0.04	23	-0.37	0.04	21	-0.69	0.04	26	-0.53	0.04	23	-0.86	0.04	23	-0.86	0.04	20	-1.02	0.04	22	-1.12	0.04	27
Iceland				01. Jan	0.05	1										0.35	0.04	1				0.61	0.04	1	0.67	0.04	1
Ireland	-0.07	0.04	13	0.55	0.03	6	0.66	0.04	3	0.05	0.04	11	-0.07	0.03	12	-0.38	0.04	14	-0.35	0.03	9	0.05	0.03	8	0.12	0.03	4
Israel	-0.06	0.04	12							0.10	0.03	9	-0.32	0.04	19	-0.71	0.04	21	-0.43	0.03	13	-0.32	0.03	15			
Italy	0.17	0.05	6													-0.43	0.05	15				-0.82	0.03	20	-0.68	0.04	21
Kosovo																-0.86	0.06	24									
Latvia										-0.49	0.04	23													-0.35	0.05	15
Lithuania													-0.26	0.03	17	-0.55	0.04	16	-0.37	0.03	11	-0.33	0.03	16	-0.57	0.04	19
Luxembourg	0.75	0.04	2	0.70	0.04	2																					
Montenegro																									-0.65	0.04	20
Netherlands	-0.14	0.03	14	0.05	0.03	12	0.28	0.03	10	0.17	0.03	7	0.08	0.03	8	-0.28	0.03	10	-0.18	0.03	7	0.06	0.03	6	-0.16	0.03	13
Norway	0.07	0.03	8	0.17	0.03	10	0.39	0.03	9	0.13	0.03	8	0.11	0.03	7	-0.13	0.03	7	-0.13	0.03	3	0.06	0.03	7	0.03	0.03	6
Poland	0.26	0.03	4	0.42	0.03	z9	0.79	0.04	1	0.43	0.03	2	0.45	0.03	2	-0.01	0.04	6	-0.13	0.03	4	-0.02	0.03	12	-0.15	0.04	12
Portugal	-0.15	0.04	15	-0.13	0.03	20	0.15	0.03	11	-0.01	0.03	14	-0.13	0.03	14	-1.07	0.04	27	-0.39	0.04	12	0.05	0.03	9	0.20	0.04	3
Romania										0.08	0.04	10															
Russia							-0.75	0.04	23	-0.92	0.04	28	-0.86	0.03	26	-1.45	0.04	28				-1.06	0.03	23			
Serbia																									-0.76	0.04	23
Slovakia				0.05	0.04	14	0.11	0.03	12	-0.28	0.03	18	-0.45	0.04	21	-0.90	0.04	25							-1.15	0.04	28
Slovenia	-0.32	0.04	19	0.05	0.04	13	0.05	0.04	14	-0.31	0.04	20	-0.39	0.04	20	-0.56	0.04	17	-0.58	0.04	18	-0.53	0.04	17	-0.76	0.04	24
Spain	0.08	0.03	7	0.43	0.04	8	0.49	0.03	7	-0.01	0.03	13	-0.02	0.03	10	-0.30	0.04	11	-0.36	0.03	10	0.16	0.03	3	-0.07	0.04	8
Sweden	0.80	0.04	1	0.68	0.03	4	0.76	0.03	2	0.67	0.03	1	0.79	0.04	1	0.31	0.03	2	0.47	0.03	1	0.46	0.03	2	0.31	0.04	2
Switzerland	0.27	0.03	3	0.57	0.03	5	0.64	0.03	5	0.30	0.03	3	0.25	0.03	4	-0.18	0.03	9	-0.15	0.03	5	0.10	0.03	5	0.11	0.03	5
Turkey				-0.48	0.04	24				-0.85	0.04	27															
Ukraine				0.03	0.04	15	-0.19	0.04	20	-0.55	0.04	25	-0.47	0.04	22	-0.77	0.04	22									
United Kingdom	-0.31	0.04	18	0.00	0.03	17	-0.11	0.03	17	-0.33	0.03	21	-0.30	0.03	18	-0.70	0.04	20	-0.58	0.04	17	-0.02	0.03	11	-0.08	0.03	10

Table A4: Measurement quality estimates of the measure of attitudes towards immigration, for each country and round

Country	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	Round 9
Austria	0.79	0.84	0.85	-	-	-	0.87	0.89	0.87
Belgium	0.74	0.77	0.78	0.79	0.79	0.78	0.8	0.8	0.79
Switzerland	0.76	0.81	0.8	0.78	0.76	0.75	0.76	0.81	0.78
Czechia	0.81	0.85	-	0.82	0.86	0.88	0.83	0.83	0.83
Germany	0.78	0.82	0.83	0.83	0.85	0.81	0.84	0.86	0.85
Denmark	0.84	0.86	0.85	0.86	0.87	0.87	0.87	-	0.85
Spain	0.78	0.84	0.83	0.86	0.83	0.85	0.82	0.86	0.86
Finland	0.78	0.82	0.8	0.81	0.84	0.83	0.85	0.85	0.86
France	0.86	0.87	0.89	0.86	0.87	0.87	0.86	0.87	0.87
United Kingdom	0.85	0.89	0.89	0.9	0.89	0.89	0.89	0.89	0.91
Greece	0.83	0.89	-	0.89	0.88	-	-	-	-
Hungary	0.8	0.83	0.84	0.82	0.82	0.86	0.83	0.87	0.88
Ireland	0.85	0.89	0.87	0.87	0.9	0.9	0.87	0.89	0.9
Israel	0.83	-	-	0.87	0.84	0.82	0.83	0.86	-
Italy	0.72	-	-	-	-	0.87	-	0.89	0.9
Luxembourg	0.68	0.76	-	-	-	-	-	-	-
Netherlands	0.72	0.77	0.77	0.76	0.76	0.78	0.76	0.78	0.74
Norway	0.78	0.81	0.8	0.81	0.81	0.81	0.8	0.82	0.83
Poland	0.76	0.73	0.8	0.79	0.78	0.81	0.8	0.78	0.84
Portugal	0.8	0.84	0.81	0.81	0.81	0.85	0.79	0.79	0.79
Sweden	0.81	0.84	0.85	0.84	0.87	0.86	0.86	0.85	0.86
Slovenia	0.74	0.83	0.82	0.83	0.85	0.84	0.83	0.87	0.87
Estonia	-	0.86	0.8	0.82	0.8	0.82	0.83	0.86	0.84
Iceland	-	0.81	-	-	-	0.81	-	0.85	0.86
Slovakia	-	0.75	0.76	0.78	0.8	0.84	-	-	0.81
Turkey	-	0.85	-	0.87	-	-	-	-	-

Country	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	Round 9
Ukraine	-	0.89	0.87	0.87	0.9	0.86	-	-	-
Bulgaria	-	-	0.88	0.87	0.87	0.88	-	-	0.94
Cyprus	-	-	0.77	0.79	0.82	0.8	-	-	0.82
Russia	-	-	0.89	0.87	0.88	0.87	-	0.85	-
Croatia	-	-	-	0.87	0.89	-	-	-	0.86
Latvia	-	-	-	0.84	-	-	-	-	0.84
Romania	-	-	-	0.85	-	-	-	-	-
Lithuania	-	-	-	-	0.82	0.85	0.83	0.86	0.85
Albania	-	-	-	-	-	0.74	-	-	-
-	-	-	-	-	-	0.87	-	-	-
Montenegro	-	-	-	-	-	-	-	-	0.89
Serbia	-	-	-	-	-	-	-	-	0.91

A Scent of Strategy: Response Error in a List Experiment on Anti-Immigrant Sentiment

Sebastian Rinken[1], Sara Pasadas-del-Amo[2] & Manuel Trujillo-Carmona[1]

[1] *Instituto de Estudios Sociales Avanzados (IESA), CSIC, Córdoba, Spain*

[2] *Universidad de Córdoba, Spain*

Abstract

This Research Note reports on a list experiment regarding anti-immigrant sentiment (n=1,965) that was fielded in Spain in 2020. Among participants with left-of-center ideology, the experiment originated a *negative* difference-in-means between treatment and control. Drawing on Zigerell's (2011) deflation hypothesis, we assess the possibility that leftist treatment group respondents may have altered their scores *by more than one* to distance themselves unmistakably from the sensitive item. We consider this possibility plausible in a context of intense polarization where immigration attitudes are closely associated with political ideology. This study's data speak to the results of recent meta-analyses that have revealed list-experiments to fail when applied to prejudiced attitudes and other highly sensitive issues – i.e., precisely the kind of issues with regard to which the technique ought to work best. We conclude that the possibility of strategic response error in specific respondent categories needs to be considered when staging and interpreting list experiments.

Keywords: item-count technique, social desirability bias, strategic response error, immigration attitudes, political polarization

The list experiment, or item-count technique (ICT), aims to obtain unbiased estimates of sensitive behaviors or attitudes. Respondents are divided randomly in treatment and control groups, administered identical lists except for the target item's addition as treatment, and asked *how many,* but not which, items apply to them. The sensitive item's prevalence is estimated by comparing both groups' differences-in-means (DiMs), and the extent of social desirability bias (SDB) assessed by contrast with an equally worded direct question (DQ) (Miller, 1984; Glynn, 2013). This paper dwells on a list experiment on anti-immigrant sentiment that obtained an apparently non-sensical *negative* difference-in-means for some respondents (but not others). Among participants with leftist ideology, the experiment's mean score was significantly *lower* when exposed to treatment (addition of "immigrants" as potentially antipathetic group) than when confronted only with an otherwise identical list of control items. The ensuing aggregate result echoes the findings of a recent meta-analysis that detects *reverse* ICT-DQ differences in studies of prejudiced attitudes (Blair et al., 2020); a second meta-analysis observes disappointing ICT results regarding highly sensitive items (Ehler et al., 2021). Our data offer a rare opportunity for exploring response patterns in specific participant categories, a line of research that might contribute to discerning why list experiments tend to fail precisely when applied to the kind of issues for which they ought to work best.

Background and Objectives

ICT has been employed to gauge the prevalence of ill-regarded behaviors and attitudes such as drug use, risky sex, vote buying, racism, or anti-Semitism, and well-regarded ones such as voting or charitable giving, among many others (Tourangeau & Yan, 2007; Holbrook & Krosnick, 2010; Krumpal, 2013; Blair et al., 2020). Four control items, one each of ample and scarce prevalence and two mutually exclusive ones, are recommended to prevent respondents from considering all items applicable (ceiling), or none (floor), situations that would compromise perceived anonymity (Kuklinski et al., 1997; Blair & Imai, 2012); sensitive controls should be avoided if possible (Droitcour et al., 1991; Ehler et al., 2021). ICT is generally rated as preferable to other unobtrusive survey pro-

Acknowledgements

We would like to thank all EASIE team members as well as IESA's Technical Unit for Applied Studies and Sara Varela (IMOP) for their varied contributions to the survey's successful implementation (cf. https://www.iesa.csic.es/en/proyectos/explaining-placid-attitudes-towards-immigrants-in-spain-easie/).

Direct correspondence to

Sebastian Rinken, Instituto de Estudios Sociales Avanzados (IESA), CSIC, Córdoba, Spain
E-mail: srinken@iesa.csic.es

cedures such as randomized response technique, which guarantees privacy by requesting a score for *either* the sensitive item *or* an unrelated one, for example – petitions that might confuse or even irritate some participants (Coutts & Jann, 2011; Hox & Lensvelt-Mulders, 2008; Rosenfeld et al., 2016; Wolter & Diekmann, 2021). Although list experiments are comparatively straightforward, a growing number of papers have voiced concerns about various kinds of non-strategic response error and ensuing instability (Tsuchiya & Hirai, 2010; Kiewiet de Jonge & Nickerson, 2014; Ahlquist, 2018; Gosen et al., 2019; Kramon & Weghorst, 2019; Jerke et al., 2019; Ehler et al., 2021; Kuhn & Vivyan, 2021; Riambau & Ostwald 2021; Jerke et al., 2022).

The list experiment's most notorious drawback is outsize variance (Miller, 1984; Blair et al., 2020; Ehler et al., 2021); Blair and colleagues (2020) estimate ICT to be 14 times (!) more variable than DQs. Hence, even for considerable differences vis-à-vis obtrusive measures, extremely large samples are required to clear customary significance thresholds. Since this problem is exacerbated in subgroups, little is known about the scope, or even direction, of ICT-DQ comparisons in specific respondent categories (Lax et al., 2016; Blair et al., 2020). A related hitch is relative opacity regarding covariates: vast standard errors arise when regressing ICT results on predictors (Corstange, 2009; Imai, 2011; Blair & Imai, 2012; Glynn, 2013).

Most list experiments obtain reduced bias as compared to obtrusive measurement. Recent meta-analyses conclude that ICT improves estimates of SDB-prone behaviors or mindsets by 8.5 (Ehler et al., 2021) to 10 percentage points (Blair et al., 2020) on average as compared to DQs. However, ICT's performance varies strongly across substantive domains (Blair et al., 2020). Startlingly, the technique has defied expectations with regard to highly sensitive items in general (Ehler et al., 2021) and prejudiced attitudes, in particular (Blair et al., 2020). Blair and colleagues (2020) even find ICT-based prejudice estimates to diverge from DQ-based ones in the *opposite* direction. How may such data be accounted for?

One possible explanation, the reverse polarity of social norms, has been documented in specific contexts, such as nativism in the US (Knoll, 2013), anti-immigrant sentiment in Japan (Igarashi & Nagayoshi, 2022), and vote-buying in Nigeria (Hatz *el al.*, 2023). However, reverse SDB seems implausible with regard to prejudiced attitudes and other highly sensitive items in general (since that proposition would presuppose the reverse polarity of social norms *tout court*), and it cannot possibly explain why treatment respondents mark *lower* scores than their control-group peers.

ICT's rationale relies on encouraging insincere norm violators to alter their score *by one* when faced with the sensitive item. Two crucial assumptions apply (Imai, 2011; Blair & Imai, 2012): sincere scores regarding the sensitive item ("no liars"), and indifference of control item scores to treatment ("no design effect"). Extant scholarship contemplates strategic response error almost exclusively with regard to the experiment's intended addressees (insincere norm violators), hence insisting on optimal anonymity safeguards (cf. ceiling/floor). However, the situation thus created may pose difficulties for respondents keen to distance themselves unequivocally from the sensitive item. This possibility –which seems especially plausible with regard to norm *adherers*– was first observed by Zigerell (2011, p. 553): to prevent any risk of being associated with the treatment item, some respondents may deflate their score "by *any number*", thereby originating *negative* differences between treatment- and control-group scores and distorting aggregate estimates of the sensitive item and related bias. Analogously, respondents keen to send an unmistakable signal of association with a socially desirable treatment item might inflate their scores *by more than one*. Such response behavior would constitute a "design effect" of sorts, yet one deriving from confrontation with the treatment item as such, rather than a flawed choice of controls. Apart from Zigerell's (2011) work on racism, deflation effects have been reported by just a handful of studies, all of which regard strongly polarizing issues such as marijuana use (García-Sánchez & Queirolo, 2020), violent extremism (Clemmow et al., 2020), or anti-immigrant sentiment (Rinken et al., 2021).

This study adds to the extant literature in three ways. First, we document a *negative* difference-in-means between treatment and control among respondents with leftist ideology – a rare opportunity to explore subgroup-level response behavior in a list experiment. Second, we argue that non-strategic error fails to explain why the *longer* list induces *lower* mean scores in this respondent category, but not others. This is important, given that such explanations are favored by the extant literature. Third, by building on Zigerell's (2011) work, we hypothesize various reasons for leftist respondents to deflate their experimental scores *by more than one* in the study context. Negative DiMs in participant subgroups entail an additional rationale, other than and potentially complementary to reverse SDB, for explaining *reverse* aggregate ICT-DQ differences (cf. Blair et al., 2020). Our data highlight the need for further research on the possibility of strategic response error in list experiments on prejudiced attitudes and other highly sensitive items.

Data and Method

A list experiment on anti-immigrant sentiment (AIS) was included in a web survey on native citizens' attitudes toward immigration and immigrants (see online appendix, Figures A1 through A3). Respondents were asked toward how many, among various social groups, they felt antipathy. "Immigrants" were added as treatment to four control items, two of which antagonist (labor unionists and multi-millionaires), one low-prevalence (compulsive gamblers) and one high-prevalence (drug dealers). Control-group respondents were subsequently asked heads-on about antipathy towards immigrants; random assignment to control or treatment ensures the comparability of both estimates. The term “antipathy” refers to the affective core of prejudiced attitudes (Allport, 1954) in a negatively charged way that seems prone to elicit desirability pressures. Hence, our baseline expectation was that the ICT estimate (DiM) would exceed the direct AIS gauge. Control items were chosen based on two pretests, one regarding the entire questionnaire (n=86) and a second one (n=220) focusing on ICT design (see section 1 of the online appendix for details). While the chosen list performed well, some pre-tested options originated negative DiMs – with hindsight, a bellwether of our study’s results.

The survey was administered in 2020 to an online sample of Spanish nationals born and resident in Spain (n=1,965). The sample was selected randomly from a probability-based online panel recruited via random digit dial surveys (see online appendix, Tables A1 and A2). Since we focus on comprehending the response patterns observed in this particular experiment rather than producing population estimates, we use unweighted data in this paper.

Randomization worked well: the covariate profiles of the experiment’s control and treatment arms are almost identical (see online appendix, Table A3). ICT non-response was negligible (1 and 2 persons respectively in treatment and control), and there are few cases at either tail of the item score distribution for both experimental groups, indicating that the experimental design avoided significant ceiling and floor effects (see online appendix, Figure A4). The test for design effects (Blair & Imai, 2012) was passed although a negative proportion is estimated for one respondent type (online appendix, Table A4). This does not prove the absence of design effects (Blair & Imai, 2012): rather, the test did not exclude the possibility of the negative value having arisen by chance. SDB was estimated with R-LIST as difference between a linear-model fit for the ICT and a logit-model fit for the DQ result (Blair & Imai, 2012). Covariates of the ICT-based AIS estimate were modeled by nonlinear least squares (NLS) and maximum likelihood (ML) regressions as implemented in R-LIST; covariates of manifest AIS were modeled as logit regression (Imai, 2011; Blair & Imai, 2012; Blair, Chou & Imai, 2018) (see online appendix for details).

Results

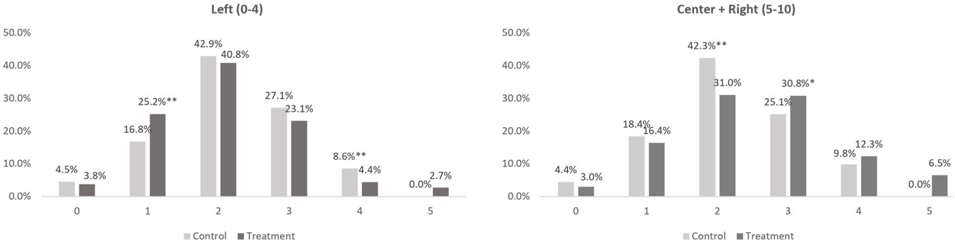
The experiment failed to generate the increased AIS estimate we had anticipated (Table 1). On aggregate, the treatment group’s mean score exceeds the control group’s mean, but the ensuing AIS estimate does not differ significantly from the DQ-based result even when lowering the customary 95% confidence interval (AIS range: 3% to 21.8%) to 90% (range: 4.5% to 20.3%). That said, the ICT-based estimate is actually 3.4 percentage points *lower* than the DQ-based one.

Table 1 Estimates of anti-immigrant sentiment (ICT vs. direct question) and SDB

	Control mean	Treatment mean	ICT estimate (DiM)	DQ estimate	SDB
	2.183	2.308	0.124	0.159	-0.034
S.E.	(0.031)	(0.036)	(0.048)	(0.012)	(0.049)
N	974	988		973	

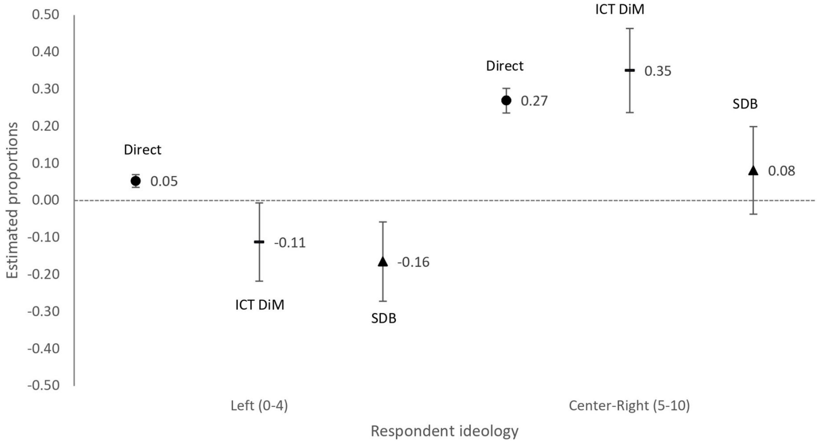
Source: EASIE survey. Abbreviations: ICT=Item-count technique; DiM=difference-in-means between control and treatment; DQ=direct question; SDB=social desirability bias (difference between ICT and DQ-based estimates).

Closer inspection reveals that the experiment generated different response patterns in distinct participant categories. This situation, discernible for several sociodemographic variables including educational attainment and age group, is observed most clearly with regard to political ideology. Treatment participants with centrist or right-of-center ideology mark less 2s and increasing proportions of higher scores (especially 3s) than their control group peers. However, among leftist treatment respondents, the share of 1s increases significantly by comparison to the control group, whereas the proportions of higher scores (especially 4s) decrease (Figure 1). Consequently, among respondents with centrist or right-of-center ideology, our ICT-based estimate of anti-immigrant sentiment exceeds the direct gauge by about 8 percentage points, a non-significant difference. In sharp contrast, an AIS estimate of *minus* 11%, as opposed to 5% in DQ, is obtained for respondents with left-of-center ideology (significant for 90% confidence interval) (Figure 2).



Source: EASIE survey. (Left Total=963, Control= 487 Treatment= 476; Center-right Total= 984, Control= 478, Treatment= 506). Categories of political ideology were derived from self-ratings on a 0-10 scale where ‘0’ means ‘completely leftist’ and ‘10’ means ‘completely rightist’. * p < 0.05; ** p< 0.01.

Figure 1 Item scores in list experiment on anti-immigrant sentiment (unweighted), by respondent ideology



Source: EASIE survey. (Left Total=963, Control= 487 Treatment= 476, DQ=487; Center-right Total= 984, Control= 478, Treatment= 506; DQ=477). Categories of political ideology were derived from self-ratings on a 0-10 scale where '0' means 'completely leftist' and '10' means 'completely rightist'. Bars represent 90% confidence intervals.

Figure 2 Estimates of anti-immigrant sentiment (DQ vs. ICT-DiM) and social desirability bias, by political ideology

Political ideology is a consistent predictor of immigration attitudes (Ceobanu & Escandell, 2010; Hainmueller & Hopkins, 2014; Dražanová, 2022): leftist ideology is generally associated with more benevolent views, and rightist ideology with more restrictive or intolerant ones. In our study, political ideology is associated, net of sociodemographic controls, to both AIS gauges (see online appendix, Table A5). Our study is not powered to assess DiM estimates for each point of the ideological self-rating scale, but those data (see Figure A6 in the online appendix) clearly support the creation of the two groupings (0-4 vs. 5-10) considered here.

Discussion

This study is hampered by ICT's notorious weakness of large variance. The *negative* aggregate difference vis-à-vis direct measurement and the *positive* difference among respondents with centrist or right-of-center ideology both fail to clear any meaningful significance threshold, and the 11-points *negative* difference-in-means between treatment and control among participants with left-of-center ideology is significant only for a 90% confidence interval. This situation might tempt some analysts to dismiss the data as spurious. However, it seems worth noting that our study's aggregate result echoes the *opposite* margin vis-à-vis DQs detected by a recent meta-analysis in list experiments on prejudiced attitudes (Blair et al., 2020); another meta-analysis reveals disappointing ICT results when applied to highly sensitive items (Ehler et al., 2021). While desirability pressures might in some cases be trivial enough for obtrusive measurement to capture such items reasonably well, it seems precipitated to extend that hypothesis to prejudiced attitudes in general (Blair et al., 2020: 1310), and it seems implausible to attribute the inverse relation between item sensitivity and ICT effects (Ehler et al., 2021) to reverse SDB. Negative DiMs in sample subgroups caution against such interpretations. From this perspective, our data offer a welcome opportunity for exploring why ICT seems prone to fail precisely when it ought to work best. Given these circumstances, we consider a suboptimal significance level justified here. Hence, in the following, we will dwell on possible reasons for leftist treatment respondents to mark *lower* mean scores than their control group peers.

Most extant scholarship attributes counter-intuitive or inconclusive ICT data to various kinds of non-strategic response error, such as comprehensibility issues (Kramon & Weghorst, 2019; Jerke et al., 2019), unequal length of lists (Tsuchiya & Hirai, 2010), perceived weirdness (Kuha & Jackson, 2014), or confounding control items (Ehler et al., 2021). We find these explanations unconvincing with regard to our data. Given the negligible incidence of non-response, we see no reason to suspect that the experiment posed excessive cognitive difficulties. Actually, negative DiMs are observed across education levels among leftist respondents (however, large variance keeps these results from attaining statistical significance). If a higher number of items, as such, were to distort results, we see no reason why this should apply only to participants with left-of-center ideology. Similarly, if erratic responses were occasioned by the potentially disconcerting nature of the experimental task (“just how many…”), they should occur regardless of participants’ ideological profiles. In both ideological groupings, response times of treatment participants increased by almost identical margins (4.9 and 4.7 seconds, respectively) as compared to controls (see online appendix, Table A6); given the need to consider a higher number of items, such an increase should be expected. However, respondent ideology might come into play with regard to control items. To prevent ceiling and floor effects, lists are required to contain two mutually exclusive items (Kuklinski et al., 1997; Blair & Imai, 2012). When inquiring about antipathy toward a varied assortment of social groups, it seems inevitable that one such might be perceived as sensitive by some respondents; specifically, in our study, some leftist respondents might have been reluctant to admit antipathy toward labor unionists. If so, though, both experimental groups should be similarly affected by such reluctance. Therefore, we do not see how the treatment arm’s *lower* mean could derive from desirability pressures regarding labor unionists.

Bearing in mind that the experiment is exactly the same for all participants, *except for inclusion of an additional item as treatment*, confrontation with this item offers the most straightforward explanation for any differential response pattern vis-à-vis control. Indeed, the expectation that list experiments ought to originate improved prevalence estimates of sensitive behaviors and attitudes is predicated on this premise: norm violators are supposed to alter their score by one (by comparison to analogous control group participants) when faced with the treatment item, whereas all other treatment respondents are supposed to be unaffected by the sensitive item. However, treatment participants who fervently adhere to the norm might react in unanticipated ways, as might stubbornly insincere norm violators. The possibility that the experimental situation might originate strategic response error has played a subdued role in the scholarly debate thus far. In an apparent nod to Zigerell's (2011) work on racism, Blair et al. (2020: 1310) acknowledge passingly "that the list experiment (might) not provide the cover it is designed to provide in this context", yet do not elaborate any further.

The experimental situation's opacity ("*just how many*") is meant to encourage insincere norm violators to lower their guard. Zigerell (2011) argued that this very opacity may prove challenging for respondents aiming to send a clear signal of dissociation from a negatively charged item. He hypothesized that such respondents may alter their score *by more than one*, thereby originating a negative DiM by comparison to their control-group peers (an analogous logic of "overacting" may apply to positively charged treatment items). Such deflation effects presuppose very intense desirability pressures, as was the case with Zigerell's data on racism in the U.S. Because unwelcoming attitudes toward immigrants are prone to be interpreted as telltale of racist or xenophobic views (Esses et al., 1998; Wilkes et al., 2008), the possibility of similarly intense desirability dynamics seems worth considering here. With regard to AIS in Spain, deflation effects have been documented among self-declared xenophiles (Rinken et al., 2021), who are by definition keen to distance themselves from anti-immigrant prejudice. Since attitudes toward immigration and immigrants correlate strongly with political ideology (in our dataset, correlation coefficients exceed 0.38 for various ATII gauges), it seems fair to assume that leftist respondents and self-declared xenophiles react similarly to a list experiment on AIS. However, in our study, negative DiMs are statistically significant for leftist respondents but not for xenophile ones; this situation suggests some additional factor driving leftist participants' response behavior.

The empirical context of our study entails discernible incentives, apart from and beyond xenophile attitudes, for leftist respondents to seek clear dissociation from AIS. For the first time since the Franco dictatorship's demise, a radical-right party featuring anti-immigrant rhetoric had recently scored significant electoral gains across Spain (Ferreira, 2019; Mendes & Dennison, 2021; Turnbull-

Dugarte et al., 2020). Consequently, immigration-related issues became supercharged ever more intensely by broader questions of ideological allegiance. This context is reflected by intensifying polarization of survey data on immigration attitudes (González Enríquez & Rinken, 2021): in direct measurement, right-wing respondents manifest increasingly unfavorable views, whereas leftwingers state increasingly favorable positions. Such data might reflect genuine trends (souring or improved attitudes, respectively), but enhanced desirability pressures might play a role too. To participants with right-wing ideology who pay lip-service to anti-immigrant rhetoric in DQ, the list experiment offers the coverage needed for revealing their true feelings.

In contrast, leftists are in a bind. In our survey's context of intense ideological polarization, it seems plausible to assume that the experimental situation might be experienced as inconveniently opaque by some leftist participants. Whatever their mindset regarding immigration and immigrants, this context makes it tempting for leftist treatment-group respondents to distance themselves unmistakably from a sensitive item that is routinely tagged, in their ideological ecosystem, as deplorable epitome of right-wing extremism. The ensuing scores do not reveal true feelings: leftist treatment respondents might opt to deflate their scores either because of particularly strong xenophile convictions, or else due to an intense wish to appear to be sharing such convictions. Our data offer no insight about the relative importance of either, but raw scores do indicate a constraint (cf. Figure 1): an overwhelming majority mark scores higher than zero. Thus, the urge to unmistakably flag anti-racist convictions does not propel leftist treatment respondents to induce any doubt about their disdain for drug-dealers. Also, it seems worth noting that the data indicate a minimum level of deflation behavior, rather than measuring its exact extent: a negative difference-in-*means* is observed net of the *increased* scores that some leftist participants may have marked when faced with the treatment item.

Who might such advertisements of norm compliance be directed to? Response behavior in survey settings is meaningful only with regard to an (imaginary or tangible) audience. Most SDB studies have considered external audiences, such as interviewers or bystanders; however, recent research retrieves interest in the self as ever-present and potentially decisive audience (Blair et al., 2020; Brenner, 2020). In our panel-based data, survey administrators cannot be discarded as salient social referent (Coutts & Jann, 2011) – be it to safeguard one's xenophile credentials, or else to counterfactually exhibit politically correct attitudes. Yet, the experimental situation may also induce respondents to "edit their report for their own benefit; that is, for their own view of themselves" (Brenner, 2020: 49). In a context of strong polarization regarding immigration-related issues, it seems plausible that ideological self-identifications may claim center stage. Thus, leftist treatment respondents may alter their list scores *by more than one* either to burnish a genuine self-image of benevolence towards immigrants, or else to dispel the dissonant chord (Festinger, 1957) struck by the sensitive item with regard to their overall ideology. Cross-tabulation of both parameters (relation with the pro-immigrant norm, on one hand, and projected audience, on the other) originates a tentative taxonomy of deflation motives that might benefit future attempts at refining their conceptualization (see online appendix, Table A7).

Conclusion

This exploratory study aims to stimulate further research on the methodological properties of list experiments. Apart from heeding the recommendation to field future list experiments with extremely large samples, survey methodologists and practitioners interested in highly sensitive issues should consider two related possibilities: (a) inconclusive or counter-intuitive aggregate data might stem from divergent response patterns in participant subgroups, and (b) strategic response error might contribute to their explanation.

Funding

This work was supported by grant CSO2017-87364-R, funded by the Spanish Ministry of Science and Innovation (MCIN/ AEI /10.13039/501100011033/) and the European Regional Development Fund (ERDF "A way of making Europe").

Ethics approval

The dataset used in this study was generated by a survey approved by the Spanish Research Council' Ethics Committee (reference nº 127/2020).

Data availability statement

The dataset is available at CSIC's institutional open-access repository (cf. Rinken et al., 2023).

References

Ahlquist, J. S. (2017). List experiment design, non-strategic respondent error, and item count technique estimators. *Political Analysis, 26,* 34–53. https://doi.org/10.1017/pan.2017.31

Allport, G. (1954). *The nature of prejudice*. Reading: Addison-Wesley.

Blair, G., & Imai, K. (2012). Statistical analysis of list experiments. *Political Analysis, 20*, 47–77. https://doi.org/10.1093/pan/mpr048

Blair, G., Chou, W., & Imai, K. (2019). List experiments with measurement error. *Political Analysis, 27*, 455-480. https://doi.org/10.1017/pan.2018.56

Blair, G., Coppock, A., & Moor, M. (2020). When to worry about sensitivity bias: A social reference theory and evidence from 30 years of list experiments. *American Political Science Review. 114*, 1297–1315 https://doi.org/10.1017/s0003055420000374

Brenner, P. S. (2020). Advancing theories of socially desirable responding: How identity processes influence answers to 'sensitive questions'. In P. S. Brenner (Ed.), *Understanding survey methodology: sociological theory and applications* (pp. 45–65). Cham: Springer. https://doi.org/10.1007/978-3-030-47256-6_3

Ceobanu, A. M., & Escandell, X. (2010). Comparative analyses of public attitudes toward immigrants and immigration using multinational survey data: A review of theories and research. *Annual Review of Sociology, 36,* 309–28. https://doi.org/10.1146/annurev.soc.012809.102651

Clemmow, C., Schumann, S., Salman, N. L., & Gill, P. (2020). The base rate study: Developing base rates for risk factors and indicators for engagement in violent extremism. *Journal of Forensic Sciences, 65*, 865–81. https://doi.org/10.1111/1556-4029.14282

Corstange, D. (2009). Sensitive questions, truthful answers? Modeling the list experiment with LISTIT. *Political Analysis, 17,* 45–63. https://doi.org/10.1093/pan/mpn013

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods & Research, 40*(1), 169–93.
https://doi.org/10.1177/0049124110390768

Dražanová, L. (2022). Sometimes it is the little things: A meta-analysis of individual and contextual determinants of attitudes toward immigration (2009–2019). *International Journal of Intercultural Relations, 87,* 85–97.
https://doi.org/10.1016/j.ijintrel.2022.01.008

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Thiowetz, & S. Sudman (eds.) *Measurement errors in surveys* (pp. 185–210). John Wiley & Sons. https://doi.org/10.1002/9781118150382.ch11

Ehler, I., Wolter. F., & Junkermann, J. (2021). Sensitive questions in surveys. A comprehensive meta-analysis of experimental survey studies on the performance of the item count technique. *Public Opinion Quarterly, 85*(1), 6–27. https://doi.org/10.1093/poq/nfab002

Esses, V., Jackson, L., & Armstrong, T. (1998). Intergroup competition and attitudes toward immigrants and immigration: An instrumental model of group conflict. *Journal of Social Issues, 54*(4), 699-724. https://doi.org/10.1111/j.1540-4560.1998.tb01244.x

Ferreira, C. (2019). Vox as representative of the radical right in Spain: A study of its ideology. *Revista Española de Ciencia Política,* (51), 73–98. https://doi.org/10.21308/recp.51.03

Festinger, L. (1957). *A theory of cognitive dissonance,* Stanford, CA: Stanford University Press. https://doi.org/10.1515/9781503620766

García-Sánchez, M., & Queirolo, R. (2020). A tale of two countries: The effectiveness of list experiments to measure drug consumption in opposite contexts. *International Journal of Public Opinion Research, 33*(2), 255–72. https://doi.org/10.1093/ijpor/edaa031

Glynn, A. N. (2013). What can we learn with statistical truth serum? Design and analysis of the list experiment. *Public Opinion Quarterly, 77*(S1), 159–72.
https://doi.org/10.1093/poq/nfs070

González Enríquez, C., & Rinken, S. (2021). Spanish public opinion on immigration and the effect of VOX. ARI 46/2021. Madrid: Real Instituto Elcano.

Gosen, S., Schmidt, P., Thörner, S., & Leibold, J. (2019). Is the list experiment doing its job? In J. Mayerl, T. Krause, A. Wahl, & M. Wuketich (eds.), *Einstellungen und Verhalten in der empirischen Sozialforschung: Analytische Konzepte, Anwendungen und Analyseverfahren* (pp. 179–205). Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-16348-8_8

Hainmueller, J., & Hopkins, D. J. (2014). Public attitudes toward immigration. *Annual Review of Political Science, 17*(1), 225–49. https://doi.org/10.1146/annurev-polisci-102512-194818

Hatz, S., Fjelde, H. & Randahl, D. (2023). Could vote buying be socially desirable? Exploratory analyses of a 'failed' list experiment. *Quality & Quantity*. https://doi.org/10.1007/s11135-023-01740-6

Holbrook, A. L., & Krosnick, J. A. (2010). Social desirability bias in voter turnout reports tests using the item count technique. *Public Opinion Quarterly, 74*(1), 37–67. https://doi.org/10.1093/poq/nfp065

Hox, J. & Lensvelt-Mulders, G. (2008). Randomized response. In Lavrakas, P. J. *Encyclopedia of survey research methods*. Sage publications. https://dx.doi.org/10.4135/9781412963947

Igarashi, A., & Nagayoshi, K. (2022). Norms to be prejudiced: List experiments on attitudes towards immigrants in Japan. *Social Science Research*, 102, 102647. https://doi.org/10.1016/j.ssresearch.2021.102647

Imai, K. (2011). Multivariate regression analysis for the item count technique. *Journal of the American Statistical Association, 106*(494), 407–16. https://doi.org/10.1198/jasa.2011.ap10415

Jerke, J., Johann, D., Rauhut, H., & Thomas, K. (2019). Too sophisticated even for highly educated survey respondents? A qualitative assessment of indirect question formats for sensitive questions. *Survey Research Methods, 13*, 319–51. https://doi.org/10.18148/srm/2019.v13i3.7453

Jerke, J., Johann, D., Rauhut, H., Thomas, K., & Velicu, A. (2022). Handle with care: implementation of the list experiment and crosswise model in a large-scale survey on academic misconduct. *Field Notes, 34*(1), 69–81. https://doi.org/10.1177/1525822x20985629

Kiewiet de Jonge, C. P., & Nickerson, D. W. (2014). Artificial inflation or deflation? Assessing the item count technique in comparative surveys. *Political Behavior, 36*(3), 659–82. https://doi.org/10.1007/s11109-013-9249-x

Knoll, B. R. (2013). Implicit nativist attitudes, social desirability, and immigration policy preferences. *International Migration Review*, 47(1), 132-165. http://dx.doi.org/10.1016/j.ssresearch.2013.07.012

Kramon, E., & Weghorst, K. (2019). (Mis)Measuring sensitive attitudes with the list experiment. *Public Opinion Quarterly, 83*(S1), 236–63. https://doi.org/10.1093/poq/nfz009

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity, 47*, 2025–47. https://doi.org/10.1007/s11135-011-9640-9

Kuha, J., & Jackson, J. (2014). The item count method for sensitive survey questions: Modelling criminal behaviour. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 63(2)*, 321-341. https://doi.org/10.2139/ssrn.2119238

Kuhn, P. M., & Vivyan. N. (2021). The misreporting trade-off between list experiments and direct questions in practice: Partition validation evidence from two countries. *Political Analysis*, published online April 16, 2021. https://doi.org/10.1017/pan.2021.10

Kuklinski, J. H., Cobb, M. D., & Gilens, M. (1997). Racial attitudes and the 'New South'. *The Journal of Politics, 59*(2), 323–49. https://doi.org/10.2307/2998167

Lax, J. R., Phillips, J. H., & Stollwerk, A. F. (2016). Are survey respondents lying about their support for same-sex marriage? Lessons from a list experiment. *Public Opinion Quarterly, 80*(2), 510–33. https://doi.org/10.1093/poq/nfv056

Mendes, M. S., & Dennison, J. (2021). Explaining the emergence of the radical right in Spain and Portugal: Salience, stigma and supply. *West European Politics, 44*(4), 752–75. https://doi.org/10.1080/01402382.2020.1777504

Miller, J. D. (1984). A new survey technique for studying deviant behavior. PhD Thesis, George Washington University.

Riambau, G., & Ostwald, K. (2021). Placebo statements in list experiments: Evidence from a face-to-face survey in Singapore. *Political Science Research and Methods, 9*(1), 172–79. https://doi.org/10.1017/psrm.2020.18

Rinken, S., Pasadas-del-Amo, S., Rueda, M., & Cobo, B. (2021). No magic bullet: Estimating anti-immigrant sentiment and social desirability bias with the item-count technique. *Quality & Quantity, 55,* 2139–59. https://doi.org/10.1007/s11135-021-01098-7

Rinken, S., Buraschi, D., Domínguez Álvarez, J. A., Godenau, D., González Enríquez, C., Lafuente, R., ... Varela, S. (2023). Survey on attitudes toward immigration and immigrants in Spain (EASIE survey) [Data set]. DIGITAL.CSIC. https://doi.org/10.20350/DIGITALCSIC/15586

Rosenfeld, B., Imai, K., & Shapiro, J.N. (2016). An empirical validation study of popular survey methodologies for sensitive questions. *American Journal of Political Science, 60*(3), 783–802. https://doi.org/10.1111/ajps.12205

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859-883. https://doi.org/10.1037/0033-2909.133.5.859

Tsuchiya, T., & Hirai, Y. (2010). Elaborate item count questioning: Why do people underreport in item count responses? *Survey Research Methods, 4*(3), 139-149. https://doi.org/10.18148/srm/2010.v4i3.4620

Turnbull-Dugarte, S. J., Rama, J., & Santana, A. (2020). The Baskerville's Dog suddenly started barking: Voting for VOX in the 2019 Spanish general elections. *Political Research Exchange* 2(1), 1781543. https://doi.org/10.1080/2474736x.2020.1781543

Wilkes, R., Guppy, N., & Farris, L. (2008). No thanks, we're full: Individual characteristics, national context, and changing attitudes toward immigration. *International Migration Review, 42*(2), 302-329. https://doi.org/10.1111/j.1747-7379.2008.00126.x

Wolter, F. and Diekmann, A. (2021). False positives and the 'more-is-better' assumption in sensitive question research. *Public Opinion Quarterly, 85*(3), 836–63. https://doi.org/10.1093/poq/nfab043

Zigerell, L. J. (2011). You wouldn't like me when I'm angry: List experiment misreporting. *Social Science Quarterly, 92*(2), 552–62. https://doi.org/10.1111/j.1540-6237.2011.00770.x

Exploring Respondents' Problems and Evaluation in a Survey Proposing Voice Inputs

Melanie Revilla[1] & Mick P. Couper[2]

[1] *Institut Barcelona d'Estudis Internacionals (IBEI)*

[2] *Survey Research Center, Institute for Social Research, University of Michigan*

Abstract

Integrating voice inputs into web surveys holds the potential for various benefits, including eliciting more comprehensive and elaborate responses or extracting additional information from vocal tones and ambient sounds. Nevertheless, important challenges persist, including technical problems, privacy concerns, and low participation rates. Given the limited knowledge on this subject, this research note addresses four research questions, distinguishing between two voice input methods (dictation and voice recording) and two approaches to presenting them (providing a choice, or pushing respondents toward voice inputs, with a text alternative offered only in the absence of response): *RQ1***.** What reasons are provided for not opting for voice inputs when they are offered? *RQ2***.** Which variables are associated with the reported use of voice inputs? *RQ3***.** What challenges do individuals answering through voice inputs report? And *RQ4***.** How do respondents evaluate the different methods of answering they employed?

Drawing on data from a survey on nursing homes conducted in February/March 2023 within the Netquest opt-in online panel in Spain (1,001 completes), where participants were offered to respond to two experimental questions through voice methods, our analyses reveal that contextual factors and the perceived challenge of oral expression are key reasons for abstaining from voice input responses. Furthermore, individuals who exhibited complete trust in the confidentiality of their responses and those already using voice input in their daily lives were significantly more likely to opt for voice inputs. Among respondents utilizing voice inputs, recurring challenges included contextual constraints and difficulties in verbal expression, alongside technical problems. Despite these hurdles, a majority of participants found answering through voice easy, although a lower proportion reported liking it. These results contribute to the limited literature and can help enhance the effectiveness of voice input surveys.

Keywords: dictation, survey question evaluation, open questions, challenges, voice recording, web surveys

In recent years, the integration of voice input technology into everyday activities has become increasingly common (Deloitte, 2018). Simultaneously, an increasing number of surveys have embraced this technology to collect responses for specific questions, typically open-ended narrative questions (see “Background section”).

It has been argued that offering voice input in web surveys could present a variety of potential benefits, such as eliciting richer and longer answers or permitting the extraction of additional information from nuances in the tone of voice or from ambient noise (Höhne et al., 2023; Revilla, 2022; Singer & Couper, 2017). Nevertheless, persistent challenges, including technical issues, data protection and privacy concerns, and low participation rates (see e.g., Revilla & Couper, 2021), underscore the need for a comprehensive understanding of the factors influencing engagement with voice input for open questions. Gaining insights into how respondents perceive these novel methods of answering is essential for enhancing their overall effectiveness.

This research note presents the outcomes of a web survey (both mobile and PC devices were allowed) on opinions about nursing homes conducted in February/March 2023 within the Netquest opt-in online panel in Spain (N=1,001 completes). It focuses on the problems and challenges encountered by participants, as well as their evaluations, when two voice inputs are proposed as a response method for open narrative questions[1]:

- *Dictation* (also called *Automatic Speech Recognition* or *ASR*): Respondents speak, and their voice is instantly transcribed into text on their device’s screen. Respondents can then edit the transcriptions using their keyboard.

[1] Data from this same survey have also been used in a different paper focused on comparing participation and data quality across the different experimental groups presented in Table 1.

Funding

This project received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 849165).

Data availability

The anonymized dataset, R script, and supplementary online materials (SOM) are accessible at: https://osf.io/3crsg/?view_only=52bc495d5007463faa8a6e56bad9bf97

Acknowledgements

We are very grateful to Ixchel Perez Duran for all her work in designing the questionnaire, and to Maria Paula Acuña Pardo, Oriol Bosch, Patricia Iglesias, Ksenija Ivanovic, Carlos Ochoa and Tracy Silva for their assistance at different stages of this study’s preparation. We also want to thank two anonymous reviewers and the journal editor who provided helpful feedback on an earlier draft of this manuscript.

Direct correspondence to

Melanie Revilla, Institut Barcelona d’Estudis Internacionals (IBEI)
E-mail: mrevilla@ibei.org

- *Voice recording*: Respondents are asked to record their voice. They can create and review multiple audio files before submitting their responses.

Furthermore, two approaches to propose these voice inputs are compared:

- *Push*: Respondents are initially presented with only one of the voice input methods. If they skip the question without answering, the question is repeated with a message emphasizing the importance of their responses, and a text alternative.
- *Choice*: Respondents are offered three options: answering by typing in a text-box, by dictating, or by recording their voice. They can choose whichever they prefer, and can use multiple methods.

Background

Some studies have explored respondents' stated willingness to use voice input to answer survey questions (Höhne, 2021; Lenzner & Höhne, 2022; Revilla et al., 2018). Others have actually asked respondents to answer open-ended narrative questions through voice input, using experimental designs. For instance, studies by Lütters and colleagues (2018) and Meitinger and Schonlau (2022) randomly assigned participants to a voice-only group, a choice group (allowing selection between voice or text), and a text-only group. Other studies compared voice recording and text responses (Gavras, 2019; Gavras & Höhne, 2022; Gavras et al., 2022; Höhne & Gavras, 2022). Revilla et al. (2020) compared text with dictation for iOS respondents and text with voice recording for Android respondents.

The findings of these studies indicate that participation tends to be lower when respondents are offered voice input methods, even when given the option to choose between voice and text. For instance, in the study by Lütters et al. (2018), 49% of the participants answered in the voice-only group, and 54% in the choice group, compared with 94% in the text-only group. Further, in cases where a choice is available, a significant majority of participants opted for the text option (e.g., 93.9% in Meitinger et al., 2022).

Nevertheless, there are indications that voice answers could have higher quality, with significantly longer answers and a greater variety of words than text responses (e.g., Höhne & Gavras, 2022). Also, certain underrepresented groups (e.g., older or lower-educated individuals) may be encouraged to respond to open-ended questions when voice inputs are proposed (Gavras, 2019).

Some studies also explored respondents' evaluations and experiences, finding that participants are more positive about text than voice answers (Lütters et al., 2018; Revilla et al., 2020).

In addition, previous research suggests that participation, data quality and respondents' evaluation of voice input methods might be affected by participants' characteristics. For example, Revilla and Couper (2021) found that gender, education, mother tongue, using voice input in daily life, trust in anonymity, multitasking, and answering from home significantly affected at least one of their dependent variables related to nonresponse, data quality and evaluation of voice recording.

Finally, Revilla and Couper (2021) tried to improve the voice input option on Android devices. Providing different instructions to help respondents using the voice recording tool had minimal impact on uptake rates. A filter question to determine whether respondents were in a setting that permitted voice recording, directing others to text input, was more successful. However, technical issues and low participation persisted.

Overall, the available studies remain sparse, and in particular, little is known about possible differences between dictation and voice recording, and between different approaches to presenting the voice input options to participants.

Research Questions and Contribution

To fill these gaps, this research note addresses four research questions regarding the integration of voice inputs for responding to open-ended narrative questions:

RQ1) What reasons are provided for not using voice inputs when they are offered?

RQ2) Which variables are associated with the reported use of voice inputs?

RQ3) What challenges do individuals answering through voice inputs report?

RQ4) How do respondents evaluate different methods of answering open questions?

This study contributes to the limited literature on utilizing voice inputs in web surveys in several ways. Firstly, it provides fresh empirical evidence on two distinct voice input methods: dictation and voice recording.

While both are voice input methods, voice recording has been studied more frequently than dictation. Besides, the methods exhibit some key differences that may affect respondents' experience with and evaluation of the methods. Notably, although respondents can review their answers in both methods, the process differs: editing the transcription versus recording a full answer again. Furthermore, privacy concerns can be less prevalent for dictation than for voice recording, since the voice file is not shared with the fieldwork company or researchers, fostering a sense of confidentiality. The cognitive load can also differ since in one case, visual support can be provided and answers can be reviewed by reading while in the other respondents can only listen to the audio files. Thus, we expect that different reasons could be provided for not using the two kinds of voice inputs (e.g., more aspects related to privacy issues could be mentioned in voice recording) and that different variables could be associated with participation in questions proposing dictation versus recording. Similarly, differences are expected in the prevalence of the problems faced by the participants and in their evaluation of such methods.

Second, this study contributes by distinguishing between two approaches of offering the dictation and voice recording options (*Push* and *Choice*).

The way of offering the voice input options could affect the results to the different research questions: in particular, the "choice" method is expected to yield fewer reported problems/challenges and slightly more positive overall evaluations, since participants can select what they prefer.

Third, this is the first study to collect voice data through the *WebdataVoice* tool (Revilla et al., 2022), which allows for either dictation or voice recording on Android and iOS devices as well as PCs and has been designed to be user-friendly. Using this new tool could produce more favorable results compared to previous studies, especially fewer technical and understanding problems, which in turn could lead to more positive evaluations.

Overall, insights from this study can help researchers and survey designers tailor voice input surveys to mitigate reported problems/challenges and enhance participant evaluations.

Method and Data

Data Collection

Data were collected between February 22 and March 30, 2023, in the Netquest online opt-in panel in Spain. The objective was to obtain 1,000 participants completing the full survey. Quotas for gender and age, education, and autonomous community were defined to match the adult online population in Spain (under 75 years old) according to the National Statistics Institute.

Of the 4,789 panelists invited, 1,860 started the survey. Of those, 577 were excluded for various reasons (170 did not provide consent, 185 quotas full, 17 did not pass basic fraud checks and 205 reported unfamiliarity with nursing homes), leading to 1,170 panelists answering the first survey question after all the filter/quota questions. Another 169 panelists broke-off during the survey, meaning that 1,001 completed the full survey. The average age of those finishing the survey is 47 years old, 50.5% are female, and 35.0% have a higher education degree. On average they have been in the Netquest panel for six years (median=5.7), and have completed 157 surveys (median=141). Most participants used smartphones (73.6%) to respond. The average survey completion time was 9.1 minutes.

Questionnaire

The online questionnaire included more than 80 questions optimized for mobile devices but accessible from any device. None of the respondents got all questions, due to routing. The full questionnaire in Spanish and its English translation are available in the Supplementary Online Material (SOM) 1.

Respondents could continue without answering the questions, except those used to control quotas and filter/tailor other questions. Following the panel's usual practice, going back was not allowed.

The survey mainly dealt with perceptions of nursing homes in Spain (e.g., to what extent they trust them or consider that they are transparent) but also included a block of questions about political opinions (e.g., trust in the government), as well as sociodemographic questions (e.g., mother tongue), questions about the context in which respondents answered the survey (e.g., presence of third parties) and about their evaluation of some questions (e.g., how easy or difficult it was to answer open-ended questions using different methods).

The survey included the following two narrative open-ended questions asking respondents to explain why they selected a given answer in the previous question:

- ***WHYTRANSP.*** Explain why you think that nursing homes provide [no information at all/very little information/some information/a lot of information/a huge amount of information[2]] about the implementation of their services. Please give as much detail as you can. In your answer, mention if you think there is a difference among **public** and **private** nursing homes.
- ***WHYTRUST.*** Explain why you personally [not at all/very little/somewhat/very much/completely] **trust** nursing homes. Please give as much detail as you can. In your answer, mention if you think there is a difference among **public** and **private** nursing homes.

For these two questions, an experimental design was used: respondents were assigned to four groups, as presented in Table 1: a *Control* group, two "push" groups (*PushDictation* and *PushRecording*) and a *Choice* group where all three options were offered. Detailed instructions for both experimental questions can be found in SOM1. Screenshots of these questions (together with the question just before and the follow-up when relevant) for each of these groups are provided in SOM2.

[2] This was tailored for each respondent depending on the previous answer.

Table 1 Experimental groups (same group for both WHYTRANSP and WHYTRUST)

Control	PushDictation	PushRecording	Choice
Text answers only.	Propose dictation, if they do not answer, also offer text.	Propose recording, if they do not answer, also offer text.	Choice between: ▪ Dictation ▪ Recording ▪ Text

In this research note, we are mainly interested in questions asking respondents a) their reasons for not using voice input methods to answer ***WHYTRANSP*** and ***WHYTRUST***, b) which kinds of problems they faced to use these answering methods and c) how they evaluate these new ways of answering and the conventional (text) one[3]. We also use questions about the respondents' profile (socio-demographics and attitudinal variables) to answer ***RQ2*** (see below).

Analyses

To answer ***RQ1***, we report the answers to a question asking respondents[4] to select all that apply of the reasons for not using voice inputs in the following list: "I preferred another of the alternatives" (only in *Choice* group), "The device I am using to answer the survey does not have a microphone", "I tried, but I had technical problems", "I tried, but I had problems understanding the function", "I did not want to use it because of the context (e.g., I was around other people)", "I did not want to use it because I found it difficult to express myself orally", "Other reasons". The proportions of panelists selecting each option are reported for both dictation and voice recording, separating the push from the choice groups.

To assess whether there are differences between dictation and voice recording, we compare:

- *PushDictation* to *PushRecording*
- *ChoiceDictation* (i.e., respondents from the *Choice* group who have stated they used dictation – whether alone or in combination with other methods) to *ChoiceRecording* (i.e., respondents from the *Choice* group who have stated they used recording – whether alone or in combination with other methods).

[3] Another narrative open-ended question asking about the perceived quality of the nursing homes was presented to the panelists. This question was placed before the two experimental ones, and all respondents were asked to answer it using a text-box.

[4] This question was asked only to those who stated "No, I never used the dictation/voice recording tool" in the questions USEDDICTATION/ USEDVOICE (see SOM1 and Appendix A).

To assess whether there are differences between push and choice groups, we compare:

- *PushDictation* to *ChoiceDictation*
- *PushRecording* to *ChoiceRecording*.

We test whether differences in proportions across groups are significant at the 5% level using exact Fisher tests.

To answer ***RQ2***, logistic regressions analyses were conducted. The dependent variables are the use of dictation or voice recording reported in the questions USEDDICTATION and USEDVOICE (see Appendix A), grouping the two "yes" options to create indicators where 1 indicates that dictation or voice recording has been used, and 0 otherwise.

The key independent variable is the experimental group: push or choice (push being used as reference category). We control for the following sociodemographic characteristics: gender, age (two dummies for respondents having less than 30 and more than 60), and education level (two dummies for low and high education).

Additionally, based on previous research (Revilla & Couper, 2021) but also, since little is known yet, logical reasoning about which factors might influence the reported use of dictation and voice recording and data availability, we include the following set of independent variables:

- Having Spanish as a mother tongue (dummy): Non-native speakers might exhibit more reluctance to answer through voice options (e.g., because of concerns about their accent).
- Social trust (values ranging from "1-You can't be too careful" to "5-Most people can be trusted") and trust in the confidentiality of answers (dummy, 1 = complete trust and 0 = the rest): Higher levels of trust may be associated with lower privacy concerns, and, consequently, increased use of voice inputs.
- Comfort in using new technologies (dummy, 1 = "quite" to "completely comfortable", and 0 = "not at all" or "little comfortable"): Being comfortable with new technologies is expected to be associated with higher participation through voice inputs.
- Lack of awareness of voice inputs existence (one dummy for each type of voice input) and occasional use of voice inputs in daily life (one dummy for each type of voice input[5]): Distinguishing between these variables is essential, as individuals aware of voice inputs but not using them are likely to dislike such features, while those unaware might be positive about using them once they are informed about these possibilities. However, the lack of awareness regarding voice inputs suggests a potential lack of technological knowledge,

[5] The four dummies for lack of awareness and use in daily life are created using FREQDICTATION and FREQVOICE.

which, in turn, may result in increased difficulties in utilizing the voice tools and subsequently lower voice participation. Overall, we expect that both individuals unaware of voice inputs and those aware but never using them are less likely to participate through voice.

- Device type (1 = smartphones/tablets, 0 = PCs): Since PCs are not always equipped with microphones, PC respondents might participate less using voice inputs.
- Place of completion (1 = home, 0 = other): Responding from home is expected to be associated with higher voice participation (e.g., lower privacy concerns at home).
- Presence of third parties (1 = people around, 0 = alone): The presence of third parties is expected to decrease voice participation, due to privacy concerns.

We report the odds ratios (OR) and 95% confidence intervals (CI) of these two logistic regressions (dummies based on USEDDICTATION and USEDRECORDING).

To answer ***RQ3***, we first report the proportion of respondents (within those who stated having used the voice input methods, see Appendix A) who reported having faced the following problems: “Technical problems (e.g., microphone not working)”, “Problems understanding the function”, “I could not speak freely because of the context (e.g., I was around other people)”, “I found it difficult to express my answers orally”, or “None of these”. The proportions of panelists selecting each option are presented for both dictation and voice recording, separating the push from the choice groups. Tests of significance are implemented, in a similar way as for ***RQ1***.

Finally, to answer ***RQ4***, we report the proportions of respondents who found it easy/difficult and who dis/liked using the voice input methods and answering by text. While these questions[6] were all asked using a five-point bipolar scale, for the analyses we combined the two positive (e.g., extremely and quite easy) and the two negative (e.g., extremely and quite difficult) answer categories, thus presenting three categories (positive, neutral, negative). Again, tests of significance are implemented as in previous analyses, although this time we additionally test for significance of the differences between text and the four other groups.

All analyses were performed using R version 4.3.1 (R Core Team, 2023).

[6] See questions EASYDICTATION, EASYVOICE, EASYTEXT, LIKEDICTATION, LIKEVOICE and LIKETEXT in the questionnaire (SOM1).

Results

Stated Reasons for not Using Voice Inputs (RQ1)

Table 2 presents the proportions of respondents who selected each of the reasons for not using voice inputs when offered, distinguishing dictation and voice recording, and push and choice groups.

Table 2 Reasons for not using dictation or voice recording for those who stated not having used them (% of those answering the question)

Reasons for not using voice input	Dictation		Recording	
Group	Push (n=130)	Choice (n=186)	Push (n=107)	Choice (n=169)
Prefer another alternative	NA	57.5	NA	51.5
Concerns about context	24.6	16.7	**30.8**	**17.2**
Hard to express orally	21.5	13.4	22.4	17.2
No microphone	**20.0**	**7.0**	**17.8**	**6.5**
Technical problems	**18.5***	**1.6**	7.5*	3.5
Problems understanding the function	**6.9**	**1.1**	4.7	1.2
Other reason	**17.7**	**9.7**	**22.4**	**9.5**

Note. The sum is not 100 because respondents could select several reasons. Bold numbers indicate significant differences between push and choice groups (5% level) within methods. Stars (*) indicate significant differences between dictation and recording (*PushDictation* vs *PushRecording* or *ChoiceDictation* vs *ChoiceRecording*). P-values of all tests are provided in SOM3.

First, focusing on the reasons offered to all groups and excluding the "other" option, the ranking is similar for all four groups: concerns about the context is the main reason for not using voice inputs, followed by the difficulty of expressing one's ideas orally. Technical and understanding problems, in contrast, are reported less often.

However, important differences exist across groups. In particular, technical problems are reported as a reason for not using voice input by a much larger proportion of respondents in the *PushDictation* group, compared to the others.

Furthermore, in the choice groups, more than half of the respondents mentioned that they "preferred another alternative". Since this option was not offered for the push groups, this creates important differences in the reported levels of other reasons between push and choice groups.

Variables Associated with the Use of Voice Inputs (RQ2)

Moving to ***RQ2***, Table 3 presents the OR and 95% CI of the two logistic regressions, with reported use of dictation or voice recording to answer at least one experimental question as dependent variables.

Table 3 OR and 95% CI of the logistic regressions

DV: reported using…	Dictation			Recording		
	OR	2.5%	97.5%	OR	2.5%	97.5%
Choice group	**0.31**	0.20	0.47	**0.22**	0.14	0.33
Female	1.44	0.95	2.21	1.14	0.76	1.73
Age (Chi^2=5.7 for Dictation and 0.4 for Recording, d.f.=2, p>.05 in both cases)						
30 or less	0.54	0.27	1.02	0.84	0.46	1.50
31 to 59	-	-	-	-	-	-
60 or more	1.39	0.80	2.38	0.92	0.54	1.54
Education (Chi^2=3.7 for Dictation and 2.0 for Recording, d.f.=2, p>.05 in both cases)						
Low	0.64	0.38	1.07	1.36	0.80	2.31
Middle	-	-	-	-	-	-
High	0.63	0.36	1.08	1.00	0.58	1.74
Spanish native language	1.14	0.52	2.64	1.44	0.71	2.96
Social trust	1.08	0.90	1.30	0.98	0.81	1.18
Complete trust in confidentiality	**1.83**	1.13	2.98	**2.13**	1.32	3.47
Comfortable with technology	1.36	0.74	2.56	1.19	0.70	2.03
Not aware dictation/recording	1.14	0.57	2.24	0.98	0.24	3.35
Use dictation/recording in daily life	**4.21**	2.68	6.71	**2.57**	1.57	4.27
Answer from mobile	**2.01**	1.20	3.42	1.29	0.76	2.21
Answer from home	1.11	0.65	1.89	0.75	0.46	1.25
People around	0.68	0.41	1.10	0.86	0.52	1.40
Intercept	**0.18**	0.05	0.69	0.35	0.10	1.18
AIC		*577.34*			*585.93*	
N		*490*			*473*	

Note: Bold numbers indicate statistically significant odds ratios.

The use of both voice inputs is influenced by several factors. Firstly, the method employed to offer the voice inputs plays an important role. As expected, individuals provided with a choice are less inclined to use voice inputs compared to those in the push groups. Secondly, individuals who completely trust that their answers are treated confidentially are more likely to use voice inputs. Moreover, respondents who already use voice inputs in their daily lives are also more likely to employ them within the survey context. Additionally, answering from a mobile device also increases the likelihood of using dictation.

Notably, only a few variables exhibit significant effects. In particular, despite the survey context being the most frequently cited reason for not using voice inputs (excluding the "prefer another alternative," which was exclusively proposed in the *Choice* group; see Table 2), factors such as being at home and the presence of third parties do not yield significant effects. Similarly, variables that one might expect to be correlated with difficulties in articulating oral responses (e.g., non-native Spanish speakers or lower education levels) do not demonstrate significant effects. Lastly, the comfort level in using new technologies, which could be associated with understanding problems, also does not show any significant effects.

Stated Problems (RQ3)

Panelists who stated they used dictation and/or voice recording to answer to at least one of the experimental questions were asked whether they faced various problems when using these tools. Table 4 reports the proportion of respondents reporting having encountered each issue, distinguishing dictation/voice recording and push/choice groups.

Table 4 Reported problems for those who stated having used dictation or voice recording (in % of those answering the question)

Reported problems	Dictation		Recording	
Group	Push (n=120)	Choice (n=60)	Push (n=145)	Choice (n=56)
None	45.8*	50.0	64.1*	58.9
Technical problems	21.7*	16.7	6.9*	8.9
Hard to express orally	20.8*	11.7	11.0*	7.1
Could not speak freely given context	10.0	20.0	15.9	19.6
Problems understanding the function	6.7	8.3	2.8	5.4

Note. The sum is not 100 because respondents could select several reasons (except if they selected "none"). There are no significant differences between push and choice groups (5% level). Stars (*) indicate significant differences between dictation and recording (*PushDictation* vs *PushRecording* or *ChoiceDictation* vs *ChoiceRecording*). P-values of all tests are provided in SOM3.

First, a majority of respondents in the voice recording groups did not report experiencing any of the difficulties we asked about. In the dictation groups, slightly fewer than half (46% and 50%) reported encountering no issues.

In particular, the *PushDictation* group exhibited a significantly higher incidence of technical problems and greater difficulty in articulating responses orally compared to the *PushRecording* group. Furthermore, 10% to 20% of respondents (contingent on the group) reported constraints in expressing themselves freely due to contextual factors. Conversely, challenges pertaining to comprehension of tool functionality were the least frequently reported.

Evaluations (RQ4)

Finally, to answer ***RQ4***, Table 5 presents the evaluations of respondents of three ways of answering: by text (used by all respondents to answer at least one open-ended narrative question), dictation and voice recording (for those reporting using them to answer at least one question).

Table 5 Evaluation of the way of answering questions

			Dictation		Recording	
Var.	Answer categories	Text (n=1,001)	Push (n=120)	Choice (n=60)	Push (n=145)	Choice (n=56)
EASY	Easy	72.9	_63.3_*	_51.7_	__77.2__*	___51.8___
	Neither easy nor difficult	20.8	__15.0__	___41.7___	__13.8__	___44.6___
	Difficult	6.3	___21.7___*	__6.7__	9.0*	3.6
LIKE	Liked	48.3	40.0	46.7	_38.6_	_33.9_
	Neither liked nor disliked	46.6	44.2	46.7	48.3	_62.5_
	Disliked	5.2	_15.8_	6.7	_13.1_	3.6

Note. Bold numbers indicate significant differences between push and choice groups (5% level). Stars (*) indicate significant differences between dictation and recording (*PushDictation* vs *PushRecording* or *ChoiceDictation* vs *ChoiceRecording*). Numbers in italics indicate significant differences compared to Text. P-values of all tests are provided in SOM3.

Overall, most respondents found it easy to answer (51.7% to 77.2%), in the case of text as well as in the case of voice inputs. In contrast, a minority of respondents reported liking answering in each of the ways (33.9% to 48.3%).

However, while there are no significant differences between experimental groups in how much respondents dis/liked answering in different ways, in the case of easiness, differences are observed. In particular, significantly more respondents found it difficult to answer through dictation than through voice recording. Also, respondents given a choice reported significantly more that it was “neither easy nor difficult” to use the voice tools, compared to those in the push groups.

Conclusions

Voice input surveys offer exciting opportunities, but several challenges persist. This study provides new empirical evidence, comparing two voice input methods (dictation and voice recording) and two ways of proposing them to participants (push and choice).

Summary of Results

The results show, first, that in the *Choice* groups, the primary reason stated for not using voice input (***RQ1***) is that respondents prefer text input. Then, in all groups, follow concerns related to the context (e.g., the presence of others) and the difficulty of orally expressing one’s ideas. Although technical and understanding problems are still present, especially in the *PushDictation* group, they are reported by smaller proportions of respondents compared to other issues. Notably, the prevalence of technical and understanding issues is lower than in the study by Revilla and Couper (2021), where technical problems were reported by 12% to 25% of the respondents and understanding problems by 14% to 17% (depending on the groups; all groups used voice recording). This reduced reporting of technical and understanding issues relative to previous studies may be attributed to the use of a new tool, *WebdataVoice*, and/or to the increasing proficiency of panelists in using their devices.

Moving on to ***RQ2***, employing logistic regression analyses, we found that only a few of the tested variables exhibit significant associations with the reported use of voice inputs to answer the experimental questions: providing a choice (as opposed to pushing to voice), having complete trust in the confidentiality of the answers, already using voice inputs in daily life, and, in the case of dictation, answering through a mobile device. In contrast, other variables, such as being at home or having people around, do not show significant effects, despite the context being cited as a key reason for not using voice inputs (see ***RQ1***).

As for the challenges posed by the use of voice input tools (***RQ3***), a majority of respondents who reported using these tools did not report experiencing any of the challenges we asked about. However, in the choice groups, around 20% of respondents reported constraints associated with the context, while in the *PushDictation* group, similar proportions reported both technical problems and difficulty of expressing answers orally.

Turning to the evaluation of different answering methods (***RQ4***), namely text, dictation and voice recording, the majority of respondents found it easy to answer in all three methods, although the specific levels varied across groups. Nevertheless, fewer participants reported liking the voice input methods. However, compared to the study by Revilla and Couper (2021), we found higher levels of liking of the tools (33.9% to 46.7% versus 22.6% to 30.8%).

Limitations and Practical Implications

This study is subject to certain limitations. Firstly, the sample size disparity among groups, particularly notable in the choice group where a small proportion opted for voice tools, might account for the limited statistical significance observed in some instances. Secondly, reliance on self-reports introduces the possibility of errors. Thirdly, we do not have detailed information on the specific nature of problems encountered, such as the specifics of “technical problems”. Finally, factors such as the topic (opinions about nursing homes), question type (probes), country (Spain), and sample source (opt-in panel) could influence the results. Therefore, further research is required to evaluate their robustness in different contexts.

Despite these limitations, this research contributes to the existing literature by shedding light on the differences between dictation and voice recording, as well as between push and choice designs. Importantly, it underscores that many obstacles to the adoption of voice input extend beyond the researcher’s control. The primary impediments, contextual constraints and difficulty in oral expression, are inherently beyond the purview of researchers conducting web surveys.

Nevertheless, some of the results can help enhance the effectiveness of voice input surveys. For instance, our analyses suggest that trust in the confidentiality of the answers is one of the few variables which significantly affects the use of voice inputs, in line with Revilla and Couper’s (2021) results. These levels of trust could be improved by joint efforts between researchers and fieldwork companies to guarantee data protection, for example by improving transparency and security measures. Also, we found that answering through mobile devices increases the likelihood of using dictation. Thus, researchers planning to propose dictation could encourage participants to answer through mobile devices. Finally, technical and understanding problems remain, even if these do not seem to be the main obstacles to the use of voice inputs to answer survey questions. Strategies to reduce them are therefore still needed. More generally, further research in this area is warranted to uncover additional insights and refine best practices for voice-based surveys.

References

Deloitte (2018). *2018 Global mobile consumer survey: US edition. A new era in mobile continues.* Retrieved from https://www2.deloitte.com/tr/en/pages/technology-media-and-telecommunications/articles/global-mobile-consumer-survey-us-edition.html

Gavras, K. (2019, March 6–8). *Voice Recording in Mobile Web Surveys – Evidence from an Experiment on Open-Ended Responses to the 'Final Comment'.* [Paper presentation] General Online Research conference, Cologne, Germany.

Gavras, K., & Höhne, J. K. (2022). Evaluating Political Parties: Criterion Validity of Open Questions with Requests for Text and Voice Answers. *International Journal of Social Research Methodology, 25*(1), 135–141. https://doi.org/10.1080/13645579.2020.1860279

Gavras, K., Höhne, J. K., Blom, A. G., & Schoen, H. (2022). Innovating the collection of open-ended answers: The linguistic and content characteristics of written and oral answers to political attitude questions. *Journal of the Royal Statistical Society Series A: Statistics in Society, 185*(3), 872–890. https://doi.org/10.1111/rssa.12807

Höhne, J. K. (2021). Are respondents ready for audio and voice communication channels in online surveys? *International Journal of Social Research Methodology, 26*(3), 335–342. https://doi.org/10.1080/13645579.2021.1987121

Höhne, J. K., & Gavras, K. (2022). Typing or speaking? comparing text and voice answers to open questions on sensitive topics in smartphone surveys. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.4239015

Höhne, J. K., Kern, C., Gavras, K., & Schlosser, S. (2023). The sound of respondents: Predicting respondents' level of interest in questions with voice data in smartphone surveys. *Quality & Quantity.* https://doi.org/10.1007/s11135-023-01776-8

Lenzner, T., & Höhne, J. K. (2022). Who Is Willing to Use Audio and Voice Inputs in Smartphone Surveys, and Why? *International Journal of Market Research, 64*(5): 594-610. https://doi.org/10.1177/14707853221084213

Lütters, H., Friedrich-Freksa, M., & Egger, M. (2018, February 28–March 2). *Effects of Speech Assistance in Online Questionnaires.* [Paper presentation] General Online Research conference, Cologne, Germany.

Meitinger, K., van der Sluis, S., & Schonlau, M. (2022, March 3–4), *Implementing Voice-Recordings in a Probability-based Panel: What We Learnt So Far.* [Paper presentation] CIPHER virtual conference https://cesr.usc.edu/cipher_2022

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Revilla, M. (2022). How to Enhance Web Survey Data Using Metered, Geolocation, Visual and Voice Data? *Survey Research Methods, 16*(1), 1-12. https://doi.org/10.18148/srm/2022.v16i1.8013

Revilla, M., & Couper, M. P. (2021), Improving the Use of Voice Recording in a Smartphone Survey. *Social Science Computer Review, 39*(6), 1159-1178. https://doi.org/10.1177/0894439319888708.

Revilla, M., Couper, M. P., Bosch, O. J., & Asensio, M. (2020). Testing the Use of Voice Input in a Smartphone Web Survey. *Social Science Computer Review, 38*(2), 207-224. https://doi.org/10.1177/0894439318810715

Revilla, M., Couper, M. P., & Ochoa, C. (2018). Giving Respondents Voice? The Feasibility of Voice Input for Mobile Web Surveys. *Survey Practice, 11.* https://doi.org/10.29115/SP-2018-0007

Revilla, M., Iglesias, P., Ochoa, C., & Antón, D. (2022). *WebdataVoice: a tool for dictation or recording of voice answers in the frame of web surveys*. [Computer software]. OSF. https://doi.org/10.17605/OSF.IO/B2WYZ

Singer, E. & Couper, M.P. (2017). Some Methodological Uses of Responses to Open Questions and Other Verbatim Comments in Quantitative Surveys. methods, data, analyses 11(2), 115-134. https://doi.org/10.12758/mda.2017.01

Appendix A

Reported Use of Dictation and Voice Recording

Table A1 presents the answers to the questions USEDDICTATION and USED-VOICE, asking respondents to report whether they used dictation (respectively, voice recording) to answer at least one of the experimental questions. Three response options were proposed: “Yes, I used only the dictation tool whenever I had this option”, “Yes, I used the dictation tool, but also other options (e.g., the keyboard)”, and “No, I never used the dictation tool” (same with voice recording).

Table A1 Reported use of voice inputs per group (in %)

Reported use of…	Dictation		Voice recording	
Group	Push (n=250)	Choice (n=246)	Push (n=252)	Choice (n=225)
Yes, only this	20.8	8.5	36.9	8.4
Yes, but not only	27.2	15.9	20.6	16.4
No	52.0	75.6	42.5	75.1

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should...
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - be submitted as MS Word documents.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
  - should be anonymized ("blinded") for review.
- Please also send us an abstract of your paper (approx. 300 words), a front page with a brief biographical note (no longer than 250 words as supplementary file), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - pdf
  - jpg (uncompressed, high quality)
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi.

- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- Please number the headings of your article. Doing so will make the work of the layout editor easier.
- If your text includes formulas we would like to ask you to upload your text also as a PDF, additional to the Word document.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.
- Response rates in research papers or research notes, where population surveys are analyzed, should be calculated according to AAPOR standard definitions.
- Please follow the APA guideline when structuring and formating your work.

When preparing in-text references and the list of references please also follow the APA guidelines:

Entire Book

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI)

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. https://doi.org/10.1037/1061-4087.45.2.10

Journal Article (without DOI)

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI)

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).