methods, data, analyses | 2026, pp. 1-28 DOI: 10.25521/mda.662

Measurement Invariance of Survey Data across Face-to-face and Video Interview Modes

Diana Zavala-Rojas[1,2], Oguz Korkut Keles[2] & Hannah Schwarz[2]

[1] *European Social Survey ERIC*

[2] *Universitat Pompeu Fabra*

Abstract

The COVID-19 pandemic posed challenges to the traditional mode of administration of surveys. In most European countries it was not feasible to conduct standardized in-person interviews due to restrictions. These events accelerated the trend towards alternative modes of survey data collection, such as live video interviewing. In this paper, we investigate whether survey data collected with two different interviewing modes, face-to-face and video interview, are comparable. Measurement invariance is a prerequisite to make meaningful comparisons across groups. Using data from the European Social Survey Round 10, we assess if the measurement invariance condition is met for two concepts, social trust and attitudes towards immigration, in six European countries: Estonia, Finland, Iceland, Italy, the Netherlands and Norway. The results are encouraging: Full metric invariance is established for all countries for social trust and for three out of six countries (Iceland, Italy and Norway) for attitudes towards immigration. For those cases, analysts can aggregate data across modes to investigate relationships among variables, e.g., using regression analysis. Scalar invariance holds in fewer cases. For the concept social trust, full scalar invariance is established for Finland, Italy, the Netherlands and Norway while for the concept attitudes towards immigration, it is established for Iceland and Italy. For those cases, data can be aggregated across modes to compare observed means at the country-language level. Generally, where measurement equivalence is not reached, we recommend excluding observations from the video interview mode before analysis.

Keywords: Measurement invariance, video mode, face-to-face mode, European Social Survey

Conducting video interviews for various purposes is becoming increasingly popular as the required technology becomes ever more accessible (Conrad et al., 2023). Video interviewing offers potential advantages compared to in-person interviewing, such as cost reduction and time efficiency. A trend towards video interviewing was already discernible before the COVID-19 pandemic (Joshi et al., 2020). With the COVID-19 pandemic posing numerous challenges to survey research relying on in-person (i.e., face-to-face) interviews, its relevance peaked as conducting in-person interviews became infeasible for entities conducting survey interviews in many countries, especially for surveys also covering respondents in health risk groups such as elderly persons. Given the apparent success of video interviews, survey researchers have continued studying this method beyond the pandemic period. The mode of interview administration typically has effects on the responses that people provide to survey questions. Recent literature investigates the possible implications the video interview mode might have on different aspects of surveys, for example in terms of interviewer effects (West et al., 2022), implementation challenges (Durrant et al., 2024), participation (Schober et al., 2023) and data quality (Conrad et al., 2023). Furthermore, survey mode effects across in-person and video interview mode have been studied (Endres et al., 2023).

These studies suggest that differences between modes may influence how respondents process questions and provide answers—through mechanisms such as interviewer behavior or perceived confidentiality. Our work contributes to this body of research by investigating the comparability of survey data across in-person and video interview mode by formally testing for measurement equivalence (Meredith, 1993). We assess measurement invariance to examine whether constructs are measured equivalently across modes, that is, whether differences in responses reflect true differences in the underlying concepts. We use measurement equivalence testing because respondents have not been randomly assigned to interview modes in the survey data we use. In this context, measurement invariance testing[1] provides a statistical means to evaluate comparability across groups while acknowledging that some residual bias from sampling differences may remain. It enables researchers to assess if the compara-

[1] *Measurement equivalence or measurement invariance* is assessed by testing whether certain parameters of a measurement model are identical across groups (Davidov et al., 2015).

Acknowledgement
We thank Carlos Poses for his comments on an early version of this study and Laura Font Rigueiro who collaborated as a research assistant, as well as the anonymous reviewers for their helpful comments.

Direct correspondence to
Hannah Schwarz, Edificio Mercè Rodoreda, Despatx 24.407, Ramón Trías Fargas 25-27, 08005 Barcelona, Spain
E-mail: hannah.schwarz@upf.edu

bility of data is not hindered by group membership (here: mode groups). Only if measurement invariance is established between in-person and video interview data, can measurements across groups be analyzed jointly in conventional statistical analysis.

There are different levels of tests for measurement invariance with increasing restrictions in the parameters of the models (Meredith, 1993; Vandenberg & Lance, 2000). Configural invariance means that the same survey items (i.e., indicators) load on the same latent construct (i.e., factor) across groups. Sometimes, this is also referred to as the measurement model. It is interpreted as a construct having similar theoretical content across groups. Because the factor model relies on only three observed variables (statistically this is known as a just identified model), we do not test for configural invariance. This situation is common with very short scales, particularly in survey research where three-indicator single-factor models are not uncommon and where the identification constraints needed to set the metric of the latent variable exhaust all available degrees of freedom. Prior methodological work (e.g., Cheung & Rensvold, 2002; Wu & Estabrook, 2016) therefore treats configural invariance in such cases as satisfied, noting that no empirical misfit can be detected at this step. In other words, the theoretical models are not tested but assumed[2].

Loading invariance, also known as metric invariance, means that the items' factor loadings are equal across groups. Where this is the case, statistical relationships of each item of a construct can be meaningfully compared across groups. In other words, the relative weight of each item is the same across groups.

Intercept invariance, also known as scalar invariance, means that not only the items' factor loadings but also the intercepts are equal across groups. When metric and scalar invariance are established across groups, observed means can be compared. In the remainder of this paper, we use the terms metric and scalar invariance.

The above also represents an ascending rank ordering of desirability of types of invariance for analysts. In order to conduct meaningful analyses across groups, here mode groups, establishing scalar invariance results in most ease for analysts as it allows them to meaningfully compare observed means across the mode groups. Where only metric but not scalar invariance is established, only statistical relationships can be meaningfully compared across mode groups. If analysts still wish to compare means having only established metric invariance, they have to resort to the strategy of comparing latent means to do so.

---

[2] Our approach in this article of assuming the theoretical model in a case of only three items is in line with other published work such as Nickel and Weber (2024); Pirralha and Weber (2020); Van de Vijver (2019); Nießen et al. (2020). We include in the Appendix the commonly reported global fit statistics for our scalar models purely for the sake of transparency, not because we actually rely on them for model evaluation.

One, furthermore, distinguishes between partial and full invariance where partial invariance means that a given type of invariance can be established for only some items constituting a construct, not all (Byrne et al., 1989). For the case of finding partial invariance, a strategy analysts can resort to is, once again, comparing latent means instead of observed means (Pokropek et al., 2019).

If the interactions between respondents and interviewers were the same in both modes (West et al., 2022), we would expect scalar invariance. However, we will argue that the social dynamics in the two modes differ due to factors such as perceptions of confidentiality and personal proximity. Furthermore, interviewers have less control over the context of the interview in video mode such that distractions through multitasking on the side of the respondent are more likely than in in-person mode (Deakin & Wakefield, 2014). Additionally, other people might be present during the interview unbeknownst to the interviewer who only sees the part of the room captured by the camera. More generally, a relevant argument for this study concerns potential differences in social desirability across modes and the concepts selected for the measurement invariance tests. Social desirability bias tends to be less pronounced in modes where interpersonal proximity between respondent and interviewer is lower (Heerwegh, 2009). This would be the case in video interviews as both of them are not physically present in the same place, but the interaction happens by the means of a screen.

For all these reasons, we expect that data collected across the two modes may not be fully comparable, that is, we do not expect full scalar invariance.

Using data from the European Social Survey (ESS) Round 10, we test if the measurement invariance condition is met for two concepts, social trust and attitudes towards immigration, in six different European countries: Estonia, Finland, Iceland, Italy, the Netherlands and Norway.

Despite having found differences across countries and concepts in the levels of invariance, if in-person interviews are the benchmark, our results are encouraging for video interviews: Full metric invariance is established for all countries for social trust and for three out of six countries (Iceland, Italy and Norway) for attitudes towards immigration. For those cases, analysts can aggregate data across modes to investigate relationships among variables, for example, using regression analysis. Full scalar invariance is established for fewer cases but still for four countries for social trust (Finland, Italy, the Netherlands and Norway) and for two countries for attitudes towards immigration (Iceland and Italy). For those cases, data can be aggregated across modes to compare observed means at the country level. Generally, where measurement equivalence is not reached, we recommend excluding observations from the video interview mode before analysis or using latent variable approaches, to account for measurement error (Saris & Gallhofer, 2014). In this article, we argue that differences in invariance levels across countries and concepts may be attributable to differences in social desirability across modes, concepts and cultures. However, future research will be necessary to elucidate the exact causes of non-invariance.

The remainder of the paper is structured as follows: First, we introduce the relevant literature and position our research within it. Then, we define the concepts that are examined across different modes, namely social trust and attitudes towards immigration. Next, we present the measurement instruments (i.e., survey questions) used to measure them. Then, in the methods section, we elaborate on the data we use, on measurement invariance testing more generally and on our analytical approach. Finally, we present our results followed by a discussion and conclusion.

Survey Modes and Potential Implications for Equivalence

Our research contributes to the growing literature that tests for measurement invariance to elucidate potential issues with data equivalence across groups. The necessity of testing for measurement invariance has been emphasized in several studies as it ensures meaningful, valid, and unbiased comparisons across data gathered in different groups (e.g., modes, countries, languages; Meuleman et al., 2023). This is particularly critical in cross-cultural studies, where invariance testing is essential to validate comparisons (Davidov et al., 2014; Steinmetz et al., 2009).

More specifically, our research contributes to the literature on measurement invariance across modes of data collection. Previous research on this includes, for instance, Boal et al. (2017) who assess measurement invariance across two different data collection modes, telephone and web self-administration. Their hypothesis is that there may be differences because the survey is presented primarily visually in web self-completion mode and primarily aurally in telephone mode, and additionally, because phone administration involves an interviewer, whilst web administration does not. Their results show configural, metric and partial scalar invariance for the Posttraumatic Stress Disorder (PTSD) Checklist across the two mode groups. The authors conclude that, in practical terms (i.e., for PTSD diagnosis) the use of both modes is unproblematic.

There is, furthermore, literature on mode effects, some of which we will review here because it helps us develop our hypotheses. For example, Buerger et al. (2016) assess mode effects between different self-completion modes hypothesizing that they could exist due to factors such as item layout, input devices, scrolling, and the likelihood of revisiting of past responses. Some of these factors might also differ between video and in-person interviews. For example, input devices and/or item layout (depending on the screen size) might differ between the two modes. Furthermore, the physical presence of an interviewer in in-person interviews may induce pressure on respondents to continue with the interviewing process swiftly which could lead to respondents in in-person

interviews being less inclined to revisit previous questions than those in video interviews.

More generally, a relevant argument for this study concerns potential differences in social desirability across modes. Social desirability bias tends to be less pronounced in modes where the interpersonal proximity between respondent and interviewer is lower (Heerwegh, 2009). While this argument is often made concerning the distinction between interviewer-administered and self-completion modes (Heerwegh, 2009; Schnell & Kreuter, 2005), it similarly applies to the comparison of a context where the interviewer is physically present in a room with respondents as compared to a context where respondent and interviewer communicate via videochat (Moallem, 2015). If this is the case, we could expect differences across the concepts. As attitudes towards immigration shows high social desirability bias (Creighton et al., 2019), it could be the case that in video interview respondents would not feel as pressured to give a socially expected answer as they would when the interviewer is present in the same place as them.

Closely linked to this is the literature exploring the varying impact of interpersonal interactions across interview modes. We argue that the type of relationship established between the respondent and the interviewer during the interview process may differ due to factors such as their perceptions of confidentiality, willingness to share information, social presence, personal proximity, and/or the frequency of interactions or dialogues (e.g., there might be less probing in one mode; Hoehne et al., 2024; de Villiers et al., 2022). The quality of the interaction between respondent and interviewer is an important aspect of the mode effect (Conrad et al., 2023) and may also be important for the comparability of the data across cultures (Van de Vijver & Tanzer, 2004). Even if a live interviewer is present in both assessed modes, there is evidence that a successful interaction through videochat requires experience (de Villiers et al., 2022). Further research demonstrates that rapport is stronger in in-person interviews or, more generally, in-person communication than in video interviews (Faucett et al., 2017; Fernandez et al., 2021; Gordon et al., 2020; Joseph et al., 2017; Sherman et al., 2013). Yet, a weaker social presence of the interviewer in video mode may also be beneficial for topics in which the respondent perceives a high expectation of a socially desirable answer.

Another aspect to consider is that video interviews can take place in a larger variety of environments than in-person interviews which are usually conducted at a respondent's home. An important consequence is the higher potential for disruptive background factors in video interviews which might lead to distraction or multitasking behavior on the side of the respondent (Deakin & Wakefield, 2014).

Furthermore, in the literature a distinction is made between two types of video interviews: pre-recorded and live interviews. Schober et al. (2020) argue

that differences between live video interviews and pre-recorded video interviews that may induce different answering behavior across the two modes due to the fact live video interviews involve interaction and pre-recorded video interviews do not.

Concepts

We test for measurement invariance of the concepts of social trust and attitudes towards immigration using the country cases of the ESS Round 10 which have data available for video as well as in-person interview mode. We specifically chose these concepts because they differ in how prone they are to social desirability bias, that is, the tendency to provide an answer that the respondent thinks aligns with a socially expected and accepted idea. The measurement of attitudes towards immigration is argued to suffer from high social desirability bias (Rinken et al., 2021; Piekut, 2021) while this is not the case for social trust. Furthermore, video and in-person interviewing should induce differing levels of social presence of the interviewer (Moallem, 2015) which, in turn, should affect how strongly answers are impacted by social desirability bias. Hence, we expect variation in patterns of answering behavior between the two modes across the two concepts.

Furthermore, we consider the selected concepts especially relevant as both are widely studied in the social sciences. There is a broad consensus among social scientists about the importance of the concept of social trust, from a political and societal perspective. For example, it has been argued that social trust is central for the functioning of democracies (Inglehart, 1999). However, even if social trust has been routinely used in cross-country research (Inglehart, 1999; Delhey & Newton, 2005; Letki & Evans, 2005; Rothstein & Uslaner, 2005; Adam, 2008; Herreros & Criado, 2008; Kaasa & Parts, 2008; Zmerli & Newton, 2008), the measurement of social trust is not necessarily comparable across different groups (Van der Veld & Saris, 2011; Reeskens & Hooghe, 2008). Studying the concept of attitudes towards immigration has also been a relevant matter of research in the social sciences for some decades (see, e.g., Ceobanu & Escandell, 2010). The determinants of these attitudes are also a matter of study, both at the individual and the country level. For a review see for instance Davidov and Semyonov (2017). Furthermore, this concept is frequently studied using cross-nationally comparative datasets (Ceobanu & Escandell, 2010), where, among many other factors, the modes of data collection vary (Smith et al., 2011).

For both concepts, cross-cultural measurement equivalence has been studied with ESS data (see, e.g., Davidov et al., 2015; Nickel & Weber, 2024; Reeskens & Hooghe, 2008). Furthermore, Revilla (2013), assesses measurement equivalence of the two concepts across in-person and web self-completion mode in sur-

vey data from the Netherlands and finds scalar invariance for both concepts. Moreover, punctual experiments have been conducted in different ESS Rounds which assessed mode effects across a variety of modes. Results from these show, for example, mode differences in attitudinal questions in general, and more specifically for the concept of attitudes towards immigration in in-person interviews versus various alternative modes (Villar & Fitzgerald, 2017).

In this study, each of the concepts is measured with three indicators. The wording of the three corresponding ESS items is shown in Table 1 and Table 2, respectively. All six survey questions are presented with 11-point item-specific scales (0-10). Figure 1 illustrates the configural model, that is, the relationship between the latent variables and the indicators for the concept of social trust, as an example. As shown by the three unidirectional arrows, the responses to the observed variables (y_{ij}) are determined by the latent construct (ξ_j). The corresponding equations to this graph are shown in Section 4.2.

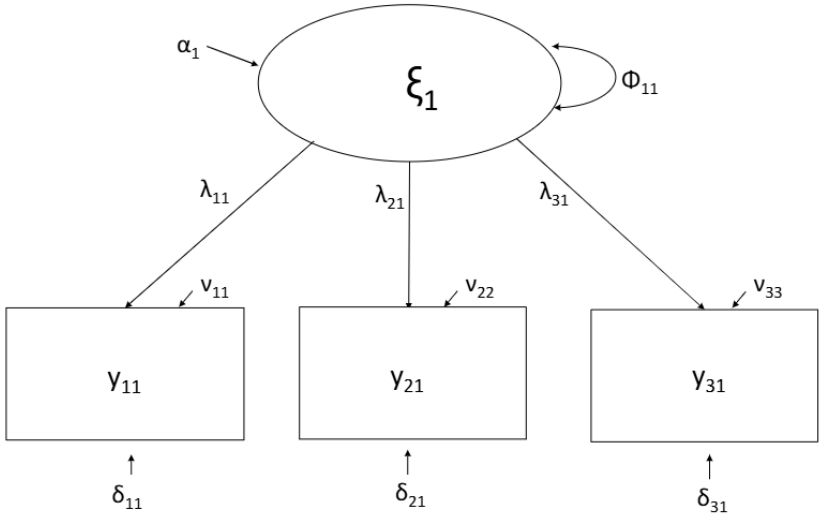

Figure 1 Path diagram of the configural model

Table 1 ESS survey instrument for measuring concept social trust

Question wording	Variable name	Response scale
Using this card, generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people? Please tell me on a score of 0 to 10, where 0 means you can't be too careful and 10 means that most people can be trusted.	ppltrst	0 (You can't be too careful) to 10 (Most people can be trusted)
Using this card, do you think that most people would try to take advantage of you if they got the chance, or would they try to be fair?	pplfair	0 (Most people try to take advantage of me) to 10 (Most people try to be fair)
Would you say that most of the time people try to be helpful or that they are mostly looking out for themselves?	pplhlp	0 (People mostly look out for themselves) to 10 (People mostly try to be helpful)

Table 2 ESS survey instrument for measuring concept attitudes towards immigration

Question wording	Variable name	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	0 (Bad for the economy) to 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	0 (Cultural life undermined) to 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	0 (Worse place to live) to 10 (Better place to live)

Methods

Data

We use data from the ESS Round 10 (ESS ERIC, 2023). Due to locally varying contact restrictions in the context of the COVID-19 pandemic, some ESS countries opted for adding a video-interviewing option to their Round 10 fieldwork protocol while others did not (Ghirelli et al., 2024). In this paper, we include all six country cases for which a video interviewing option was offered to respondents as an alternative to in-person interviewing, namely Estonia, Finland, Iceland, Italy, the Netherlands and Norway. Fieldwork in these countries was conducted

between June 2021 and May 2022. The ESS reports the following response rates: 47.2% (Estonia), 41.1% (Finland), 33.6% (Iceland), 49.8% (Italy), 35.7% (the Netherlands) and 37.9% (Norway). These response rates are calculated as the number of complete and valid interviews over the number of issued eligible sample units (for more detailed information, see Ghirelli et al., 2024). Distributions of the variables of main interest are presented in Tables 3 and 4. These Tables also show that the proportions of cases missing due to item nonresponse are low with a mean of 0.7% for the concept of social trust and somewhat higher, with a mean of 3.2%, for the concept of attitudes towards immigration. This is unsurprising given that the latter concept can be regarded as more sensitive, frequently being at the center of political debate and given that previous studies have found higher item nonresponse for this concept (for a more extensive discussion, see, e.g., Piekut, 2021).

Table 3 Descriptive statistics of variables measuring social trust (ppltrst, pplfair, pplhlp) by country and mode

Country	Mode	Valid *N*	*N*	ppltrst *M (SD)*	pplfair *M (SD)*	pplhlp *M (SD)*
Estonia	in-person	1296	1302	5.70 (2.17)	5.88 (2.13)	5.16 (2.15)
	video	239	240	6.10 (2.01)	6.44 (1.84)	5.76 (2.02)
Finland	in-person	1333	1337	6.93 (1.87)	7.10 (1.84)	6.37 (1.92)
	video	239	240	7.00 (1.62)	7.18 (1.58)	6.37 (1.64)
Iceland	in-person	560	567	6.53 (2.11)	7.14 (1.69)	6.73 (1.82)
	video	331	333	6.71 (1.99)	7.20 (1.59)	6.55 (1.71)
Italy	in-person	2166	2183	4.96 (2.24)	4.97 (2.04)	4.56 (2.04)
	video	447	457	5.21 (2.12)	5.22 (1.95)	4.69 (2.02)
Netherlands	in-person	1216	1221	6.50 (1.81)	6.76 (1.57)	5.89 (1.75)
	video	245	248	6.76 (1.71)	6.93 (1.34)	6.02 (1.51)
Norway	in-person	917	920	6.90 (1.87)	7.09 (1.74)	6.37 (1.87)
	video	488	491	6.69 (1.81)	6.95 (1.67)	6.22 (1.79)

As respondents self-selected into either the in-person or the video interviewing mode, it can be expected that the sociodemographic composition of these two groups differs. We provide information on the distributions of variables age, sex and university education in Table 5. We find respondents who self-selected into the video interviewing mode to be, on average, younger and more highly educated. We find no strong or consistent differences in terms of sex.

Table 4 Descriptive statistics of variables measuring attitudes towards immigration (imbgeco, imueclt, imwbcnt) by country and mode

Country	Mode	Valid *N*	*N*	imbgeco *M (SD)*	imueclt *M (SD)*	imwbent *M (SD)*
Estonia	in-person	1261	1302	5.25 (2.40)	5.22 (2.42)	5.16 (2.25)
	video	238	240	6.41 (2.27)	6.48 (2.16)	5.76 (1.91)
Finland	in-person	1314	1337	6.12 (2.19)	7.29 (1.93)	6.37 (2.00)
	video	236	240	6.40 (2.07)	7.40 (1.96)	6.37 (1.98)
Iceland	in-person	544	567	7.04 (1.99)	7.39 (1.96)	6.73 (1.98)
	video	323	333	7.14 (1.95)	7.62 (1.98)	6.55 (1.84)
Italy	in-person	2089	2183	5.00 (2.39)	5.13 (2.55)	4.56 (2.24)
	video	442	457	5.60 (2.46)	5.86 (2.59)	4.69 (2.27)
Netherlands	in-person	1156	1221	5.91 (1.90)	6.56 (1.97)	5.89 (1.76)
	video	230	248	6.18 (1.59)	6.93 (1.65)	6.02 (1.68)
Norway	in-person	898	920	6.20 (1.97)	6.68 (2.21)	6.37 (1.97)
	video	483	491	6.06 (1.92)	6.73 (2.08)	6.22 (1.95)

Table 5 Descriptive statistics of sociodemographic variables by country and mode

Country	Mode	Age *M (SD)*	Proportion women	Proportion university education
Estonia	in-person	53.88 (18.63)	.55	.29
	Video	39.53 (12.61)	.55	.61
Finland	in-person	54.70 (19.40)	.50	.33
	Video	40.98 (14.05)	.55	.58
Iceland	in-person	54.83 (18.86)	.51	.33
	Video	42.15 (15.79)	.53	.45
Italy	in-person	52.37 (19.01)	.53	.15
	Video	47.87 (16.61)	.51	.24
Netherlands	in-person	50.22 (18.81)	.48	.37
	Video	40.87 (14.69)	.53	.55
Norway	in-person	50.28 (18.84)	.48	.41
	Video	41.76 (15.37)	.52	.50

Measurement Invariance

Measurement invariance testing is suited for situations, such as ours, where respondents self-select into groups that differ in observed and unobserved characteristics. It tests whether the latent constructs are recovered equivalently despite such possible compositional differences (Meredith, 1993; Vandenberg & Lance, 2000; Davidov et al., 2014). Several families of measurement invariance tests are available (e.g. Comparative Fit Index/Root Mean Square Error of Approximation rules, Bayesian Multi-Group Confirmatory Factor Analysis, alignment, Multiple Indicator Multiple Cause model), but their performance varies with sample size among groups and model identification strategies. Because our mode groups are highly unbalanced (as Tables 3 and 4 show) and each construct is measured with only three indicators (yielding just-identified configural models), fit-index–based criteria risk inflated Type II error in detecting non-invariance (Chen, 2007) and Bayesian priors cannot be incorporated. We therefore adopt the bias-corrected bootstrap factor-ratio test proposed by Cheung and Lau (2012) within multi-group confirmatory factor analysis (MG-CFA), which remains robust when group sizes differ substantially.

The model represented in Figure 1 is equivalent to Equation 1, which specifies the configural model (Bollen, 1989; Meredith, 1993).

Equation 1 specifies a model for the relationship between three observed variables $y_{ij}^{g}, i = 1, 2, 3,$ the answers to the survey items and a latent (unobserved) variable, $\xi_j^g, j = 1, 2,$ representing the concepts we measure by the survey questions, namely, 1) social trust and 2) attitudes towards immigration. As multiple groups are involved, a multi-group confirmatory factor analysis (MG-CFA) is used, in our case, this is represented by $g = 1, 2,$ the mode groups: in-person or video interview, respectively. In this model, ν_{ij}^{g}, represents the intercepts, λ_{ij}^{g}, represents the factor loadings between the *i*th observed variable and the *j*th latent variable and δ_{ij}^{g} represents the disturbance terms. It is assumed that the disturbance terms have a mean of zero, they are uncorrelated with each other and with ξ_j^g as shown by (4) to (6).

$$y_{ij}^{g} = \nu_{ij}^{g} + \lambda_{ij}^{g}\xi_j^g + \delta_{ij}^{g} \tag{1}$$

For the latent means, α_j, we establish that

$$\alpha_j = E(\xi_j) \tag{2}$$

Furthermore, for the variance of ξ_j, we establish that

$$\phi_{jj} = var(\xi_j) \tag{3}$$

Moreover, for all models we make the following assumptions:

$$E(\delta_{ij}) = 0; \tag{4}$$

$$E(\delta_{ij}\xi_1) = 0; \tag{5}$$

$$E(\delta_{ij}\delta_{ij'}) = 0 \; or \; i \neq i' \tag{6}$$

Measurement invariance holds where the parameters (ν_{ij}^{g}, λ_{ij}^{g}) of the factor model for ξ_j^g are equal across groups. Metric invariance is established when λ_{ij}^{g} are equal for $g = 1, 2$ and scalar invariance is established when ν_{ij}^{g} are equal for $g = 1, 2$. Furthermore, for metric and scalar invariance, partial invariance can be achieved. Partial metric invariance refers to the loadings for some but not all items being equal. Partial scalar invariance refers to loadings and intercepts only being equal for some of the items, not all (Byrne et al., 1989).

When equality constraints are imposed on loadings and intercepts, the constrained model is usually compared with the unconstrained (configural) model using a likelihood-ratio test based on the change in chi-square, $\Delta\chi^2$. Yet $\Delta\chi^2$ is highly sample-size sensitive and may inflate type-I error even for minor parameter differences (Cheung & Rensvold, 2002). Consequently, to mitigate this problem, researchers also inspect other fit indices, such as comparative fit index (ΔCFI), root mean square error of approximation (ΔRMSEA) and standardized root mean squared residual (ΔSRMR), but these statistics lose power when group sizes are unbalanced, increasing type-II error risk and potentially masking non-invariance (Chen, 2007).

As dependency of global fit measures on sample size is a long historical challenge of structural equation modelling (SEM), other approaches have been suggested, for instance, to evaluate the models combining the fit index (ΔFI) with the detection of misspecifications using local fit measures (Van der Veld & Saris, 2011). However, in both global and local fit approaches, the employed cut-off values for defining misspecifications and thresholds for concluding that one model has a worse fit than another has been criticized as arbitrary (Cheung & Rensvold, 2002; Kline, 2016; Millsap, 2011). Recently, other more flexible alternatives to MG-CFA such as Bayesian MG-CFA (Muthén & Asparouhov, 2017) have been suggested. However, they are unsuitable for just-identified models, as there are no degrees of freedom for the incorporation of priors. Additionally, the degree of the sensitivity of Bayesian MG-CFA to sample size is being debated (Gelman, 2006) hence we do not consider this a suitable approach.

Another consideration is that invariance testing, and more generally, structural equation models, require using an item as a reference to define the unit of measurement and identify the models, that is, the loading of an arbitrarily chosen item is fixed to 1, e.g. $\lambda_{21} = 1$ and its intercept to zero, e.g. $\nu_{21} = 0$. Cheung and Rensvold (1999) argue that, with this choice, the selected item is implicitly assumed to be invariant. This implies that the choice of the item is critical, and if a non-invariant item is selected, the conclusions about the invariance of other items are likely to be incorrect (Johnson et al., 2009).

All the considerations outlined above led us to work with the bias-corrected (BC) bootstrap confidence interval approach (Cheung & Lau, 2012). This approach uses a single model as opposed to the likelihood ratio test (LRT) and Δ FI approaches, which estimate and compare different models. It offers a solution to test for measurement invariance using data with very unequal sample sizes. It uses maximum likelihood estimation with bootstrapping to evaluate the models, in combination with a factor-ratio test, that allows overcoming the challenges of the reference item, especially relevant in our data as the models have only three items each. The specific implementation of the BC bootstrap confidence interval approach is described in the next section.

Analytical Strategy

To test for metric and scalar measurement invariance following the BC bootstrap confidence interval approach (Cheung & Lau, 2012), we constrained the estimates of the parameters λ_{ij}^{g} for both mode groups, $g = 1$ for in-person interviews, $g = 2$, for video interviews to equality, within each country model. The factor-ratio test uses three constraints which are created using each item as reference. Equation 7 to Equation 9 illustrate the test for the concept of social trust. When y_{11}^{g} is the referent item, the hypotheses to test for metric invariance are:

$$\frac{\lambda_{21}^{1}}{\lambda_{11}^{1}} - \frac{\lambda_{21}^{2}}{\lambda_{11}^{2}} = 0 \tag{7}$$

$$\frac{\lambda_{31}^{1}}{\lambda_{11}^{1}} - \frac{\lambda_{31}^{2}}{\lambda_{11}^{2}} = 0 \tag{8}$$

When y_{2j}^{g} is the referent item, an additional hypothesis is tested:

$$\frac{\lambda_{31}^{1}}{\lambda_{21}^{1}} - \frac{\lambda_{31}^{2}}{\lambda_{21}^{2}} = 0 \tag{9}$$

As there are only three items, there are no other hypotheses to test when y_{3j}^{g} is the referent item. If zero falls outside of a confidence interval, the null hypothe-

sis of invariance can be rejected. Should zero fall outside of the confidence interval for only one of these constraints, we can establish partial invariance. Using the same procedure, a factor-ratio test is defined to test for scalar invariance, and confidence intervals are computed and evaluated, using the same rules: Should zero fall outside of the confidence interval for one of these constraints, we can establish partial invariance and should zero fall outside of the confidence interval for at least two of these constraints, this shows that there is non-invariance[3]. We estimate confidence intervals using the bias-corrected bootstrapping approach, for which 10,000 bootstrap samples are created, parameters are estimated for each bootstrap sample and calculations are made to the bootstrap distribution of the parameters to form the confidence intervals.

Results

Table 6 summarizes our findings. We show results by country, concept and type of invariance. We show the point estimates and the confidence intervals of the factor-ratio tests in a series of graphs. In order to facilitate the interpretation of the graphs, we simplify the notation of the equations from Section 4.3 in the Y-axis of the graphs.

Below, we present the original labels from our factor-ratio tests along with explanations.

1. $\lambda_{2.1}^{g1} - \lambda_{2.1}^{g2}$ represents the difference in the factor loading for item 2 between group 1 (in-person) and group 2 (video), when item 1 is used as the reference (i.e., its loading is fixed). Under measurement invariance, the relative contribution of item 2 compared to item 1 should be identical across groups.

Y-axis label: *Loading Diff. for Item 2 (Ref: Item 1)*

2. $\lambda_{3.1}^{g1} - \lambda_{3.1}^{g2}$ this is the analogous difference for item 3's loading, again with item 1 as the reference. It tests whether the relative loading of item 3 is equivalent across groups.

Y-axis label: *Loading Diff. for Item 3 (Ref: Item 1)*

3. $\frac{\lambda_{3.1}^{g1}}{\lambda_{2.1}^{g1}} - \frac{\lambda_{3.1}^{g2}}{\lambda_{2.1}^{g2}}$ this ratio compares the loading of item 3 relative to item 2 across groups.

It serves as an alternative anchoring check (using item 2 as the reference) to ensure that invariance holds regardless of which item is selected as the anchor.

Y-axis label: *Loading Ratio Diff. for Item 3 vs. Item 2*

[3] For a detailed mathematical derivation of these ratio-based constraints, see Cheung and Lau (2012), Appendices B and D.

4. $\nu_2^{g1} - \nu_2^{g2}$ this indicates the difference in the intercept for item 2 between groups when item 1 is used as the reference (with its intercept fixed). Under scalar invariance, the baseline (intercept) of item 2 should be identical across groups.

Y-axis label: *Intercept Diff. for Item 2 (Ref: Item 1)*

5. $\nu_3^{g1} - \nu_3^{g2}$ this is the difference in the intercept for item 3 between groups, with item 1 as the reference. It reflects whether the baseline response level for item 3 is comparable across modes.

Y-axis label: *Intercept Diff. for Item 3 (Ref: Item 1)*

6. $\Delta_{\text{int},3} - \left(\frac{\lambda_{3,1}}{\lambda_{2,1}}\right)\Delta_{\text{int},2}$ here, $\Delta_{\text{int},3}$ is the difference in intercept for item 3 and $\Delta_{\text{int},2}$ is the difference in intercept for item 2 (both computed with item 1 as the reference). Multiplying $\Delta_{\text{int},2}$ by the ratio $\lambda_{3,1}/\lambda_{2,1}$ re-scales the intercept difference for item 2 to reflect the relationship between items 3 and 2. This adjusted difference tests whether, when item 2 is used as the reference, the intercept for item 3 remains invariant.

Y-axis label: *Adjusted Intercept Diff. for Item 3 (Ref: Item 2)*

Table 6 Overview of findings by country, concept and type of invariance

	Social trust		Attitudes towards immigration	
	Metric	Scalar	Metric	Scalar
Estonia	invariance	partial invariance	non-invariance	–
Finland	invariance	invariance	partial invariance	–
Iceland	invariance	non-invariance	invariance	invariance
Italy	invariance	invariance	invariance	invariance
Netherlands	invariance	invariance	non-invariance	–
Norway	invariance	invariance	invariance	partial invariance

The graphs show differences in the parameters of interest, point estimates of loadings for metric invariance models and point estimates of intercepts for scalar invariance models. We include their confidence intervals. We assess if the estimates and their confidence intervals include zero or not. Where confidence intervals overlap with zero, we conclude invariance, because there are no significant differences across parameters. Where zero falls outside of the confidence intervals for one parameter comparison, we regard this as partial invariance. Where zero falls outside of the confidence intervals for at least two

parameter comparisons, we regard this as non-invariance. In the graphs, for the cases where metric invariance is not established, intervals are not shown for scalar invariance, as it is not even tested for, and the respective cells remain empty in Table 6, as well.

For Italy, we find metric as well as scalar invariance across the two modes for both concepts as Figure 2 does not show statistically significant differences in parameters. For Norway, both metric and scalar invariance is established for the concept of social trust (see Figure 3). For attitudes towards immigration, metric and partial scalar invariance is established. Iceland shows a different result, in that metric and scalar invariance are established for attitudes towards immigration but only metric invariance for the concept of social trust (see Figure 4). For Finland, we find metric and scalar invariance across the two modes for social trust but we conclude only partial metric invariance for the concept of attitudes towards immigration (see Figure 5). For the Netherlands, results show metric and scalar invariance for the concept of social trust, but metric non-invariance for the concept of attitudes towards immigration (see Figure 6). The latter also holds for Estonia (see Figure 7). Lastly, for the concept of social trust, metric and partial scalar invariance could be established for Estonia (see Figure 7).

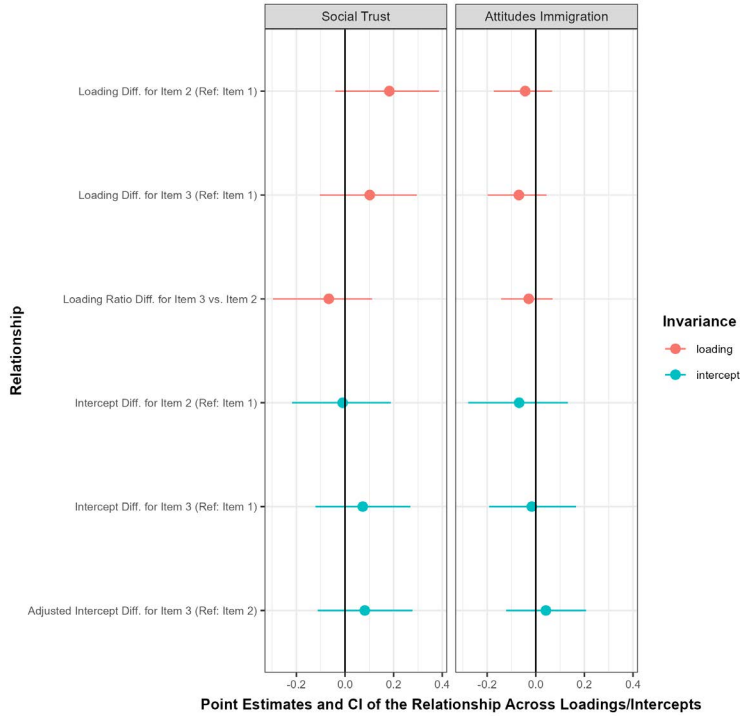

Figure 2 Results invariance testing Italy

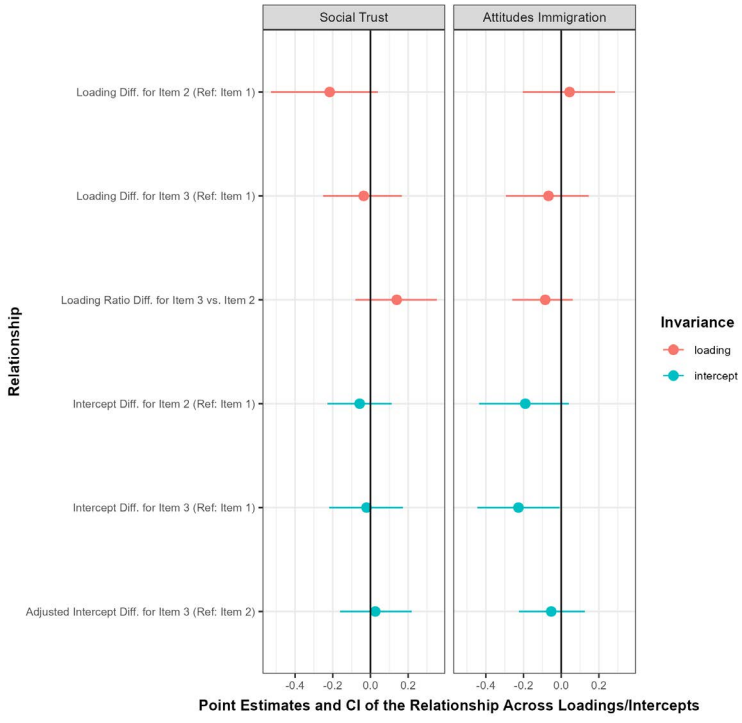

Figure 3 Results invariance testing Norway

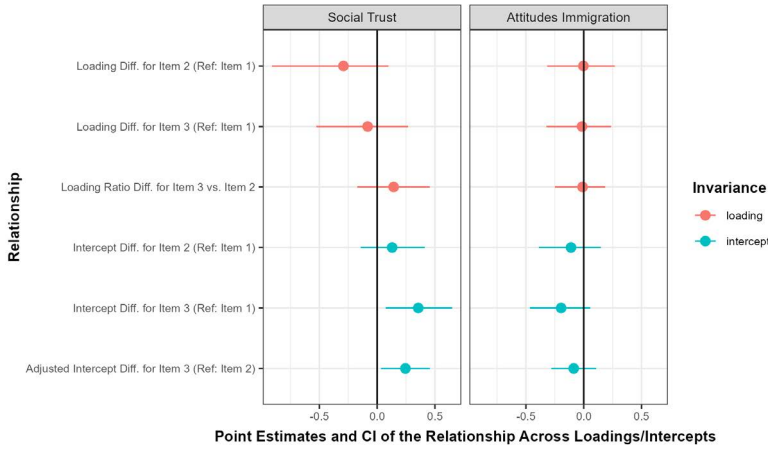

Figure 4 Results invariance testing Iceland

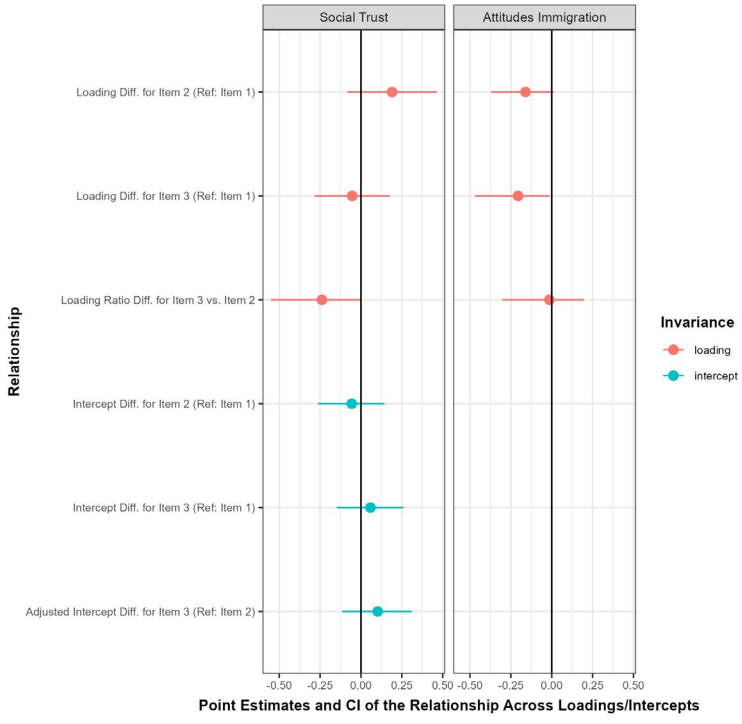

Figure 5 Results invariance testing Finland

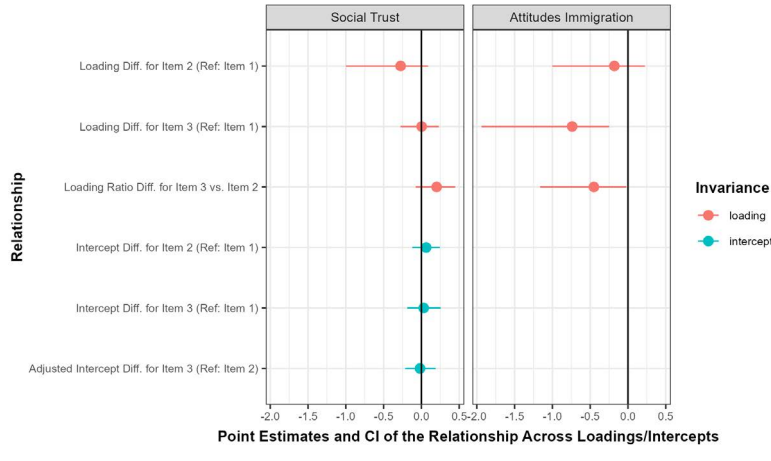

Figure 6 Results invariance testing Netherlands

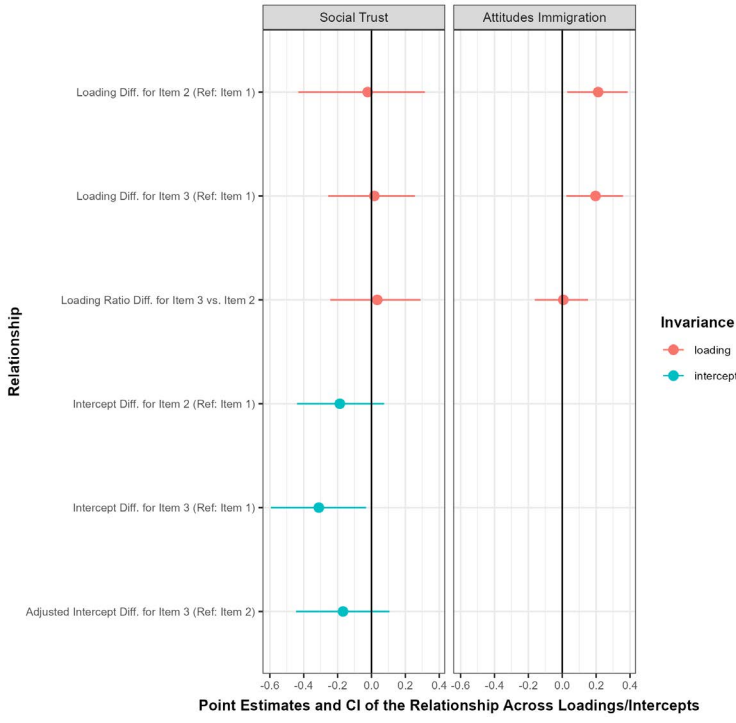

Figure 7 Results invariance testing Estonia

Discussion and Conclusion

The COVID-19 pandemic increased the diffculties of collecting high-quality human data, especially by the means of in-person interviews. New interview formats that do not require respondents and interviewers being in the same place emerged or consolidated, and are nowadays still in use. As noted earlier, ESS Round 10 offered respondents the option of an online video interview as an alternative to an in-person interview (Ghirelli et al., 2024). A body of research also emerged, such as this special issue, evaluating the costs and benefits of the new interview formats and their potential effects on the answers (Deakin & Wakefield, 2014; Oates et al., 2022). We contribute to this research using the ESS Round 10 data and studying the measurement equivalence of data collected via live video and in-person interviewing. As respondents were not randomly allocated to the two groups, but rather self-selected into an interview mode, it cannot be assumed that the data would be equivalent, hence this should be tested (Meredith, 1993). Ensuring the equivalence of the data derived using these modes is crucial for their effective employment in the analysis of social attitudes. This is

especially important when the characteristics of the concepts make them prone to bias by response behaviors. The concepts of social trust and attitudes towards immigration differ in the level of social desirability bias among them, with attitudes towards immigration showing larger bias (Creighton et al., 2019).

Measurement invariance between live video and in-person interviews is particularly relevant when considering how different modes of data collection influence response behavior. Video interviewing and in-person interviewing differ regarding the interaction with the administrator and regarding tools used in the interviews, mainly, the intermediation of a screen in video interview. We argue that for concepts where social desirability is present, such as attitudes towards immigration, it is less pronounced in video interviews. Social desirability is lower in modes of data collection where the social presence of an interviewer is weaker (Heerwegh, 2009). The social presence may be perceived as weaker in a context when respondent and interviewer communicate via videochat (Moallem, 2015) compared to a context where the interviewer is physically present in a room with respondents.

Overall, the results can be seen as encouraging. In general, measurement invariance can indeed be established between in-person and video interview mode, across countries and concepts in the ESS Round 10 data. In consequence the data is highly comparable. Particularly for Italy, both metric and scalar invariance are established for both assessed concepts. In practical terms, this means that the data can be aggregated and used as a single dataset for the analysis of the concepts of interest, namely social trust and attitudes towards immigration, and that the dataset can be aggregated as well for the analysis of the means of the concepts.

Measurements for the concept of social trust have shown full metric invariance across in-person and video interviewing modes for all six assessed countries: Estonia, Finland, Iceland, Italy, the Netherlands and Norway (see overview in Table 6). This means that each item contributes consistently to the concept of social trust across modes and countries. The measurements for the concept of social trust are full scalar invariant across video and in- person mode for four out of the six countries (Finland, Italy, the Netherlands and Norway), with Estonia showing partial invariance and Iceland non-invariance. This means that, only for four out of the six countries, the data can be aggregated, and observed mean responses can be computed across the two modes.

For the concept of attitudes towards immigration, the results are somewhat less encouraging. Three out of the six countries show full metric invariance across mode groups. Where full metric invariance is present, the structure of the underlying construct is preserved regardless of the mode of data collection. As a result, analysts can confidently combine data from in-person and video interviewing to study the contribution of each item into the construct in Iceland, Italy and Norway, but not in Estonia, Finland and the Netherlands. In

cases where metric invariance is not established, we would advise using latent models, or, as a less optimal solution because it decreases the number of cases, excluding observations from the video interview mode since the relationship of the variables differs across modes. In two out of the three cases showing metric invariance, scalar invariance is also established, namely for Iceland and Italy. For Norway, only partial scalar invariance could be established. Where scalar invariance is not established, we advise against aggregating the data of the two modes to calculate mean values. Instead, analysts could calculate the latent means, taking into account the method variance contribution of both mode groups. This is because the differences in the mean response to items may not purely reflect true differences in the underlying construct across groups, and method variance should be accounted for.

Our findings are neither homogeneous across countries nor across concepts. Measurement invariance across in-person and video interview mode are sensitive to the concept being measured.

It has been found before that measurement invariance is not easily established for attitudes towards immigration if measured in different contexts. Nickel and Weber (2024) study the concept of attitudes towards immigration, and report that metric invariance was established for most ESS countries in Rounds 1 to 9 but scalar invariance only for about half of them. Measurement invariance seems sensitive to contextual influences. Our findings are in line with this result, though the different contexts in our study refer to different modes within countries. Throughout this article, we have presented the argument that it is precisely because of the differences in social desirability between the two concepts that we observe differences in the levels of invariance achieved. On the one hand, attitudes towards immigration presents a larger bias than social trust, on the other, video interviews imply a weaker social presence of the interviewers, than in-person interviews, potentially reducing this bias.

We indeed observe higher levels of invariance for social trust, this means that the data is more comparable than the data for attitudes towards immigration. Our interpretation for this result is that in video mode, respondents feel less pressured to give a socially expected answer, therefore the data is less comparable to an in-person interview. However, more research is needed to explore whether this interpretation is correct.

There are other potential explanations. It could be revealing to investigate questionnaires, interviewer instructions, protocols and similar documentation of the cases where invariance was not consistently reached (especially Estonia, Finland, Iceland and the Netherlands) to see if one can find hints as to how the implementation of these modes might have differed here from those countries where invariance was consistently reached.

Assessing measurement invariance in video and in-person interviews and assessing mode effects are two different subjects of study. In this research, we

cannot test simultaneously for mode effects and measurement invariance across modes. Testing for mode effects would require random allocation of respondents into either video or in-person mode. Invariance testing assesses for the comparability of the data in the presence of non-random allocation into groups, in which case characteristics of the survey participants, such as age, gender, employment status, home ownership, and ethnicity, may present differences across interviewing modes (Rowen et al., 2022).

References

Adam, F. (2008). Mapping social capital across Europe: findings, trends and methodological shortcomings of cross-national surveys. *Social Science Information, 47*(2), 159-186. https://doi.org/10.1177/0539018408089077

Boal, A. L., Vaughan, C. A., Sims, C. S., & Miles, J. N. V. (2017). Measurement invariance across administration mode: Examining the posttraumatic stress disorder (PTSD) checklist. *Psychological Assessment, 29*(1), 76–86. https://doi.org/10.1037/pas0000301

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons. https://doi.org/10.1002/9781118619179

Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The transition to computer-based testing in large-scale assessments: Investigating (partial) measurement invariance between modes. *Psychological Test and Assessment Modeling, 58*(4), 597–616.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

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*, 309-328. https://doi.org/10.1146/annurev.soc.012809.102651

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods, 15*(2), 167–198. https://doi.org/10.1177/1094428111421987

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. J*ournal of Management, 25*(1), 1-27. https://doi.org/10.1177/014920639902500101

Claassen, J., Lenzner, T., Höhne, J. K., & Ziller, C. (2025). A survey mode of the future? Investigating respondents' willingness to participate in self-administered video-based web surveys. *Methods, Data, Analyses*, 1–30. https://doi.org/10.12758/mda.2025.10

Conrad, F. G., Schober, M. F., Hupp, A. L., West, B. T., Larsen, K. M., Ong, A. R., & Wang, T. (2023). Video in survey interviews: Effects on data quality and respondent experience. *Methods, Data, Analyses*, 17(2), 135–170. https://doi.org/10.12758/mda.2022.13

Creighton, M. J., Schmidt, P., & Zavala-Rojas, D. (2019). Race, wealth and the masking of opposition to immigrants in the Netherlands. *International Migration, 57*, 245-263. https://doi.org/10.1111/imig.12519

Davidov, E., & Semyonov, M. (2017). Attitudes toward immigrants in European societies. *International Journal of Comparative Sociology, 58*(5), 359-366. https://doi.org/10.1177/0020715217732183

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review Sociology, 40*, 55-75. https://doi.org/10.1146/annurev-soc-071913-043137

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: Exact versus approximate measurement equivalence. *Public Opinion Quarterly, 79*(S1), 244–266. https://doi.org/10.1093/poq/nfv008

de Villiers, C., Farooq, M. B., & Molinari, M. (2022). Qualitative research interviews using online video technology – challenges and opportunities. *Meditari Accountancy Research, 30*(6), 1764–1782. https://doi.org/10.1108/MEDAR-03-2021-1252

Deakin, H., & Wakefield, K. (2013). Skype interviewing: Reflections of two PhD researchers. *Qualitative Research, 14*(5), 603-616. https://doi.org/10.1177/1468794113488126

Delhey, J., & Newton, K. (2005). Predicting cross-national levels of social trust: Global pattern or nordic exceptionalism? *European Sociological Review, 21*(4), 311–327. https://doi.org/10.1093/esr/jci022

Durrant, G., Kocar, S., Brown, M., Hanson, T., Sanchez, C., Wood, M., Taylor, K., Tsantani, M., & Huskinson, T. (2024). *Live video interviewing: Evidence of opportunities and challenges across seven major UK social surveys* (Survey Futures Working Paper No. 1). University of Essex, Institute for Social and Economic Research. https://www.iser.essex.ac.uk/wp-content/uploads/files/working-papers/survey-futures/2024-01.pdf

Endres, K., Hillygus, D. S., DeBell, M., & Iyengar, S. (2023). A randomized experiment evaluating survey mode effects for video interviewing. *Political Science Research and Methods, 11*(1), 144–159. https://doi.org/10.1017/psrm.2022.30

European Social Survey European Research Infrastructure (ESS ERIC). (2023). *ESS10 integrated file*, edition 3.2 [Data set]. Sikt – Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ess10sce03_2

Faucett, H. A., Lee, M. L., & Carter, S. (2017). I should listen more: Real-time sensing and feedback of non-verbal communication in video telehealth. *Proceedings of the ACM on Human-Computer Interaction, 1*(CSCW), 1–19. https://doi.org/10.1145/3134679

Fernandez, E., Woldgabreal, Y., Day, A., Pham, T., Gleich, B., & Aboujaoude, E. (2021). Live psychotherapy by video versus in-person: A meta-analysis of efficacy and its relationship to types and targets of treatment. *Clinical Psychology & Psychotherapy, 28*(6), 1535–1549. https://doi.org/10.1002/cpp.2594

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis, 1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Ghirelli, N., Lynn, P., Xena, C., Dorer, B., Ambler, M. L., Schwarz, H., Hanson, T., Kappelhof, J., Flore, P., Kessler, G., Lebedev, D., Briceno-Rosas, R., Frank, L.-H., Rød, L.-M., & Øvrebø, O.-P. (2024). *ESS10 overall face-to-face fieldwork and data quality report*. GESIS – Leibniz Institute for the Social Sciences. https://www.europeansocialsurvey.org/sites/default/files/2024-09/ESS10_Quality_Report.pdf

Gordon, H. S., Solanki, P., Bokhour, B. G., & Gopal, R. K. (2020). "I'm not feeling like I'm part of the conversation" Patients' perspectives on communicating in clinical video telehealth visits. *Journal of General Internal Medicine, 35*(6), 1751–1758. https://doi.org/10.1007/s11606-020-05673-w

Gorgievski, M. J., Van der Heijden, B. I. J. M., & Bakker, A. B. (2019). Effort-reward imbalance and work-home interference: a two-wave study among European male nurses. *Work & Stress, 33*(4), 315–333. https://doi.org/10.1080/02678373.2018.1503358

Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research, 22*(1), 111–121. https://doi.org/10.1093/ijpor/edn054

Herreros, F., & Criado, H. (2008). The state and the development of social trust. *International Political Science Review, 29*(1), 53-71. https://doi.org/10.1177/0192512107083447

Inglehart, R. (1999). Trust, well-being and democracy. In M. E. Warren (Ed.), Democracy and Trust (pp. 88–120). chapter, Cambridge: Cambridge University Press.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(4), 642–657. https://doi.org/10.1080/10705510903206014

Joseph C., Garruba M., & Melder A. (2018). Patient satisfaction of telephone or video interpreter services compared with in-person services: A systematic review*. *Australian Health Review, 42*, 168–177. https://doi.org/10.1071/AH16195

Joshi, A., Bloom, D. A., Spencer, A., Gaetke-Udager, K., & Cohan, R. H. (2020). Video interviewing: a review and recommendations for implementation in the era of COVID-19 and beyond. *Academic Radiology, 27*(9), 1316–1322. https://doi.org/10.1016/j.acra.2020.05.020

Kaasa, A., & Parts, E. (2008). Individual-level determinants of social capital in Europe: Differences between country groups. *Acta Sociologica, 51*(2), 145-168. https://doi.org/10.1177/0001699308090040

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Methodology in the social sciences. Guilford Press.

Letki, N., & Evans, G. (2005). Endogenizing social trust: Democratization in east-central Europe. *British Journal of Political Science, 35*(3), 515–529. https://doi.org/10.1017/S000712340500027X

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825

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, 52*(3), 1401-1419. https://doi.org/10.1177/00491241221091755

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

Moallem, M. (2015). The impact of synchronous and asynchronous communication tools on learner self-regulation, social presence, immediacy, intimacy and satisfaction in collaborative online learning. *The Online Journal of Distance Education and e-Learning, 3*(3), 55–77.

Muthén, B., & Asparouhov, T. (2017). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research, 47*(4), 637-664. https://doi.org/10.1177/0049124117701488

Nickel, A., & Weber, W. (2024). Measurement invariance and quality of attitudes towards immigration in the European Social Survey. *Methods, Data, Analyses, 18*(2), 213–248. https://doi.org/10.12758/mda.2024.04

Nießen, D., Beierlein, C., Rammstedt, B., & Lechner, C. M. (2020). An English-language adaptation of the interpersonal trust short scale (KUSIV3). *Measurement Instruments for the Social Sciences, 2*(1), Article e11019. https://doi.org/10.1186/s42409-020-00016-1

Oates, M., Crichton, K., Cranor, L., Budwig, S., Weston, E. J., Bernagozzi, B. M., & Pagaduan, J. (2022). Audio, video, chat, email, or survey: How much does online interview mode matter? *PLoS ONE, 17*(2), Article e0263876. https://doi.org/10.1371/journal.pone.0263876

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

Pirralha, A., & Weber, W. (2020). Correction for measurement error in invariance testing: An illustration using SQP. *PLoS ONE, 15*(10), Article 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

Reeskens, T., & Hooghe, M. (2008). Cross-cultural measurement equivalence of generalized trust. Evidence from the European social survey (2002 and 2004). *Social Indicators Research, 85*, 515–532. https://doi.org/10.1007/s11205-007-9100-z

Revilla, M. A. (2012). Measurement invariance and quality of composite scores in a face-to-face and a web survey. *Survey Research Methods, 7*(1), 17–28. https://doi.org/10.18148/srm/2013.v7i1.5098

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*, 2139–2159. https://doi.org/10.1007/s11135-021-01098-7

Rothstein, B., & Uslaner, E.M. (2005). All for one: Equality, corruption, and social trust. *World Politics 58*(1), 41-72. https://dx.doi.org/10.1353/wp.2006.0022

Rowen, D., Mukuria, C., Bray, N., Carlton, J., Longworth, L., Meads, D., O'Neill, C., Shah, K., & Yang, Y. (2022). Assessing the comparative feasibility, acceptability and equivalence of video-conference interviews and face-to-face interviews using the time trade-off technique. *Social Science & Medicine, 309*, Article 115227. https://doi.org/10.1016/j.socscimed.2022.115227

Saris, W. E., & Gallhofer, I. (2014). *Design, evaluation, and analysis of questionnaires for survey research*. John Wiley & Sons. https://doi.org/10.1002/9781118634646

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

Schober, M. F., Conrad, F. G., Hupp, A. L., Larsen, K. M., Ong, A. R., & West, B. T. (2020). Design considerations for live video survey interviews. *Survey Practice, 13*(1). https://doi.org/10.29115/SP-2020-0014

Schober, M. F., Okon, S., Conrad, F. G., Hupp, A. L., Ong, A. R., & Larsen, K. M. (2023). Predictors of willingness to participate in survey interviews conducted by live video. *Technology, Mind, and Behavior, 4*(2), 215–229. https://doi.org/10.1037/tmb0000100

Sherman, L. E., Michikyan, M., & Greenfield, P. M. (2013). The effects of text, audio, video, and in-person communication on bonding between friends. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 7*(2), Article 3. https://doi.org/10.5817/CP2013-2-3

Smith, S. N., Fisher, S. D., & Heath, A. (2011). Opportunities and challenges in the expansion of cross-national survey research. *International Journal of Social Research Methodology, 14*(6), 485–502. https://doi.org/10.1080/13645579.2011.611386

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality & Quantity, 43*, 599–616. https://doi.org/10.1007/s11135-007-9143-x

Van de Vijver, F., Avvisati, F., Davidov, E., Eid, M., Fox, J. P., Le Donné, N., Lek, K., Meuleman, B., Paccagnella, M., & van de Schoot, R. (2019). *Invariance analyses in large-scale studies* (OECD Education Working Paper No. 201). OECD Publishing. https://doi.org/10.1787/254738dd-en

Van de Vijver, F. J., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology, 54*(2), 119–135. https://doi.org/10.1016/j.erap.2003.12.004

Van der Veld, W. M., & Saris, W. E. (2011). Causes of generalized social trust. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 207–247). Routledge.

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

Villar, A., & Fitzgerald, R. (2017). Using mixed modes in survey research: Evidence from six experiments in the ESS. In M. J. Breen (Ed.), *Values and identities in Europe* (pp. 299–336). Routledge.

West, B. T., Ong, A. R., Conrad, F. G., Schober, M. F., Larsen, K. M., & Hupp, A. L. (2022). Interviewer effects in live video and prerecorded video interviewing. *Journal of Survey Statistics and Methodology, 10*(2), 317–336. https://doi.org/10.1093/jssam/smab040

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika, 81*(4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0

Zmerli, S., & Newton, K. (2008). Social trust and attitudes toward democracy. *Public Opinion Quarterly, 72*(4), 706–724. https://doi.org/10.1093/poq/nfn054

Appendix

Global Fit Indices of Scalar Invariance Models by Concept and Country

Table A1 Global fit indices of scalar invariance models for concept social trust by country

Model	x^2	df	p	rmsea	cfi
Estonia	0.07	2	.96	0.00	1
Finland	4.91	2	.09	0.04	1
Italy	3.84	2	.15	0.03	1
Netherlands	3.33	2	.19	0.03	1
Norway	3.72	2	.16	0.03	1

Table A2 Global fit indices of scalar invariance models for concept attitude towards immigration by country

Model	x^2	df	p	rmsea	cfi
Iceland	0.02	2	.99	0.00	1
Italy	1.75	2	.42	0.00	1
Norway	1.39	2	.50	0.00	1

Online First publication date: 2026-05-13