European Journal of Social Psychology

WILEY

EASP

RESEARCH ARTICLE OPEN ACCESS

Measurement Invariance of Human Values Across Time and Countries—An Analysis Using 11 Rounds of the European Social Survey

Ronald Fischer[1] | Johannes A. Karl[2]

[1]Cognitive Neuroscience and Neuroinformatics Unit, D'Or Institute for Research and Education, São Paulo, Brazil | [2]Graduate School of Business, Stanford University, Stanford, California, USA

Correspondence: Ronald Fischer (ronald.fischer@idor.org)

Received: 19 December 2024 | **Accepted:** 19 June 2025

Funding: The first author would like to acknowledge the support by the IDOR/Pioneer Science Initiative (www.pioneerscience.org) and financial support from the University of Western Australia and Advanced Studies Institute of the University of Western Australia, which facilitated the work resulting in this manuscript.

Keywords: culture | European social survey | invariance | longitudinal | values

ABSTRACT

Values have become central to the study of social and psychological processes across cultures and across time. To date, there has been no conclusive analysis of the cross-cultural and cross-temporal comparability of values in Europe. We conduct invariance tests of the Portrait Values Questionnaire collected as part of the European Social Survey (ESS, $k = 261$ samples, $N = 374{,}565$) from 2002 to 2024. Using confirmatory factor analysis, we found that the theoretical value model fits the overall data. Ignoring temporal information, covariances and mean value patterns can be compared across countries. Testing separate country samples collected in specific rounds, we found evidence that sample-specific results may not be comparable. Considering cross-temporal invariance, only 17 out of 35 countries with multiple participation in the ESS showed metric invariance. We observed patterns of deteriorating model fit and changes in the basic value structure over time.

1 | Introduction

Values are expressed goals of importance for individuals and communities and are a cornerstone of public and cultural life. Not surprisingly, researchers have long focused on values when trying to study similarities and differences within and across societies (Hofstede 1980; Inglehart 1997; Parsons and Shils 1951). The most widely used contemporary theory of human values was proposed by Schwartz (1992). A measure based on this theory has been included in the European Social Survey (ESS) (Schwartz et al. 2015), which has been running since 2002. Drawing nationally representative samples on a bi-annual basis, this offers unprecedented insights into the value priorities of populations. In order to compare insights from these data across either countries or time, it is important to test whether the measurement properties are comparable across samples and across time. The question of comparability of data is a cornerstone of science and carries implications for any inferences about the data (Leitgöb et al. 2023; Vandenberg and Lance 2000). Comparability is often tested via statistical invariance, with different levels of invariance carrying implications for the interpretation of any differences in correlations and means. The previous research has examined cross-cultural comparability of values in earlier rounds of the ESS (Bilsky et al. 2011; Cieciuch et al. 2018; Davidov et al. 2008, 2018), but to the best of our knowledge, there is no study that systematically examines both the cross-cultural and cross-temporal invariance of the value measures using the full data set.

Therefore, our primary aim is to provide reference parameters on the cross-cultural and cross-temporal invariance of the 21-item version of the Portrait Value Questionnaire over rounds 1–11 in the ESS, effectively spanning the period from 2002 to 2024. We use multi-group confirmatory factor analysis (MG-CFA), which is the most widely used tool for invariance tests (Leitgöb et al. 2023). To guide researchers on what comparisons may be possible, we provide information on the fit of the value theory slicing the data in various different ways: (a) a CFA model fit to the overall ESS data set (ignoring all cultural and temporal information), (b) a cross-cultural invariance test comparing all countries with each other (ignoring time and pooling data across all rounds), (c) testing the fit of the theory to the data via a separate CFA model for each individual country sample in each round as well as constraining the loadings of items on the latent factors according to different empirical solutions (see the method section for details) and (d) importantly, a cross-temporal invariance test across all rounds available within each country. This information is essential for interpreting findings from value research across countries, for specific rounds and across time when using the ESS. The data will be useful for researchers when evaluating with which level of confidence what kind of comparisons can be made with these data. To contextualize the importance of the measure and the associated data set, the proposal for this version of the value survey (Schwartz 2003) has been cited more than 1475 times (November 2024), and the original instrument (Schwartz et al. 2001) has been cited more than 3500 times (metrics from Google Scholar). The widespread usage of this instrument and data set makes it important to provide some guidance on the interpretability of the scores across countries and across time.

We would like to emphasize that we consider assessments of invariance as an important part of any empirical work, with invariance metrics providing essential information about the theory as well as temporal and cultural dynamics (Fischer et al. 2021, 2023; Fischer and Rudnev 2024). Even if measures were to show adequate levels of invariance or if measures fail to meet commonly accepted levels of invariance for a specific invariance comparison threshold, variability in fit estimates can be highly informative (Fischer and Karl 2023; Karl and Fischer 2022), because an invariance test constitutes a test of a causal model (Sterner et al. 2024). To provide some guidance for further exploration, we provide preliminary evidence of systematic variability in fit indices across the eleven rounds of the ESS using time and mean survey responses as predictors.

In the following, we present the Schwartz value theory, provide a brief overview of the invariance paradigm and the methods that we are employing, and review previous explorations of invariance with values in the ESS.

1.1 | The Schwartz Value Theory and Portrait Value Questionnaire

In his influential work, Schwartz (1992) outlined 10 individual-level value types, each organized within a circular framework to reflect their relationships. These values are thought to be organized in a quasi-circle. Moving around the circle, power (PO) values focus on social status, prestige and exerting control over others and resources; achievement (AC) values emphasize personal success by demonstrating competence in accordance with social standards; hedonism (HE) values focus on the pursuit of pleasure and personal gratification; stimulation (ST) values involve seeking excitement, novelty and challenge in life; self-direction (SD) values reflect a preference for independence in thought and action, with a drive to choose, create and explore; universalism (UN) values emphasize a commitment to understanding, appreciating and protecting all people and nature; benevolence (BE) values prioritize preserving and enhancing the well-being of those in close, personal relationships; tradition (TR) values reflect respect, commitment and acceptance of cultural and religious customs and beliefs; conformity (CO) values restrain actions and impulses that might disrupt or harm others and maintain social expectations; and security (SE) values focus on ensuring safety, harmony and stability within society, relationships and oneself. The circular structure is organized along two primary axes (Schwartz and Cieciuch 2022).

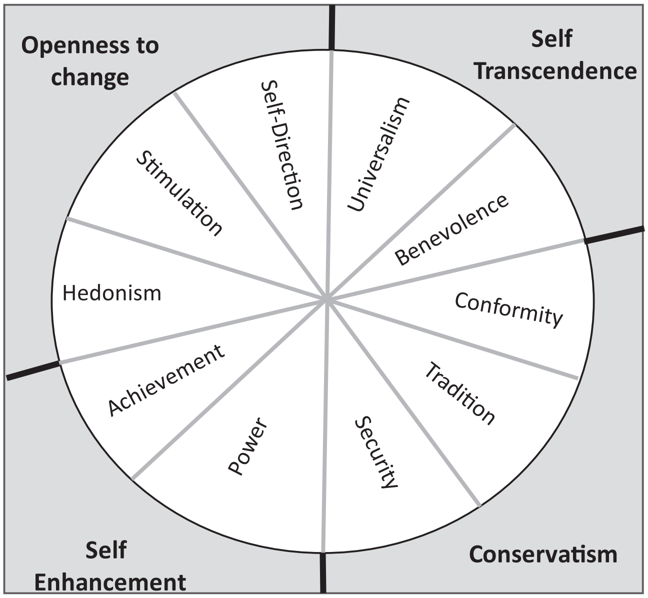

FIGURE 1 | Schematic representation of the Schwartz Value Structure.

The first axis, labelled openness to change versus conservation, contrasts values such as stimulation and self-direction (openness) with tradition, conformity and security (conservatism), emphasizing the pursuit of one's own impulses and inclinations over concern for the concerns of others and the preservation of the traditional order. The second dimension is labelled self-enhancement versus self-transcendence and separates power and achievement (self-enhancement) from benevolence and universalism (self-transcendence). Thus, this axis contrasts a motivation to put the interests of others ahead of personal interests with a motivation to get ahead of others. The motivation associated with hedonism values is consistent with both openness to change and self-enhancement, but empirically it is more closely associated with openness to change (Schwartz and Boehnke 2004). The circular structure (see Figure 1) highlights the compatible and conflicting nature of the values within this arrangement, with values that share a motivational concern placed closer together and values that are motivationally independent placed orthogonally to each other (Schwartz 1992).

Schwartz and Bilsky (1987) suggested that these 10 values and their organization respond to three universal human needs: biological needs, requirements for interpersonal coordination and social demands for group welfare and smooth functioning. The theory has been proposed and supported by survey research conducted on all inhabited continents (Schwartz 2006; Schwartz et al. 2012), and there is also emerging evidence that the same value conflicts and compatibilities can be identified in reaction time experiments, studies of neural activation and mental representations (Brosch et al. 2018; Coelho et al. 2019; Leszkowicz et al. 2017, 2021; Teed et al. 2020).

A number of different instruments have been developed over the years that measure values in-line with this descriptive theory. The Portrait Value Questionnaire has become the most widely used measure (Schwartz et al. 2001), and a short version containing 21 items has been included in the ESS since its inception (Schwartz 2003; Schwartz et al. 2015). It presents short scenarios that describe individuals and their actions and then asks respondents to indicate how similar or dissimilar these fictitious individuals are to them (see Table 1).

Overall, this descriptive theory provides a promising starting point for comparing individuals' motivations across societies and over time. There has been a significant investment in resources and efforts in collecting data from human populations since the emergence of the full theory in 1992, with different instruments inspired by the theory being implemented in a number of large-scale surveys (Bilsky et al. 2011; Davidov et al. 2008; Welzel 2010). In order to interpret these data sets across populations and across time, it is, however, first necessary to examine the comparability of the data (Poortinga 1989; van de Vijver and Leung 2000, 2021). If the measurement of this theoretical structure were found to be fully invariant, we would be able to track changes in values and relate them to a variety of theoretical processes, including differences in social organization and economic or cultural cycles. In the presence of some level of non-invariance or non-comparability, greater caution is required because additional processes may be at play that influence how value results can be interpreted. In the next section, we describe the main levels of invariance that have been discussed in the literature and the inferences that are permitted at each level of invariance.

1.2 | Measurement Invariance

Invariance testing helps determine whether a measurement tool (such as a survey or test) functions similarly across different groups. Research typically discusses three main levels of invariance, although additional levels can be identified (Boer et al. 2018; Fischer and Karl 2019; Fontaine 2005; Leitgöb et al. 2023; Meredith 1993; van de Vijver and Leung 2021; Vandenberg and Lance 2000).

There are two distinct traditions that have led to essentially analogous classifications of comparability. First, Poortinga (1989) and Van De Vijver and Poortinga (1982) developed a hierarchy of measurement equivalence based on an analysis of sources of bias in measurement and linking comparability concerns to classical scaling types (Stevens 1946). Measurement equivalence in this perspective examines a researcher's confidence in interpreting the numerical values of theoretical concepts represented by specific indicators in different measurement situations. It is concerned with assessing whether, for example, a score of 3 on a scale of 1–6 on a particular value item is comparable across groups or over time in its expression of the underlying value type.

Central to this discussion of equivalence is scaling theory (Stevens 1946), which defines measurement scales by the transformations each scale permits (e.g., nominal, ordinal, interval and ratio scales). Functional equivalence, in its broadest form, is concerned with the quality of a construct and whether or not it can be considered present in a given cultural context. This maps most directly onto nominal scales, which represent a binary qualitative classification (e.g., quality present or absent). Applying this argument to values research, we could ask whether or not a particular value is cognitively available to individuals in a particular community to guide decisions. No decisions can be made about the overall levels of the value, only its presence or absence in different cultural groups or across time. For example, some authors have argued that the value of equality only emerged as a guiding human concern during the Enlightenment (Graeber and Wengrow 2021). Structural equivalence assesses whether the same observed variables can measure a construct similarly across groups. Thus, it is about the specific value instantiation and whether the same stimuli allow ranking of individuals with respect to the intended construct. This maps onto ordinal scales that allow ranking of individuals but do not assume equal intervals between points. One classic example of potential structural non-equivalence is freedom, which may imply different notions of freedom: a negative form that implies freedom from oppression or constraints and a positive form of freedom to say or express whatever a person wants without being constrained or controlled (Berlin 2000).

Metric equivalence is the next level up and is concerned with the equidistance of measurement observations. It can therefore be associated with interval scales, where numerical scores maintain equal differences in the underlying construct across groups. Thus, a 1-point difference in scores on a value inventory (e.g., for universalism values) can be meaningfully interpreted across groups, implying an increased level of universalism in one group compared to another. This level of equivalence allows comparisons of associations (correlations or regressions) or patterns of means, but not absolute scores. Mean comparisons in terms of the implied theoretical variable are only permissible when full score or scalar equivalence is established. Here, a ratio scale is implied that has an absolute zero point and preserves ratio equivalence between any two observations. At the level of full score equivalence, scores across groups reflect true construct levels, and therefore, means can be directly interpreted with confidence as expressing the underlying level of theoretical value orientation. This discussion of equivalence based on scale types does not imply or require statistical methods such as multi-group CFA or IRT models. Instead, it is concerned with what individual observations reveal about the theoretical concept measured by an item. This perspective emphasizes the precision of each measurement in a data set.

TABLE 1 | Overview of the value types, their definition and an example item in the European Social Survey (ESS).

Value type	Definition	Example item
Openness to change		
Self-direction	Independence of choosing, creating and exploring ideas and actions	It is important to her/him to make her/his own decisions about what she/he does. She/he likes to be free and not depend on others
Hedonism	Pleasure and gratification of desires	She/he seeks every chance she/he can to have fun. It is important to her/him to do things that give her/him pleasure
Stimulation	Seeking excitement, novelty and challenges in life	She/he looks for adventures and likes to take risks. She/he wants to have an exciting life
Conservatism		
Security	Emphasizing safety, harmony and stability of the self, relationships and society	It is important to her/him to live in secure surroundings. She/he avoids anything that might endanger her/his safety
Conformity	Restraining actions and impulses that may upset or harm others and violate social norms or expectations	It is important to her/him always to behave properly. She/he wants to avoid doing anything people would say is wrong
Tradition	Emphasizing respect, commitment and acceptance of traditional customs, ideas and culture	Tradition is important to her/him. She/he tries to follow the customs handed down by her/his religion or her/his family
Self-transcendence		
Benevolence	Concern with the wellbeing of people close to oneself	It is very important to her/him to help the people around her/him. She/he wants to care for their well-being
Universalism	Emphasis on tolerance, understanding and protection of the wellbeing of all people and nature	She/he thinks it is important that every person in the world should be treated equally. She/he believes everyone should have equal opportunities in life
Self-enhancement		
Power	Pursuit of social status and prestige, control and dominance over resources and people	It is important to her/him to be rich. She/he wants to have a lot of money and expensive things
Achievement	Demonstrating competence in-line with social standards and pursuit of personal success	It's important to her/him to show her/his abilities. She/he wants people to admire what she/he does

At the operational level, a second invariance paradigm associated with latent variable modelling has become dominant in psychology (Leitgöb et al. 2023). This perspective implies similar levels of comparability. The basic level typically considered in this statistical paradigm is configural invariance, which corresponds to structural equivalence according to Van de Vijver and Leung (2021). It tests whether the same structural model holds for each of the groups or measurement occasions. The concern is that the direction of the item loadings (how much each item measures the implied latent variable) is consistent, although the strength of the loadings may vary across groups. This is generally taken as the baseline model for all further comparisons. The next level up is metric invariance, in which the relative factor loading patterns are constrained to be identical across groups or measurement occasions. This means that items discriminate equally well between individuals with the same underlying trait across groups. It is therefore operationally similar to metric equivalence. Once metric invariance can be assumed, it is possible to constrain the item intercepts to be equal across groups. If this condition shows no deterioration in fit, this indicates scalar invariance. At this level, it is possible to compare the mean scores across groups, because it implies that any variation in the observed means is caused by variation in the latent variable.

Although numerous latent variable methods exist for testing invariance, MG-CFA is the most popular approach in the literature (Boer et al. 2018; Leitgöb et al. 2023; Van Herk and Goldman 2022). Recently, the concept of partial invariance has gained traction (Boer et al. 2018; Byrne et al. 1989; Leitgöb et al. 2023). Here, only a subset of parameters (such as some factor loadings or item intercepts) is required to be invariant, whereas others are allowed to vary. This approach allows for greater leniency in how many parameters need to be invariant, and by loosening the constraints on the model, the fit is improved.

Despite its popularity, it poses some conceptual challenges. In an MG-CFA, a key assumption is that the items are interchangeable—that is, they represent a random selection from a larger pool of items and are locally independent, that is, any variation in item scores is due to the latent variable. However, most measurement tools contain a limited number of items, making it unlikely that they represent a truly random sample from a large universe of possible indicators. In addition, the items are often not conditionally independent of each other when the latent variable is taken into account (Robitzsch and Lüdtke 2023). Given a limited item pool, removing even one or two items or loosening parameter constraints can shift the meaning of the construct itself (Singh 1995; Steenkamp and Baumgartner 1998; Van Herk and Goldman 2022). This issue becomes even more complex in multi-group comparisons, where different sets of items may vary across groups (Robitzsch and Lüdtke 2023). For example, if the items used to compare the United Kingdom and the United States differ from those used to compare the United States and Spain, any conclusions about pairwise differences are open to different interpretations. Therefore, Robitzsch and Lüdtke (2023) argued that interpreting a poorly fitting multi-group factor model—where the parameters are consistent but do not fit perfectly across groups—can sometimes be conceptually clearer than interpreting a model that fits well but requires different parameters for each group.

1.3 | Previous Work on Invariance of Values in the ESS

Several studies have primarily examined the cross-country invariance of specific rounds of the ESS. Bilsky et al. (2011) used the first three rounds and theory-driven multidimensional scaling with Procrustes rotation and identified two main dimensions in all 71 samples from 32 countries. The circular structure and the organization of the individual value types within the structure were generally supported, but they also reported a clearer separation of the main value types according to theoretical predictions in countries with higher levels of economic development, education and democratic participation.

In contrast, Davidov et al. (2008) analysed data from 20 countries collected during the first round with MG-CFA. They report acceptable levels of fit for configural and metric invariance of a modified value model, which allows for comparison of correlations of these modified values across the samples included in the first round of the ESS. The modifications nevertheless implied that the original theoretical structure was not empirically supported. Further investigations comparing data from rounds two and three suggested problems with some of the value types, which needed to be merged and resulted in a four to seven value type structure that could be identified in country samples instead of the originally proposed 10-value type structure (Davidov 2008, 2010). Across these three waves, Davidov suggested that temporal invariance could be achieved with the modified models, and therefore, value change could be reliably investigated.

The most comprehensive test to date was reported by Cieciuch et al. (2018), using data from 15 countries that participated in the first six rounds. They focused on the four higher order values (instead of the 10 value types) in an MG-CFA, using an approximate measurement procedure within a Bayesian framework (Inglehart 1997). In a Bayesian approximate measurement approach, a distribution of fit parameters is assumed, and the variance of the crucial parameters at each level of invariance is assessed. They found approximate metric invariance for the higher order values of openness to change and self-enhancement in most rounds, but no scalar invariance. For conservation and self-transcendence values, metric invariance was only observed in small subsets of countries. These analyses therefore imply that across the first six rounds of the ESS, only correlations and mean patterns of some value types can be compared across countries. Furthermore, value means cannot be compared across cultures within rounds.

Across these different analysis methods and rounds, it appears that at least two of the higher value groups can be reliably compared across countries up to the first six rounds of the ESS. There is limited information on the 10 value types across countries and no information on their measurement stability over the whole period of the ESS database. The data are widely used, but it is not clear whether the value theory is applicable in European contexts. Can we compare data over the full 11 rounds covering representative samples within Europe and bordering countries?

1.4 | Treating Invariance as Useful Information on Data Quality

We conducted our analyses with the intention of providing some basic information about the confidence with which value data in the ESS data set can be interpreted. We do not suggest that a lack of invariance using thresholds such as ours should preclude further investigation. First, we used a rather liberal threshold, which could be criticized as being too lenient. Second, we strongly believe that invariance parameters provide important information about the data that should be a greater focus of research to determine whether systematic factors are at play.

To make some initial forays in this direction, we decided to focus initially on two variables: time and response behaviour. Temporal changes at the societal level (e.g., economic cycles, political instability and shifts in the zeitgeist) can affect values and their conceptualization. The postmodernization thesis (Inglehart 1997) argues that increased economic development systematically shifts values towards those that emphasize the well-being of others and the environment. Such shifts in mean values have now been well demonstrated (Inglehart 2018; Welzel 2014). An underappreciated side effect of such effects may be how individuals interpret values, that is, such economic shifts may change the way values are perceived and cognitively organized (Fontaine et al. 2008; van de Vijver and Poortinga 2002). Effects consistent with this possibility have now been reported in a number of studies (for a review of this line of research, see Fischer 2017).

Second, the response behaviour itself may also change over time and affect model fit. A classic response behaviour is response style. Both shifts in mean responses across all values, similar to acquiescent response styles, and changes in overall standard deviations, implying greater differentiation of responses or narrower use of the response scale, as implied by extreme or cautious response styles, can be distinguished (Cheung and Rensvold 2000). Such behaviours do not necessarily imply bias but may be indicative of substantive processes. For example, shifts in response means may indicate an increased concern for maintaining harmony (Smith 2004). Similarly, variability in response standard deviations of values has been taken as an indicator of cultural tightness versus looseness, that is, individuals are more likely to restrict their behavioural options or freely express their desires and wishes (Gelfand et al. 2011; Ikizer et al. 2024; Uz 2018, 201). Increasing value dispersion may also imply increasing polarization in the political domain (Ollerenshaw 2023). We are agnostic about the possible causal processes leading to change. We are interested in whether changes in either the total value means or the total standard deviation are associated with changes in measurement fit.

Therefore, we use both time and response patterns across all scores as a first indication of possible systematic effects that may affect model fit and thus invariance parameters. We provide this as a first case study, using the information from our invariance tests to increase transparency and to highlight that invariance information can have substantive meaning (Fischer et al. 2021, 2025; Fischer and Rudnev 2024; Sterner et al. 2024), challenging recent calls to abandon invariance testing (Funder and Gardiner 2024; Welzel et al. 2023). We aim to demonstrate that invariance information is valuable and can open doors for further theoretical research.

1.5 | The Current Study

Our aim is to (a) provide a series of tests of the invariance of the 10 value types measured by the ESS-21 across countries for all currently available ESS rounds, (b) explore patterns of mean differences over time and (c) provide some hypotheses for further exploration of invariance patterns in the ESS. By bringing together data from all rounds in a single set of analyses, our aim is to inform interested users about caveats in the interpretation of value patterns and to provide directions for more targeted analyses. The overall results across ESS rounds and countries will allow interested readers to assess whether and at what level value results from individual rounds can be meaningfully interpreted across rounds and countries. For example, if we find metric invariance across time or across countries, correlations reported in the previous research based on ESS data can be interpreted with confidence. If we find a poor fit of the data when conducting metric invariance tests, this would imply that results from studies using data from different countries or rounds may not be directly comparable with each other (Leitgöb et al. 2023). Therefore, our results will provide important background information on the interpretability of published value research.

2 | Methods

We used data from the ESS database. It consists of strict probability samples of at least 1000 respondents from the general population in each country aged 18 and over. The first survey was conducted in 2002, and since then there have been eleven rounds, conducted every 2 years. Not all countries are represented in all rounds. Figure 2 shows the representation of countries over time. A total of 443,655 responses were available, and data for which the value inventory was fielded were available from 408,398.

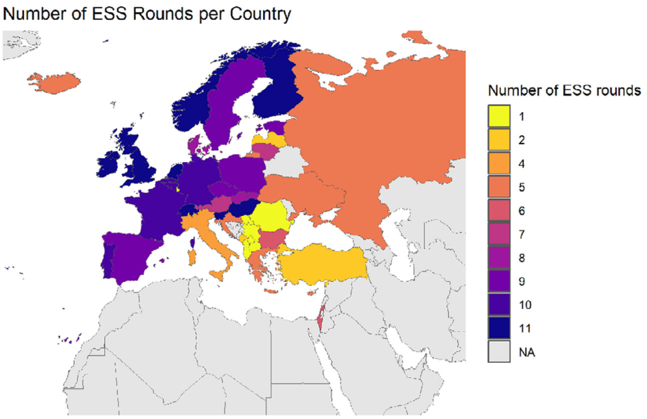

FIGURE 2 | Number of ESS rounds per country. ESS, European Social Survey.

2.1 | Instruments

A 21-item version of the Portrait Value Questionnaire (Schwartz et al. 2001, 2015) has been included as part of the ESS in each of the rounds. It consists of short verbal portraits of individuals, gender-matched to the respondent. Each statement describes a person's goals, aspirations or wishes, which are thought to encapsulate values relevant to one of the 10 value types. Respondents are asked to indicate how similar the description of the person is to them, on a scale from 'not at all like me' to 'very much like me'. The inventory includes two items per value type, except for universalism, which has three value items. From rounds 1 to 7, the value questionnaire was included in the supplement of the survey, which was administered after the sociodemographic questions and rotating modules. From round 8 onwards, it was included in the main section. The module has been applied both in paper-and-pencil and online versions. We do not separately test the invariance across application modes, and the previous work has suggested that this matters less (Davidov and Depner 2011). In round 11, the treatment of missing and non-response data was changed.

Data for value responses were downloaded in csv format on 11 October, 2024 (European Social Survey European Research Infrastructure [ESS ERIC], 2024). No power analysis was conducted. All measures, manipulations and data exclusion criteria are reported. We used all complete data (no missing information on any question for each individual), leading to the removal of 33,833 responses.

2.2 | Analytical Approach

We used confirmatory factor analysis because this has become the most common method for examining invariance (Boer et al. 2018; Leitgöb et al. 2023; Van Herk and Goldman 2022). We chose this approach because it is the most widely used, researchers will be familiar with this method, and the results can be easily understood in the current research environment.

We tested the original theoretical model with 10 latent variables representing each value type. There are two value items per latent value type, except for universalism, which has three value items. Each value item was allowed to load on their respective proposed theoretical latent variable. We did not specify cross-loadings, no uncorrelated bifactor was modelled, and we did not model the latent variable structure (all covariances between the latent variables were free). In other words, our tested model conforms to the original 1992 theoretical basic human value model proposed by Schwartz.

We used raw data for all our analyses. No weighting was applied.

All analyses were performed with lavaan 0.6–17 (Rosseel 2012) in R (R Core Team 2021). We used maximum likelihood estimation with robust (Huber–White) standard errors (MLR) and a scaled test statistic that is (asymptotically) equal to the Yuan–Bentler test statistic (Rosseel 2012). There has been some discussion about whether it might be more appropriate to use ordinal estimation methods. We initially explored solutions using weighted least squares with robust standard errors (WLSMV) and diagonally weighted least squares, which are appropriate for ordinal data. The initial results were qualitatively similar, but the estimation time was significantly longer. Because other studies (Davidov et al. 2018; Lee and Soutar 2010) have previously reported similar results using methods appropriate for ordinal and interval scale types, we proceeded with MLR estimation. For identification purposes, we always set the variance of the latent variable to 1. We did explore whether results were different when setting a factor loading to 1 for identification purposes. The results for the overall multi-group invariance analysis for countries as well as a simple simulation resulted in practically identical estimates. For practical purposes, we do not believe that changing the setting of model identification will result in qualitatively different results.

2.2.1 | Overall Model Fit in Total Sample

We first computed an overall model pooling all data from all rounds and countries. This is used as a baseline model to examine the overall fit of the data to the theoretical model. We used a number of indicators to examine model fit. Strict fit indicators such as χ^2 are known to depend on sample size, which is an issue for our sample. Therefore, we report the χ^2 but do not elaborate on the absolute value of this parameter. We consider MLR-based robust versions of the Comparative Fit Index (CFI) (Bentler 1990), robust versions of the Tucker–Lewis Index (TLI) (Bentler 1990; Tucker and Lewis 1973), robust versions of the Root Mean Square of Approximation (RMSEA) (Browne and Cudeck 1992), and the Standardized Root Mean Residual (SRMR) (Bollen 1989). The CFI assesses the relative improvement in model fit compared to a baseline (usually the 'null' or 'independence' model, which assumes no relationships between variables). The calculation is based on the χ^2 statistic, adjusted for model complexity, and is scaled between 0 and 1, with higher values indicating better fit. A CFI value of 0.90 or higher is typically considered reasonable, and 0.95 or higher has been suggested as adequate fit (Hu and Bentler 1999). The TLI (also known as the non-normed fit index, NNFI (Bentler 1990) compares the fit of the target model to the null model, adjusting for model complexity to prevent overfitting. It has been shown to be more robust to sample size (Marsh et al. 1988). Values closer to 1 (usually ≥0.90) have historically been considered to indicate adequate fit (Bentler and Bonett 1980). The RMSEA measures the approximate error of the model per degree of freedom, estimating how well the model would fit the population covariance matrix if it were available. As with the other indices, it is calculated from the χ^2 statistic and adjusted for sample size and complexity. As a general guideline, RMSEA values below 0.05 or 0.06 indicate an appropriate fit, values up to 0.08 are acceptable, and values above 0.10 indicate a poor fit (Browne and Cudeck 1992). The root mean residual evaluates the average difference between observed and predicted correlations, quantifying the residuals of the model. We used the standardized version, with values below 0.08 generally considered a good fit. Because these fit indices take into account different information and are calculated in different ways, it is recommended to consider their overall pattern.

Here, we adopted a threshold in which we considered a model to fit if three of the four fit indices were within commonly accepted limits. We considered values of CFI and TLI above 0.90 and values of RMSEA and SRMR below 0.08 to indicate acceptable fit. In considering information from different fit indices that are sensitive to different types of misspecification and considering a combination of indices, we aim to capture which models perform comparatively worse. We acknowledge that these criteria may be on the lenient side (Hu and Bentler 1999; Sivo et al. 2006; Ximénez et al. 2022), and there may be uncertainty about judging model fit absent model-specific simulations. However, given the complexity of the model with 10 latent variables and the large sample sizes, we believe our strategy is defensible given the state of the literature and practically meaningful for researchers interested in interpreting results from the ESS. At the minimum, the reporting of the fit indices can be used as a reference for other researchers on empirical model fit of the data to the theoretical model.

2.2.2 | Overall Fit Using the Average Within-Country

The previous model uses the total data set, which includes variability within and between countries in the overall data set as well as disadvantages for countries that participated less often in the ESS (e.g., some countries participated in only round, whereas several countries participated in all 11 rounds). We therefore calculated the average within-country covariance matrix using the *psych* package (Revelle 2024), which gives the same weight to each country and averages the covariance matrix among all value items (Muthen 1994). We then used this matrix as an input to a confirmatory factor analysis with ML estimation and fixed the variances of the latent variables to 1. This can be interpreted as the fit of the theoretical structure to the average within-country structure. We judged fit with the same standards: Three of the four fit indices showed acceptable fit (e.g., values of CFI and TLI above 0.90 and values of RMSEA and SRMR below 0.08).

2.2.3 | Overall Fit Using the Average Sample Across Rounds and Countries

We also computed an average sample covariance matrix, which computes the average covariance across all samples in all rounds. We computed this averaged matrix because some of the rounds varied in sample sizes, ranging from an effective sample size of 478 to 2906. Therefore, we computed the average covariance matrix among all value items that gave equal weight to each sample across all countries and rounds. This covariance matrix was submitted to a CFA with ML estimation, and the solution was identified via standardizing the latent variables. This solution can be interpreted as the fit of the theoretical model to the average sample across all rounds and countries. We utilized the same fit criteria (three of the four indicators having an acceptable fit according to our criteria of CFI and TLI above 0.90 and RMSEA and SRMR below 0.08).

2.2.4 | Cross-Cultural Invariance (Ignoring ESS Round Information)

We conducted an overall cross-country invariance test, pooling all data across rounds for each country. This test concerns the cross-cultural invariance overall, ignoring temporal information. We first estimated a configural model as a baseline using MLR estimation and fixed the variance of the latent variable to 1 for identification purposes. In lavaan, when setting the variance of the latent variable for identification purposes and imposing equality constraints on factor loadings, only the latent variable of the first group will be fixed to 1.0, and latent variances of the other groups are freely estimated. We estimated the fit using the same three index presentation strategy (three fit indices meet our threshold: CFI > 0.90, TLI > 0.90, SRMR < 0.08, RMSEA < 0.08).

For those samples that met this criterion, we then constrained factor loadings to be equal to assess metric invariance. We evaluated fit deterioration for CFI and RMSEA and considered a model with a difference of less than 0.01 to pass our fit threshold. For those samples that passed the metric invariance test, we constrained the intercepts to be identical. We used the same two fit indicators (difference in fit <0.01 for CFI and RMSEA).

2.2.5 | Model Fit in Each Sample of Each Round

The previous analyses pooled data across samples. We tested model fit separately in each individual sample, that is, for each round in each country. We ran a CFA model with MLR estimation and assessed the fit using the same three-indicator strategy. This provides the equivalent of a configural model. This information is useful for understanding whether the data from a particular round and sample can be interpreted easily. For example, misfit for a particular sample in a round would mean that any correlations of values with other variables based on that sample and round should be interpreted with caution. This information is essential reference information for all published studies using ESS score data from the last 20 years.

We then used a different strategy to approximate metric invariance across samples, given the resource demands of running an invariance test with over 200 samples, which may lead to convergence issues and model instability. Pooled data approaches have been proposed to circumvent these problems and provide more precise estimation of parameters (Saris and Satorra 2018). In-line with this general approach, we computed three different alternatives.

First, we used the unstandardized loading matrix from the overall model using the total data set across all samples as a reference and tested the fit in each individual sample when specifying loadings theoretically in the model based on this loading matrix (see Table 2). Thus, we did not constrain the loadings to be equal but rather tested whether the unstandardized loadings were equivalent to the point estimates from the overall solution across countries (see Appendix 2 in Saris and Satorra 2018 for a general introduction and Monte Carlo evaluation of robustness).

Second, we used the unstandardized loading matrix from the pooled within-country model as a reference and tested the fit of each individual sample to point estimates of this reference solution. Third, we used the unstandardized loading matrix from the pooled within-sample and country matrix as a reference and tested the fit of each individual sample to the point estimates of this reference solution.

Hence, each model used is based on a pooled data approach (Saris and Satorra 2018) but utilizes a slightly different reference solution: The first uses the total sample which may advantage larger samples and those countries that participated in all rounds; the second gives equal weight to all countries but ignores temporal information; and the third gives equal weight to each of the samples (for the specific point estimates, please see Table 2).

To judge model fit, we again utilized the same three-indicator strategy. This information provides important information for interpreting the published data with the ESS scores. If the metric invariance test is satisfied, correlations or mean patterns can be interpreted with confidence. The different reference solutions may also point to potential sources of biases that may need consideration.

2.2.6 | Temporal Invariance per Country

The unique advantage of the ESS is that data have been collected since 2002. The data make it possible to monitor changes in values over a significant period of time. However, this requires evidence that the value measures have remained stable and that any changes can be attributed to substantive changes. This requires temporal invariance.

We first tested a configural model in each country across all available rounds. In this baseline model, items were allowed to freely lead on their respective factors in each round. We used the three-indicator strategy to assess model fit (three fit indices meet our threshold: CFI > 0.90, TLI > 0.90, SRMR < 0.08, RMSEA < 0.08). For each temporal analysis within each country that met configural invariance, we then constrained factor loadings to be identical across time. As we identified the model via setting the variance of the latent variable, in a metric invariance test in lavaan in which the factor loadings are constrained to be equal, only the variance of the latent variable for the first observation in each country is fixed. Model fit was assessed by changes in CFI and RMSEA, with changes of less than 0.01 indicating adequate fit. For those countries where metric invariance over time was met, we constrained the intercepts to be identical. Again, we used the threshold change of less than 0.01 for CFI and RMSEA as an indicator of adequate fit.

TABLE 2 | Fully standardized factor loadings for the overall MLR model, ML model using the pooled within-country and ML model with the pooled within-sample and country covariance matrix.

Latent factor	Indicator	Total sample MLR loading	95% CI	Within-country	95% CI	Within-sample-country	95% CI
Achievement	ac1	0.714	0.712–0.717	0.696	0.694–0.698	0.693	0.691–0.695
Achievement	ac2	0.771	0.768–0.773	0.763	0.761–0.765	0.763	0.761–0.765
Benevolence	be1	0.697	0.694–0.700	0.678	0.675–0.680	0.673	0.671–0.675
Benevolence	be2	0.666	0.663–0.669	0.657	0.655–0.659	0.653	0.651–0.655
Conformity	co1	0.538	0.535–0.541	0.540	0.537–0.543	0.538	0.535–0.541
Conformity	co2	0.711	0.708–0.715	0.699	0.696–0.702	0.696	0.693–0.698
Hedonism	he1	0.700	0.698–0.703	0.724	0.721–0.726	0.722	0.720–0.724
Hedonism	he2	0.733	0.730–0.735	0.722	0.720–0.725	0.723	0.721–0.725
Power	po1	0.551	0.547–0.555	0.521	0.518–0.525	0.516	0.513–0.520
Power	po2	0.536	0.533–0.540	0.498	0.495–0.501	0.496	0.492–0.499
Self-direction	sd1	0.573	0.569–0.576	0.568	0.565–0.571	0.566	0.563–0.569
Self-direction	sd2	0.549	0.545–0.552	0.534	0.531–0.537	0.531	0.528–0.534
Security	se1	0.680	0.677–0.683	0.648	0.645–0.650	0.646	0.643–0.649
Security	se2	0.663	0.660–0.666	0.638	0.635–0.641	0.635	0.632–0.637
Stimulation	st1	0.669	0.667–0.672	0.665	0.663–0.667	0.665	0.663–0.668
Stimulation	st2	0.703	0.700–0.705	0.712	0.710–0.714	0.711	0.709–0.713
Tradition	tr1	0.484	0.480–0.487	0.493	0.489–0.496	0.490	0.487–0.494
Tradition	tr2	0.477	0.474–0.481	0.460	0.456–0.463	0.461	0.458–0.464
Universalism	un1	0.541	0.538–0.544	0.531	0.529–0.534	0.527	0.525–0.530
Universalism	un2	0.599	0.596–0.602	0.580	0.578–0.583	0.576	0.573–0.578
Universalism	un3	0.588	0.585–0.590	0.590	0.587–0.592	0.587	0.585–0.590

Note: All loadings are $p < 0.001$.

2.2.7 | Exploration of Misfit Patterns

We regressed the fit indicators on two predictor variables: time and response patterns. For time, we used the year of data collection for each round of the ESS and designated 2002 as year 0 for our analysis. Therefore, the regression equation can be interpreted in terms of fit in 2002. We also included a quadratic trend over time, centred on 2008. We used this as a centring point because the global financial crisis that began in 2008 may have affected the response behaviour and thus the model fit (for an earlier study demonstrating some indirect evidence of unemployment on value changes, see Vecchione et al. 2016). Importantly, there were sufficient data points prior to and after the global crisis, allowing us to study relative recovery of structure in relation to a temporal impact point. Therefore, the regression parameters can be interpreted in relation to 2008, indicating whether the fit has worsened or improved compared to 2008.[1]

For response behaviour, we calculated the mean and standard deviation for each individual respondent across all value scores and then averaged the mean and standard deviation of the total scores of the respondents within the sample (Fischer 2004). Therefore, we have an average mean and average standard deviation across all scores for each sample in each round and country. As an important observation, in this analysis, we test whether response behaviour as indicated by the overall mean across all value items is associated with model fit. This response mean is typically used for adjustment (ipsatization) purposes with the individual items (Fischer 2004; Rudnev 2021). Here, we used the same mean and rationale to see if the response behaviour is influencing model fit.

2.3 | Transparency and Open Data

The data are available at https://www.europeansocialsurvey.org/data-portal. The code and further results are available at https://osf.io/4tma8/?view_only=0d496c12b59344cb823838efe612f773. The analyses were not preregistered.

3 | Results

3.1 | Value Model Fit in Total Sample Ignoring Country and ESS Round Information

We first ran a CFA model with MLR estimation on the full data set. The fit of the model was relatively good: χ^2 (N = 443,655, df = 144) = 193,505.82, robust CFI = 0.920, robust TLI = 0.884, robust RMSEA = 0.055 (95% CI 0.055–0.055), SRMR = 0.047. The chi square is significant, and TLI is below 0.9. Table 2 shows the completely standardized factor loadings. All loadings were significant at $p < 0.001$. The lowest loadings overall were observed for tradition values (both standardized loadings below 0.5). The R^2 values varied between 0.594 (ac2) to a low of 0.228 (tr2).

For the averaged within-country solution, the fit was comparable to the overall data: χ^2 (df = 144) = 180,850.37, CFI = 0.922, TLI = 0.886, RMSEA = 0.053 and SRMR = 0.046. The lowest loadings were again observed for tradition value items (both items with standardized loadings below 0.5). The R^2 values varied between a low of 0.211 for the second tradition item and a high of 0.582 for the second achievement item.

For the averaged sample (averaged across rounds and countries), the fit was highly similar: χ^2 (df = 144) = 178,155.10, CFI = 0.922, TLI = 0.886, RMSEA = 0.053 and SRMR = 0.046. The lowest loading was again observed for tradition items (standardized loadings below 0.5), and the R^2 values again varied between a low of 0.212 for the second tradition item and a high of 0.583 for the second achievement item. The loadings for the different solutions are shown in Table 2.

Previous research had encountered identification problems when estimating the full 10-value structure. All models converged. Examining the covariances in the fully standardized models, the average correlation was 0.44 for each of the models. The pattern of the correlations was practically identical; therefore, we focus on the overall data here. Two negative correlations were observed for conformity with stimulation (r = −0.03) and tradition and stimulation (r = −0.12). Focusing on correlations above 0.80, we observed a few correlations of concerns (conformity × tradition r = 1.01; universalism × benevolence r = 0.93; power × achievement r = 0.98; tradition × security r = 0.87; hedonism × stimulation r = 0.83). These patterns confirm earlier reports that some of the value types are strongly associated (Davidov 2008; Davidov et al. 2008, 2010). We decided not to change the theoretical model in order to provide reference parameters for other researchers that have used those value types separately.

We also examined modification indices to identify possible sources of misfit. Across the three model instantiations, the largest modification indices (MI > 10,000) were observed for the two power and two stimulation items. Model fit could be improved if including cross-loadings from the power items were included to conformity, tradition, security or benevolence latent variables or for stimulation items to universalism, self-direction, benevolence, tradition, security or conformity. These items may need more careful examination in revisions.

3.2 | Analyses Comparing Model Fit Within Each Country (Ignoring ESS Round Information)

We ran a CFA with MLR estimation for each country across all ESS rounds. We did not constrain any model parameters, and each sample was analysed separately (single-country model). Using our criterion of at least 3 fit indicators passing our thresholds, 24 out of 39 countries passed (see Table 3). Focusing on the individual fit indicators, 24 samples had a robust CFI above 0.9, no sample had a robust TLI above 0.9, and all samples had a robust RSMEA below 0.08 and an SRMR below 0.08, respectively. All chi square values were significant. The average robust CFI was 0.90 (min = 0.88, max = 0.93), the average robust TLI was 0.86 (min = 0.82, max = 0.89), the average robust RMSEA was 0.06 (min = 0.05, max = 0.08), and the average SRMR was 0.06 (min = 0.04, max = 0.08).

3.3 | Cross-Country Multi-Group Invariance Analysis

We proceeded with a classic multi-group invariance analysis, starting with a configural model in which the same model was tested as a baseline in a multi-group analysis. Across all countries, we found reasonable fit: χ^2 (df = 5616) = 247,399.42, robust CFI = 0.904, robust TLI = 0.859, robust RMSEA = 0.061, SRMR = 0.051. Configural invariance was met according to our thresholds, even including countries that did not pass our fit criterion when tested individually (see Table 3).

We next constrained the factor loadings to be identical across all countries. The fit was still acceptable: χ^2 (df = 6034) = 258,870.428, robust CFI = 0.899, robust TLI = 0.863, robust RMSEA = 0.061 and SRMR = 0.053. The χ^2 difference was highly significant though: difference χ^2 (df = 418) = 8869, $p < 0.0001$. However, given the sample sizes, this significance is not surprising. The differences in CFI and RMSEA were below the 0.01 threshold. Although the CFI was within rounding margin of the threshold, overall the performance suggests that metric invariance can be assumed across countries, when pooling all data across available rounds.

When constraining item intercepts to be identical, the model fit deteriorated substantively: χ^2 (df = 6452) = 436,285.578, robust CFI = 0.828, robust TLI = 0.782, robust RMSEA = 0.076 and SRMR = 0.062. The difference between the models was also substantively worse: $\Delta\chi^2$ (df = 418) = 436,286, $p < 0.0001$; delta robust CFI = 0.071 and delta robust RMSEA = −0.015. We did not compute the delta TLI because the fit for the previous model was already below our threshold.

TABLE 3 | Model fit for each country, using MLR.

Country	CFI	TLI	RMSEA	SRMR	χ^2	Fit threshold passed
Albania	0.902	0.857	0.052	0.048	571.537	1
Austria	0.905	0.862	0.062	0.058	8431.235	1
Belgium	0.895	0.846	0.054	0.045	7296.758	0
Bulgaria	0.923	0.887	0.067	0.054	7313.698	1
Switzerland	0.891	0.842	0.055	0.045	7705.301	0
Cyprus	0.893	0.843	0.068	0.058	3414.918	0
Czech Republic	0.916	0.877	0.065	0.057	10,684.015	1
Germany	0.901	0.855	0.055	0.048	11,644.325	1
Denmark	0.913	0.873	0.050	0.041	4359.627	1
Estonia	0.896	0.848	0.058	0.054	8040.978	0
Spain	0.922	0.886	0.053	0.047	6684.357	1
Finland	0.920	0.884	0.054	0.043	7759.714	1
France	0.881	0.827	0.060	0.051	9413.426	0
United Kingdom	0.899	0.853	0.058	0.048	10,406.606	0
Greece	0.917	0.879	0.061	0.055	6659.412	1
Croatia	0.880	0.824	0.073	0.066	5821.426	0
Hungary	0.875	0.818	0.074	0.063	13,728.139	0
Ireland	0.908	0.865	0.063	0.056	12,105.391	1
Israel	0.925	0.891	0.054	0.039	5526.774	1
Iceland	0.915	0.877	0.050	0.043	1456.447	1
Italy	0.880	0.824	0.078	0.076	7150.264	0
Lithuania	0.892	0.842	0.077	0.066	10,184.319	0
Luxembourg	0.885	0.832	0.060	0.052	947.775	0
Latvia	0.876	0.819	0.072	0.069	2176.658	0
Montenegro	0.916	0.878	0.070	0.083	1854.486	0
North Macedonia	0.902	0.857	0.066	0.069	1005.364	1
Netherlands	0.908	0.865	0.054	0.043	8213.568	1
Norway	0.916	0.878	0.052	0.042	6482.696	1
Poland	0.909	0.867	0.061	0.058	7596.208	1
Portugal	0.908	0.866	0.066	0.058	10,781.450	1
Romania	0.886	0.834	0.081	0.077	2078.025	0
Serbia	0.901	0.855	0.061	0.061	1116.154	1
Russia	0.903	0.859	0.069	0.065	7468.011	1
Sweden	0.907	0.865	0.056	0.048	6582.152	1
Slovenia	0.880	0.824	0.065	0.057	8576.427	0
Slovakia	0.918	0.880	0.063	0.055	6849.710	1
Turkey	0.909	0.867	0.069	0.060	2828.467	1
Ukraine	0.901	0.855	0.071	0.072	5430.584	1
Kosovo	0.880	0.826	0.075	0.060	1053.022	0

Note: CFI, TLI and RMSEA are Huber–White robust estimates; 1 = passed; 0 = failed.

3.4 | Individual Model Fit for Each ESS Round and Country Sample

These previous analyses averaged across all ESS rounds, which may either introduce or obscure variability over time. Therefore, we ran a separate CFA model for each ESS round for each country. Because we only tested whether the theoretical model shows adequate fit in each round in each country, this can be seen as a test of configural invariance for each ESS round sample. Of these 261 analyses, 251 models passed our threshold of at least 3 appropriate fit criteria (for an overview of the results, see Table 4, for detailed results, see https://osf.io/4tma8/).

A different way to examine the data is to count the number of rounds in each country that passed configural invariance. On average, 94.1% of the rounds passed configural invariance in each country. The lowest levels were found in Croatia (4 out of 5 passed, 80%), Hungary (10 out of 11 passed, 90.9%), Italy (2 out of 4 passed, 50%), Lithuania (6 out of 7 passed, 85.7%), Latvia (1 out of 2 passed, 50%), Montenegro (1 out of 2 passed, 50%), Slovenia (9 out of 11 passed, 81.8%) and Ukraine (4 out of 5 passed, 80%). This implies that there are measurement occasion-specific effects for some of these rounds. In other words, one round (Hungaria, Lithuania, Latvia, Macedonia and Ukraine) or two rounds (Slovenia and Italy) at the maximum did not pass this threshold. Four of these samples were from round 9, and two were from round 6. Data from other rounds from each of the countries passed our threshold. Therefore, there was no obvious pattern to these results. Table 4 shows the overall result, listing how many rounds passed the threshold for each country. Table 8 shows the specific rounds that showed problems at this stage.

At this stage, we can only conclude that there was better fit to the overall model for the specific rounds, but this is not sufficient for making comparisons between any of them, either across country samples within each round or across rounds within each country. Therefore, we used a pooled data approach via the three different reference solutions we described at the outset of our analysis and tested the fit in each round and country sample when forcing the loadings to the respective loading matrix (Saris and Satorra 2018). This is conceptually similar to testing metric invariance (except that we did not constrain the loadings to be equal across samples but rather tested the relative fit of each round when specifying the loading pattern in the theoretical model).

Using this approach, only 22 samples out of 39 passed the threshold when using the overall data as reference, and 33 samples passed when either using the pooled within-country or pooled within-sample and country matrix as reference. Among the samples with more than two rounds, the following countries showed metric invariance in at least one round: Austria, Czech Republic, Germany, Spain, Finland, Britain, Ireland, The Netherlands, Norway, Sweden and Slovakia. Examining the differential solutions across the reference matrices, Czech samples fit better with the overall data as reference, whereas Spain, Ireland, The Netherlands Norway, Sweden and Slovakia performed better when using either of the within-averaged solutions. Concerning ESS rounds, there was no clear pattern favouring a specific round as more likely to show better fit, nor was the pattern similar across the different reference matrices. See https://osf.io/4tma8/?view_only=0d496c12b59344cb823838efe612f773 for full results by round and country.

3.5 | Temporal Invariance Within Each Country

Focusing more specifically on temporal effects within each country separately, we conducted a temporal invariance test within each country. We tested three increasingly restrictive models. First, we ran a configural model as a baseline. For this model, we judged appropriate fit with the same three-indicator threshold that we have described above for the individual country models. Then, we constrained factor loadings to be equal across all rounds. We considered the fit adequate if the deterioration in fit for the robust versions of the CFI, TLI and RMSEA were not exceeding a 0.01 threshold for two of the three indicators. Finally, we constrained the intercepts to be identical across time. We again considered a deterioration in two of the three indicators as decision criterion. We did not include Albania, Macedonia, Romania and Kosovo in this temporal analysis because these countries were included only once in the overall ESS data set.

First, 17 of the 35 countries passed the threshold for configural temporal invariance: Bulgaria, Czech Republic, Denmark, Spain, Finland, Israel, Ireland, Montenegro, The Netherlands, Norway, Poland, Serbia, Russia, Sweden, Slovakia and Turkey. Not passing our configural threshold were Austria (9 rounds), Luxembourg (2 rounds), Belgium, France and Portugal (10 rounds), Latvia (3 rounds), Greece, Croatia and Italy (5 rounds each), Cyprus (6 rounds), Lithuania (8 rounds), Estonia (9 rounds) and Switzerland, Germany, Great Britain, Hungary and Slovenia (11 rounds).

Among those countries that passed the configural threshold, all countries passed the metric invariance threshold.

Restricting the intercepts to be equal, Turkey, Iceland, Bulgaria, Denmark, Czech Republic, Spain, Finland, The Netherlands and Norway did not pass our invariance threshold. Therefore, these countries showed metric invariance, allowing a comparison of correlations and mean patterns, but absolute mean differences are open to alternative interpretations. In contrast, Montenegro, Serbia (2 rounds each), Russia (5 rounds), Israel (7 rounds), Slovakia (8 rounds), Poland, Sweden (both 10 rounds) and Ireland (11 rounds) passed the scalar invariance threshold, which means that mean differences over time could be safely interpreted. See Table 5 for specific results and Figure 3 for a geographical distribution of the temporal invariance statistics.

3.6 | Exploration of Invariance Patterns

As noted in the beginning, we see invariance analyses as an integral part of any scientific inquiry in the social and behavioural sciences. Information on invariance parameters can be further queried and may point towards important new research directions. To demonstrate some possible avenues for further research, we conducted some exploratory analyses of the configural and implied metric invariance fit indices (using the three different reference matrices) for each sample per country and ESS round.

TABLE 4 | Summary results showing how many samples passed configural and metric invariance (using the total sample, the averaged sample and country and averaged country matrix as reference).

Country	Rounds available	Configural invariance	Metric invariance (total sample)	Metric invariance (sample and country)	Metric invariance (country)
Albania	1	1	0	0	0
Austria	7	7	2	2	2
Belgium	10	10	0	0	0
Bulgaria	6	6	0	0	0
Switzerland	11	11	0	0	0
Cyprus	5	5	0	0	0
Czech Republic	9	9	4	2	2
Germany	10	10	1	1	1
Denmark	8	8	0	0	0
Estonia	9	9	0	0	0
Spain	9	9	1	4	4
Finland	11	11	6	7	7
France	10	10	0	0	0
United Kingdom	11	11	1	1	1
Greece	5	5	0	0	0
Croatia	5	4	0	0	0
Hungary	11	10	0	0	0
Ireland	11	11	0	0	0
Israel	6	6	2	3	3
Iceland	5	5	0	0	0
Italy	4	2	0	0	0
Lithuania	7	6	0	0	0
Luxembourg	1	1	0	0	0
Latvia	2	1	0	0	0
Montenegro	2	1	0	0	0
North Macedonia	1	1	0	0	0
Netherlands	11	11	0	2	2
Norway	11	11	3	7	7
Poland	9	9	0	0	0
Portugal	10	10	0	0	0
Romania	1	1	0	0	0
Serbia	1	1	0	0	0
Russia	5	5	0	0	0
Sweden	9	9	0	1	1
Slovenia	11	9	0	0	0
Slovakia	8	8	2	3	3
Turkey	2	2	0	0	0
Ukraine	5	4	0	0	0
Kosovo	1	1	0	0	0

TABLE 5 | Temporal measurement invariance (ESS rounds 1–11).

Country	No.	CFI	TLI	RMSEA	SRMR	ΔCFI metric	ΔRMSEA metric	ΔCFI scalar	ΔRMSEA scalar
Albania	1	0.902	0.857	0.052	0.046				
Austria	8	0.899	0.853	0.064	0.056				
Belgium	10	0.892	0.843	0.055	0.047				
Bulgaria	6	0.907	0.865	0.073	0.059	0.002	0.002	0.025	−0.007
Switzerland	11	0.885	0.832	0.057	0.047				
Cyprus	6	0.874	0.817	0.075	0.064				
Czech Republic	9	0.908	0.866	0.068	0.057	0.001	0.002	0.010	−0.001
Germany	11	0.896	0.849	0.056	0.048				
Denmark	8	0.912	0.872	0.050	0.043	0.000	0.002	0.011	−0.001
Estonia	9	0.890	0.840	0.060	0.055				
Spain	10	0.918	0.880	0.055	0.048	0.002	0.001	0.016	−0.003
Finland	11	0.918	0.880	0.055	0.045	0.001	0.001	0.016	−0.003
France	10	0.880	0.825	0.061	0.052				
United Kingdom	11	0.898	0.851	0.059	0.048				
Greece	5	0.899	0.853	0.067	0.056				
Croatia	5	0.876	0.819	0.075	0.065				
Hungary	11	0.873	0.815	0.075	0.063				
Ireland	11	0.900	0.854	0.066	0.057	0.002	0.002	0.009	−0.001
Israel	7	0.917	0.878	0.058	0.041	0.002	0.001	0.008	−0.001
Iceland	5	0.914	0.874	0.051	0.046	0.000	0.001	0.018	−0.003
Italy	5	0.874	0.817	0.079	0.074				
Lithuania	7	0.883	0.830	0.081	0.067				
Luxembourg	2	0.885	0.832	0.060	0.050				
Latvia	3	0.869	0.809	0.073	0.068				
Montenegro	2	0.911	0.871	0.073	0.073	0.003	0.000	0.006	−0.001
North Macedonia	1	0.902	0.857	0.066	0.066				
Netherlands	11	0.905	0.861	0.055	0.045	0.000	0.002	0.014	−0.002
Norway	11	0.914	0.875	0.053	0.044	0.001	0.001	0.017	−0.003
Poland	10	0.905	0.861	0.062	0.058	0.001	0.002	0.007	0.000
Portugal	10	0.900	0.854	0.071	0.060				
Romania	1	0.886	0.834	0.081	0.073				
Serbia	2	0.901	0.855	0.061	0.058	0.000	0.000	0.000	0.000
Russia	5	0.900	0.855	0.070	0.064	0.001	0.002	0.004	0.000
Sweden	10	0.904	0.860	0.057	0.048	0.001	0.001	0.008	−0.001
Slovenia	11	0.869	0.808	0.067	0.057				
Slovakia	8	0.912	0.871	0.066	0.056	0.002	0.001	0.007	0.000
Turkey	2	0.906	0.863	0.072	0.058	0.001	0.001	0.016	−0.004
Ukraine	5	0.895	0.847	0.073	0.071				
Kosovo	1	0.880	0.826	0.075	0.057				

Abbreviations: CFI, Comparative Fit Index; RMSEA, Root Mean Square of Approximation; SRMR, Standardized Root Mean Residual; TLI, Tucker–Lewis Index.

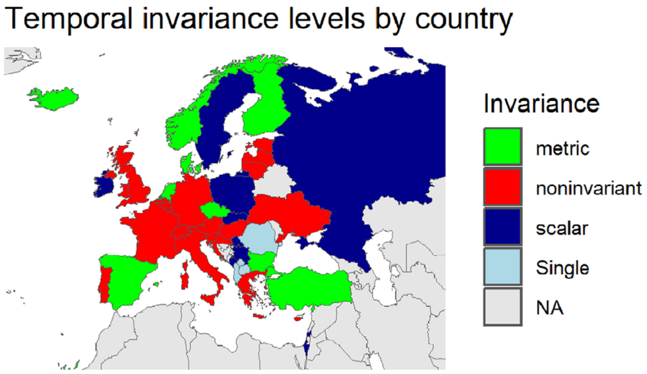

FIGURE 3 | Overview of the geographical distribution of temporal invariance.

As a first step, we computed composite scores based on the intercorrelation of the fit indices. We computed an index including all the χ^2 estimates ($\alpha = 0.98$), a composite of all the RMSEA and SRMR indicators ($\alpha = 1$), approximate fit index averaging the robust CFI and robust TLI from fitting the base model in each sample in each round and country ($\alpha = 1$), and a composite including all the CFI and TLI indicators from the metric model with the constrained factor loadings ($\alpha = 1$).

We then ran two multilevel regression models with country as a random factor. First, we used the value mean, value standard deviation and a linear time variable (year) as predictors for each of the fit indicator composites. We then ran a separate model in which we also included the curvilinear time effect centred on 2008.

Focusing on the mean value rating, it showed one significant effect on the CFI and TLI fit for the loading constrained models, suggesting that higher mean value ratings were associated with lower approximate fit when constraining the loadings to fit reference matrices. The variability of the mean value ratings had a significant effect on fit in all models. The effect suggested that higher variability was associated with lower χ^2 values and lower RMSEA/SRMR values (suggesting better fit for more varied ratings), but also a reduced CFI and TLI for both configural and loading constrained models, suggesting worse fit with increased value rating divergence.

Focusing on time effects, five of the models showed some significant effect for the linear time effect, in balance suggesting a deterioration of fit over time (CFI/TLI composite for configural and loading constrained models and overall χ^2 fit). Three of the curvilinear time predictors were significant, suggesting that fit prior to and post-2008 was better when examining the composites for χ^2, RMSEA/SRMR and CFI/TLI for configural model (Table 6).

4 | Discussion

We report on the measurement invariance of the Portrait Values Questionnaire included in 11 rounds of the ESS with 261 samples and responses from 443,655 participants from 2002 to 2024. Overall, the theoretical value model, which separates 10 value types, fits the full data set well. When testing for configural and metric invariance across country samples, we found acceptable levels of fit, suggesting that covariances and mean patterns can be compared across countries. However, this masks the fact that a number of countries did not show acceptable fit to the theoretical model when tested individually. In addition, samples from particular rounds may not show adequate fit, even though the theoretical model fits well when all rounds for that country are examined together. When testing invariance over time, only a subset of countries showed sufficient levels of invariance to allow comparison of correlations and mean patterns over the 21-year interval of the ESS. Looking at an exploratory analysis of systematic effects, we found some evidence that the fit of the model has deteriorated over time. We will discuss some of the implications next. Some key findings and answers to the main research questions are outlined in Table 7.

4.1 | Theory

Overall, the theoretical model fits the data well when tested (a) on the overall data, (b) using the average country sample, (c) the average sample across rounds and countries, (d) when testing for metric invariance across countries and (e) for configural invariance for most individual samples. This suggests that the 10 value types can be distinguished in these samples.

The caveat to this claim is that some of the latent variable intercorrelations were quite high. These patterns had been observed in the first few rounds of the ESS already (Davidov 2008; Davidov et al. 2008, 2010). Yet, other work with a more extensive value inventory has suggested that some value types should be split due to motivationally distinct content being captured (Lilleoja and Saris 2015). Taking into account observed modification indices, it suggests that adjustments either to the theoretical model or the operationalization might be desirable. Conceptually, it is interesting to reflect on when an observed correlation within a circumplex theoretical model is too high to warrant merging adjacent value types. For example, conformity and tradition values were practically merged, and their empirical distinctiveness is somewhat questionable based on the current data. Within the theoretical model, these two values share the same angular position, but tradition is positioned outside the conformity value type (Schwartz 1992). Similarly, universalism and benevolence were correlated above $r > 0.90$ in our overall models, suggesting that they are empirically highly similar. To what extent does it make sense to separate the theoretical motivation to care for close versus distant others? Conceptually, it makes sense to separate these values, but empirically these correlations suggest that individuals do not discriminate between these values in the same way. Similarly, some values measure highly diverse content, and it might be advisable to separate value types into more fine-grained components. For example, universalism in particular captures very diverse content, and it might be advisable to differentiate subcomponents of universalism (Lilleoja and Saris 2015; Schwartz et al. 2012).

With these caveats in mind, by averaging over time, it is possible to compare covariances and mean patterns across countries included in the ESS when using the full data set. This is really encouraging news for theory. Researchers interested in value effects can confidently compare value correlations and mean patterns using the combined data.

TABLE 6 | Exploration of invariance parameters.

	χ^2 factor		SRMR/RMSEA		CFI/TLI configural		CFI/TLI metric		χ^2 factor		SRMR/RMSEA		CFI/TLI configural		CFI/TLI metric	
Predictors	**Estimates**	***p***	**Estimates**	**p**	**Estimates**	***p***	**Estimates**	***p***	**Estimates**	***p***	**Estimates**	***p***	**Estimates**	***p***	**Estimates**	***p***
Intercept	1359.13 (1222.60–1495.67)	**<0.001**	0.06 (0.06–0.07)	**<0.001**	0.93 (0.92–0.93)	**<0.001**	0.86 (0.85–0.87)	**<0.001**	1215.13 (1055.78–1374.47)	**<0.001**	0.06 (0.06–0.06)	**<0.001**	0.93 (0.93–0.94)	**<0.001**	0.87 (0.86–0.88)	**<0.001**
Year (linear)	−1.95 (−8.05 to 4.15)	0.529	−0.00 (−0.00 to −0.00)	**0.041**	0.00 (−0.00 to 0.00)	0.870	−0.00 (−0.00 to −0.00)	**<0.001**	12.81 (2.41–23.21)	**0.016**	0.00 (−0.00 to 0.00)	0.072	−0.00 (−0.00 to −0.00)	**0.028**	−0.00 (−0.00 to −0.00)	**0.001**
Mean value rating	−35.19 (−525.90 to 455.52)	0.888	−0.00 (−0.01 to 0.01)	0.867	−0.01 (−0.03 to 0.02)	0.515	−0.04 (−0.07 to −0.01)	**0.017**	−307.34 (−811.68 to 197.00)	0.231	−0.01 (−0.01 to.00)	0.189	0.00 (−0.02 to.03)	0.808	−0.03 (−0.07 to 0.00)	0.070
SD value rating	−3013.30 (−3678.32 to −2348.29)	**<0.001**	−0.03 (−0.04 to −0.02)	**<0.001**	−0.06 (−0.10 to −0.03)	**<0.001**	−0.06 (−0.11 to −0.02)	**0.005**	−0.2924.80 (−0.3576.30 to −0.2273.31)	**<0.001**	−0.03 (−0.04 to −0.02)	**<0.001**	−0.07 (−0.10 to −0.03)	**<0.001**	−0.07 (−0.11 to −0.02)	**0.003**
Year (squared)									−0.106.81 (−0.168.41 to −0.45.21)	**0.001**	−0.00 (−0.00 to −0.00)	**<0.001**	0.00 (0.00–01)	**0.005**	0.00 (−0.00 to.01)	0.129
Random effects																
ICC	0.63		0.78		0.53		0.52		0.64		0.79		0.55		0.52	
Marginal R^2/Conditional R^2	0.106/0.669		0.039/0.792		0.024/0.546		0.053/0.548		0.121/0.685		0.050/0.803		0.038/0.565		0.057/0.549	

Note: Data based on 261 samples from 39 countries.

Abbreviations: CFI, Comparative Fit Index; RMSEA, Root Mean Square of Approximation; SRMR, Standardized Root Mean Residual; TLI, Tucker–Lewis Index.

TABLE 7 | Major research questions and implications from the current analysis.

Key theoretical question	Affirmative empirical evidence	Implications for researchers	Remaining uncertainty
Does the ESS value data support the Schwartz value model?	Full data set Averaged within-country data Average sample data	Ten value types can be differentiated in empirical data	Specific data sets deviate from the theoretical model when tested individually (Belgium, Switzerland, Cyprus, Estonia, France, United Kingdom, Croatia, Hungary, Italy, Lithuania, Luxemburg, Latvia, Montenegro, Romania, Slovenia and Kosovo); systematic and random effects need to be further explored; latent variable correlations and modification indices suggest both theoretical and measurement adjustments are desirable
Can correlations of values with other measures be compared across countries?	When using the full data set (including all rounds)	Correlations of values with 3rd variables are interpretable	Correlations for samples from specific rounds and countries are open to alternative interpretations (see Table 8, columns 'no configural invariance' and 'configural invariance')
Can mean patterns be compared across countries?	When using the full data set (including all rounds)	Mean patterns (relative importance of value types to each other) are interpretable	Mean patterns for samples from specific rounds and countries are open to alternative interpretations (see Table 8, columns 'no configural invariance' and 'configural invariance')
Can value means be directly compared across countries?	No	Direct mean comparisons are open to alternative interpretations for all countries	Identify levels of practical invariance thresholds
Can correlations with values be compared across time?	Partially, 17 countries showed metric invariance over time	Correlations can be directly and safely compared for some countries (Bulgaria, Czech, Spain, Finland, Iceland, Ireland, Israel, Montenegro, Netherlands, Norway, Poland, Russia, Sweden, Slovakia and Turkey; see Figure 3, Tables 5 and 8)	Correlations across time for some samples (Austria, Belgium, Switzerland, Cyprus, Germany, Estonia, France, United Kingdom, Greece, Croatia, Hungary, Ireland, Italy, Lithuania, Portugal, Slovenia and Ukraine; see Figure 3, Tables 5 and 8) need to be interpreted with care
Can mean patterns of values be compared across time?	Partially, 17 countries showed metric invariance over time	Mean patterns can be directly and safely compared for some countries (Bulgaria, Czech, Spain, Finland, Iceland, Ireland, Israel, Montenegro, Netherlands, Norway, Poland, Russia, Sweden, Slovakia and Turkey; see Figure 3, Tables 5 and 8)	Mean patterns across time for some samples need to be interpreted with care (Austria, Belgium, Switzerland, Cyprus, Germany, Estonia, France, United Kingdom, Greece, Croatia, Hungary, Ireland, Italy, Lithuania, Portugal, Slovenia and Ukraine; see Figure 3, Tables 5 and 8)

(Continues)

TABLE 7 | (Continued)

Key theoretical question	Affirmative empirical evidence	Implications for researchers	Remaining uncertainty
Can value means be compared across time?	Partially, 8 countries showed scalar invariance over time	Means can be directly compared for a small number of countries (Ireland, Israel, Montenegro, Poland, Russia, Sweden and Slovakia; see Figure 3, Tables 5 and 8)	Mean comparisons across time for most countries are open to alternative interpretations (see Figure 3, Tables 5 and 8
Are there systematic effects on model fit in the ESS value data?	Yes, there are systematic time and mean value response effects (see Table 6)	Temporal patterns of model fit and possible drivers need greater attention; mean value rating may imply substantive effects	Identify systematic drivers of variability in model fit

Abbreviation: ESS, European Social Survey.

To some extent, this pattern may seem at odds with the previous research suggesting that structure may be less well differentiated in less economically advantaged societies (Bilsky et al. 2011; Fontaine et al. 2008). It is important to remember that our CFA models were not well suited to directly address these issues, as some of the effects may play out in increasing covariations among the latent variables, which we did not constrain to being identical across samples or over time.

At the same time, the pattern with individual samples from each round and country suggests greater variability, implying that data from particular rounds are not necessarily comparable across countries. Similarly, time trends in correlations and mean patterns within countries may not be comparable over time.

The separation between the overall data set and specific rounds raises some interesting questions. First, there may be random variation for each round that cancels out in the overall data set but may negatively affect the results for specific rounds. There are various data collection parameters within each round that may affect overall data quality (e.g., changes in item translations, location of the value module within the overall data collection and mode of data collection). These effects may not individually be salient (Davidov and De Beuckelaer 2010; Davidov and Depner 2011), but they may interact with each other in yet unexplored ways.

Second, there may be systematic effects that vary over time or due to some external factors that cancel out in the overall data set. Consistent with this possibility, the final examination of the fit indices for each sample, using both the baseline configural and the constrained factor loading using different reference matrices, suggests that temporal effects may be at work, affecting model fit. The curvilinear results raise the possibility that economic factors within and between countries and over time merit further investigation. We centred the curvilinear effect on the year 2008 due to the effects of the global economic crisis. As a consequence, our interpretation needs to be treated with suitable caution in the absence of examining proper economic effects. Nevertheless, even as indicated by these curvilinear time effects, temporal effects may cancel each other out when looking at the full data set and may appear at first glance to be random fluctuations. Taking all evidence in consideration, we believe it is worthwhile to investigate such effects further and to examine possible temporal predictors.

Furthermore, the variability in mean scores also suggests that response patterns for particular rounds may affect model fit. In the context of increasing polarization and shifting norms about values, this is also a fruitful avenue for further work. We used the same information that is typically applied for standardization or ipsatization purposes. As indicated by our results, the information captured by the means may carry useful information that is worth exploring further.

4.2 | Hidden Biases and Avenues for Future Research

The results were less ideal and the pattern somewhat surprising when examining the model fit of the theoretical model to the country data overall (despite the metric invariance found in the multi-group CFA), as well as for the pattern of temporal invariance. For temporal non-invariance, it is striking that Central, Western and Southern European countries did not show temporal invariance. In other words, this suggests that the concept of values or the measurement properties, or both, have changed in central and southern Europe. This certainly challenges attempts to track value changes by looking at simple mean patterns or correlation strength of values in these countries.

On the other hand, it opens very interesting avenues for a deeper exploration of what may have changed about values. Some of these countries are core democracies that have driven the European project. They are also highly developed economically. Both factors should lead to a clearer value structure (Fischer 2017; Fontaine et al. 2008; Strack and Dobewall 2012; van de Vijver and Poortinga 2002). However, it may be the recent economic and social changes that are partly responsible for changes in how individuals interpret values (for some evidence in a broader cultural context, see van de Vijver and Poortinga (2002). For example, the economic challenges in these countries may have led to a reinterpretation and change in the meaning of values, resulting in systematic shifts in structure that lead to a poor fit of the theoretical model to the data. The time effects in our exploratory analyses certainly point in this direction. Such trends can be queried with response shift models (Rapkin and Schwartz 2019) when linked to measurement invariance analysis in data using the same individuals (Leitgöb et al. 2023; Oort 2005). Separating true change from reprioritization and recalibration in data from the same individuals allows a deeper understanding of how individuals make sense of survey items over time. This is an exciting possibility that deserves further attention.

There are two other points we would like to make here. First, we want to make clear that this requires a rethinking of the standard approach of simply comparing means when discussing temporal, economic or democratic effects. The point is that values as theoretical constructs have changed substantially. There has been a cognitive shift within a population that might lead to a reinterpretation and different internal organization of values. The classical philosophical concept is freedom, with the noted differences of freedom to versus freedom from (Berlin 2000). We believe that such a focus represents an interesting and important theoretical reorientation of traditional comparative work and can do much conceptual work for understanding evolutionary change in human societies. Reinterpreting a social construct can open up new ways of thinking and acting that lead to further downstream changes that cannot be explained by looking at the mean patterns alone.

Second, we agree with the recent critique of partial invariance (Robitzsch and Lüdtke 2023) that freeing parameters to improve fit may obscure rather than illuminate meaningful patterns. When freeing parameters, we begin to test different theoretical models (Sterner et al. 2024), and it becomes unclear which model is being compared with which other model when multiple parameters are changed across a large number of samples. It is intellectually more interesting to carefully examine possible sources of misfit than to empirically change model parameters without a better understanding of the underlying patterns. We believe it is worthwhile exploring such patterns more systematically, as they may imply systematic cognitive changes. Given the circularity of inferences in the absence of a reference standard, it may be useful to conduct more focused qualitative studies that examine how people have understood values in these countries (Behr et al. 2014, 2017). Alternatively, media and textual analyses of values may hold promise for understanding changes in value meanings (e.g., changes in the word embeddings of specific values; Bamler and Mandt 2017; Hamilton et al. 2016).

4.3 | Evaluating Fit in Large Cross-Cultural Studies

Previous simulation research had pointed towards TLI as a crucial fit indicator (Hu and Bentler 1999; Marsh et al. 1988), being relatively independent of sample size and sensitive to model misspecifications. Yet, value data may be more complex than traditional CFA models, having a large number of latent variables which are systematically linked to each other via two major underlying dimensions. Using data with more than 10 samples may create additional problems for estimating fit (Rutkowski and Svetina 2014). In our data set, TLI consistently trailed in terms of overall fit and was nearly consistently below a considerably lenient threshold of 0.9. On the other hand, RMSEA and SRMR nearly consistently showed acceptable to even excellent fit. In light of these patterns, we strongly encourage systematic simulation studies with more complex models that approximate the complexity of value research and identify suitable fit indices.

Furthermore, we also encourage work that examines practically meaningful thresholds of invariance, showing what level of non-invariance may challenge interpretation of the data and change inferences about specific theoretical or practical issues (Oberski 2014). As we noted above, some of the overall country samples did not fit when tested individually but then passed the overall configural and metric invariance test using multi-group CFA. Although fit in individual samples has been considered a precondition prior to any further invariance tests, it is not clear to what extent researchers rigorously test and report such tests, especially if the number of samples is moderate to large. As we noted in our study, the results may look rather different. We could claim both solid evidence of metric invariance across cultures and simultaneously evidence of cultural relativism based on the fact that a little less than half of the countries did not show acceptable fit in separate analyses. We need to have clearer guidelines for running and reporting multi-group CFA results. We report the full results in the supplement to provide some guidance for other researchers.

4.4 | Extending Invariance Testing With a More Diverse Methodological Tool Kit

Although we report invariance tests using a widely known method (MG-CFA), there have been significant advances in this area in recent years (Leitgöb et al. 2023). In our last section in the results, we reported some exploratory models. One promising statistical approach for more systematically testing parameters that may influence model fit at the individual and sample level are nonlinear moderated factor analysis models (Bauer et al. 2020; Belzak 2023). We strongly encourage researchers to develop appropriate theoretical models about plausible causal variables that influence model fit (Sterner et al. 2024) and to then test them with appropriate statistical tools (Bauer et al. 2020; Belzak

TABLE 8 | Specific information on invariance status for specific rounds and analyses.

Country	ISO2	No. of rounds available	No configural invariance	Configural invariance	Metric invariance (full data)	Metric invariance (pooled country)	Metric invariance (pooled country and round)	Temporal invariance
Albania	AL	1		6				NA
Austria	AT	7		3, 8, 9, 11	1, 2	2, 7	2, 7	None
Belgium	BE	10		1, 2, 3, 4, 5, 6, 7, 8, 9, 10				None
Bulgaria	BG	6		3, 4, 5, 6, 9, 10				Metric
Switzerland	CH	11		1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11				None
Cyprus	CY	5		3, 4, 5, 6, 9				None
Czech Republic	CZ	9		1, 4, 8, 9, 10	2, 5, 6, 7	5, 7	5, 7	Metric
Germany	DE	10		2, 3, 4, 5, 6, 7, 8, 9, 11	1	1	1	None
Denmark	DK	8		1, 2, 3, 4, 5, 6, 7, 9				Metric
Estonia	EE	9		2, 3, 4, 5, 6, 7, 8, 9, 10				None
Spain	ES	9		2, 4, 7, 8	1	3, 5, 6, 9	3, 5, 6, 9	Metric
Finland	FI	11		7, 8, 9, 11	1, 2, 3, 4, 5, 10	1, 2, 3, 4, 5, 6, 10	1, 2, 3, 4, 5, 6, 10	Metric
France	FR	10		1, 2, 3, 4, 5, 6, 7, 8, 9, 10				None
United Kingdom	GB	11		1, 2, 3, 4, 5, 6, 7, 8, 9, 11	10	10	10	None
Greece	GR	5		1, 2, 4, 5, 10				None
Croatia	HR	5	5	4, 9, 10, 11				None
Hungary	HU	11	7	1, 2, 3, 4, 5, 6, 8, 9, 10, 11				None
Ireland	IE	11		1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11				Scalar
Israel	IL	6		1, 5, 6	7, 8	4, 7, 8	4, 7, 8	Scalar
Iceland	IS	5		2, 6, 8, 9, 10				Metric
Italy	IT	4	8, 9	6, 10				None
Lithuania	LT	7	9	5, 6, 7, 8, 10, 11				None
Luxembourg	LU	1		2				NA
Latvia	LV	2	4	9				None
Montenegro	ME	2	9	10				Scalar
North Macedonia	MK	1		10				NA
Netherlands	NL	11		1, 2, 3, 6, 7, 8, 9, 10, 11		4, 5	4, 5	Metric
Norway	NO	11		2, 6, 8, 9	1, 3, 4	1, 3, 4, 5, 7, 10, 11	1, 3, 4, 5, 7, 10, 11	Metric
Poland	PL	9		1, 2, 3, 4, 5, 6, 7, 8, 9				Scalar
Portugal	PT	10		1, 2, 3, 4, 5, 6, 7, 8, 9, 10				None
Romania	RO	1		4				NA
Serbia	RS	1		9				NA

(Continues)

TABLE 8 | (Continued)

Country	ISO2	No. of rounds available	No configural invariance	Configural invariance	Metric invariance (full data)	Metric invariance (pooled country)	Metric invariance (pooled country and round)	Temporal invariance
Russia	RU	5		3, 4, 5, 6, 8				Scalar
Sweden	SE	9		1, 2, 3, 4, 5, 7, 8, 9		6	6	Scalar
Slovenia	SI	11	6, 9	1, 2, 3, 4, 5, 7, 8, 10, 11				None
Slovakia	SK	8		2, 3, 4, 5, 6	9, 10	9, 10, 11	9, 10, 11	Scalar
Turkey	TR	2		2, 4				Metric
Ukraine	UA	5	6	2, 3, 4, 5				None
Kosovo	XK	1		6				NA

Note: Numbers denote the specific round (from 1 to 11), unless otherwise noted. Countries are ordered alphabetically by ISO2.

2023). We chose not to focus on these more complex models here because our primary goal was to provide empirical evidence on the theoretical fit.

We are aware of discussions about the appropriateness of using CFA latent variable models for the empirical fit to theories like values that involve a circumplex structure and may not be latent variables (Bilsky et al. 2011; Schwartz 1992; Schwartz and Boehnke 2004). At the same time, a large number of recent studies have used CFA as a tool to assess model fit and invariance, and we believe it is important to provide baseline information about how well these data fit using these widely reported statistical models (Cieciuch et al. 2018; Schwartz et al. 2012; Schwartz and Cieciuch 2022). It would be informative to use other statistical tests that are more theoretically aligned with values, involve less statistical assumptions and provide intuitively interpretable results (such as multidimensional scaling or smallest space analysis) to evaluate the overall model fit (Bilsky et al. 2011; Fischer and Karl 2019; Fontaine et al. 2008; Schwartz 1992).

In addition, different approaches, such as alignment or approximate invariance (Asparouhov et al. 2015; Asparouhov and Muthén 2014; Leitgöb et al. 2023; Muthén and Asparouhov 2012), appear promising for further exploration. These methods follow an approximate measurement invariance approach which allows for small measurement differences in the model that are presumed to be small, may cancel each other out within the model and therefore are substantially insignificant (Leitgöb et al. 2023). Their advantage is that they accommodate relatively minor differences between groups and still allow to use the power of latent variable models to be used for cross-group comparisons.

Focusing on specifics, Bayesian invariance methods assume a joint probability distribution of the parameters (e.g., loadings and intercepts) across groups. The researcher needs to specify these distributions a priori. This is one of the most salient challenges, given that the choice of these priors will affect the results. One option is to run models with different priors, followed by sensitivity analyses (van Erp et al. 2018) and comparisons of model results for group means and rankings (Arts et al. 2021). In the absence of clear theoretical priors of these distributions, more work on the practical implications with complex constructs such as human values is desirable.

In contrast, alignment assumes that observed differences in measured indicators are mainly due to differences in the latent variable. Conceptually similar to the rotational indeterminacy problem in exploratory factor analysis, the goal of alignment is to rotate parameters so that a maximal number of parameters becomes invariant, without altering the overall model fit (Asparouhov and Muthén 2014). As noted by the developers, the method is likely to work best if less than 20% of the parameters are non-invariant. Given the observed non-invariance problems in the value data, it is questionable whether this method may be informative for the full ESS data. Furthermore, as noted by Leitgöb et al. (2023), it can be considered a one-step automation of partial invariance testing. Furthermore, given the conceptual closeness to partial invariance testing, this method suffers from the interpretational problems noted by Robitzsch and Lüdtke (2023).

Dimensional reduction techniques discussed so far rely primarily on the covariance structure dictated by the observations (with the exception of Bayesian approaches which additionally require specification of priors). One important avenue for future research is to combine external quality information with the data from survey studies (Lilleoja and Saris 2015; Pirralha and Weber 2020; Saris and Gallhofer 2007; Saris and Revilla 2016). The most widely used system with social and political science data has been the Survey Quality Predictor (Lilleoja and Saris 2015; Pirralha and Weber 2020; Saris and Gallhofer 2007). It involves independent ratings of the survey items, application procedure and study context in terms of validity, reliability and common method variance, which can then be included in the measurement model. This approach has shown promising results in an earlier application with value data from one large European sample (Lilleoja and Saris 2015). Current challenges are that there are no quality ratings available for the ESS values in the most recent version (Felderer et al. 2024), all quality ratings are based on human coding and therefore may not capture all relevant information for the specific country samples and rounds (Pirralha and Weber 2020), and the estimation of complex models such as values may run into non-convergence issues (Lilleoja and Saris 2015).

The combination of qualitatively derived information on data quality nevertheless is one important further avenue that has been largely underutilized (Fischer et al. 2025; Leitgöb et al. 2023). Web probing (Behr et al. 2017) in particular shows some promise in combination with statistical invariance techniques to advance our understanding of the sources of non-invariance (Fischer et al. 2025; Leitgöb et al. 2023; Meitinger 2017).

An additional option to consider in future research would be to explore if there are subgroups of either countries or time points or a mixture of both which show invariance of specific parameters. At the most basic level, the idea is to run some form of clustering analysis on pairwise parameter estimates to identify groups of samples that show invariance (Cheung and Rensvold 2000; Welkenhuysen-Gybels and van de Vijver 2007). More recently, these ideas have been further developed via mixture multi-group factor analysis (De Roover et al. 2022) and the measurement invariance explorer (Rudnev 2024). The mixture multi-group factor analysis approach uses a latent class approach based on the model parameters, whereas the measurement invariance explorer visualizes distances between samples based on multidimensional scaling or network analysis using fit indices. Both methods are promising, but not without challenges. First, the selection of the number of classes or clusters can be challenging, especially when classifying samples (either countries or time points or both) on more than one parameter. A second, more practical issue is that robust identification requires enough groups, which is restricted by the number of countries within Europe. A third issue that needs some attention is complexity of the models, because these methods tend to work better with a smaller number of factors and indicators as well as assume equal number of factors across all classes or clusters (for a general review of the advantages and disadvantages, see Leitgöb et al. 2023).

In summary, each of the aforementioned approaches has distinct advantages and disadvantages. Drawing inspiration from adversarial collaborations and many-analyst approaches, we believe large-scale comparative studies that test diverse methodological approaches competitively are a promising avenue. Such studies would enable researchers to identify parameters that consistently demonstrate misfit across methods with varying levels of complexity and different underlying statistical assumptions, an approach analogous to multiverse analysis (Steegen et al. 2016). Implementing a multiverse framework would involve systematically applying multiple analytical strategies to the same data set and examining the convergence (or divergence) of findings across approaches. Parameters showing consistent patterns of misfit across methodologically diverse analyses would warrant particular attention, as they likely represent genuine model inadequacies rather than method-specific artefacts. Conversely, parameters showing method-dependent results would highlight the importance of analytical choices and assumptions. We believe such comparative studies adopting a multiverse perspective represent a crucial methodological advancement for the field, offering a more robust foundation for drawing substantive conclusions about model fit and parameter estimation.

5 | Limitations

As outlined in the previous paragraph, our approach was based on one popular method. We discuss alternative methods, which should ideally be tested competitively to identify which samples and parameters are invariant as well as further explore any source of such invariance. One central challenge has been that invariance becomes a perceived barrier for social and behavioural science (Funder and Gardiner 2024; Welzel et al. 2023). Yet, in our opinion, the information provided by these tests is not fully explored from a theoretical perspective (Fischer et al. 2025; Fischer and Rudnev 2024).

Our model and parameter choices might also be questioned for being overly lenient. We did not apply more stringent thresholds given the complexity of the theoretical model. We also did not further explore modification indices (Saris et al. 2009). Such explorations may have resulted in sample-specific modifications, and it raises interesting questions on the theoretical implications (Robitzsch and Lüdtke 2023).

A further limitation is that we only included individuals that answered all value items. It might be informative in future analyses to examine whether some items show differential non-response patterns which may indicate specific sensitive issues for that language or sociocultural context.

6 | Conclusions

Our analysis provides essential information for evaluating the robustness of value research embedded within the ESS. We do find encouraging patterns of cross-cultural invariance that suggest comparability of correlations and mean patterns. At the same time, specific samples within each round may not show sufficient robustness to allow secure comparison across countries. Furthermore, temporal invariance was not widely supported, suggesting that the conceptualization of values has changed in at least some of the countries that have been measuring values regularly. The fit has worsened over time, which requires careful further explorations of the theoretical processes that may drive these effects.

Author Contributions

Ronald Fischer: designed the study idea, developed the analysis plan, wrote the analysis code and interpreted the results, wrote the first draft, revised the draft and agreed on the final version. **Johannes A. Karl**: developed the analysis plan, wrote the analysis code and interpreted the results, revised the draft and agreed on the final version.

Acknowledgements

We would like to thank the editor, an anonymous reviewer, Maksim Rudnev, Larissa Hartle and Giovanna Bortolini for useful comments on earlier drafts of this paper. The first author would also like thank Joanne Sneddon for the inspirations and encouragement which provided the impetus for the current project.

The Article Processing Charge for the publication of this research was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) (ROR identifier: 00x0ma614).

Ethics Statement

The European Social Survey European Research Infrastructure Consortium (ESS ERIC) subscribes to and abides by the Declaration on Ethics of the International Statistical Institute (https://isi-web.org/declaration-professional-ethics).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data are available at https://www.europeansocialsurvey.org/data-portal.

Transparency Statement

The code and further results are available at https://osf.io/4tma8. The analyses were not preregistered. A preprint of an earlier version is available here: https://doi.org/10.31234/osf.io/hrwvb_v2.

Preregistration

The analyses were not preregistered.

Endnotes

[1] We considered testing parameters centred on the COVID-19 effects (centred around 2020), but because the pandemic only affected two rounds and it is not clear how these pandemic effects play out over time (Daniel et al. 2022; Sneddon et al. 2022), we decided against it for our current analysis.

References

Arts, I., Q. Fang, R. van de Schoot, and K. Meitinger. 2021. "Approximate Measurement Invariance of Willingness to Sacrifice for the Environment Across 30 Countries: The Importance of Prior Distributions and Their Visualization." *Frontiers in Psychology* 12: 624032. https://doi.org/10.3389/fpsyg.2021.624032.

Asparouhov, T., and B. Muthén. 2014. "Multiple-Group Factor Analysis Alignment." *Structural Equation Modeling: A Multidisciplinary Journal* 21, no. 4: 495–508. https://doi.org/10.1080/10705511.2014.919210.

Asparouhov, T., B. Muthén, and A. J. S. Morin. 2015. "Bayesian Structural Equation Modeling With Cross-Loadings and Residual Covariances: Comments on Stromeyer et al." *Journal of Management* 41, no. 6: 1561–1577. https://doi.org/10.1177/0149206315591075.

Bamler, R., and S. Mandt. 2017. Dynamic Word Embeddings (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1702.08359.

Bauer, D. J., W. C. M. Belzak, and V. Cole. 2020. "Simplifying the Assessment of Measurement Invariance Over Multiple Background Variables: Using Regularized Moderated Nonlinear Factor Analysis to Detect Differential Item Functioning." *Structural Equation Modeling: A Multidisciplinary Journal* 27, no. 1: 43–55. https://doi.org/10.1080/10705511.2019.1642754.

Behr, D., K. Meitinger, M. Braun, and L. Kaczmirek. 2017. "Web Probing – Implementing Probing Techniques From Cognitive, Interviewing in Web Surveys With the Goal to Assess the Validity of Survey Questions." GESIS Survey Guidelines. https://doi.org/10.15465/GESIS-SG_EN_023.

Belzak, W. C. M. 2023. "The regDIF R Package: Evaluating Complex Sources of Measurement Bias Using Regularized Differential Item Functioning." *Structural Equation Modeling: A Multidisciplinary Journal* 30, no. 6: 974–984. https://doi.org/10.1080/10705511.2023.2170235.

Bentler, P. M. 1990. "Comparative Fit Indexes in Structural Models." *Psychological Bulletin* 107, no. 2: 238–246. https://doi.org/10.1037/0033-2909.107.2.238.

Bentler, P. M., and D. G. Bonett. 1980. "Significance Tests and Goodness of Fit in the Analysis of Covariance Structures." *Psychological Bulletin* 88, no. 3: 588–606. https://doi.org/10.1037/0033-2909.88.3.588.

Berlin, I. 2000. "Two Concepts of Liberty." In *Reading Political Philosophy*. Routledge.

Bilsky, W., M. Janik, and S. H. Schwartz. 2011. "The Structural Organization of Human Values-Evidence From Three Rounds of the European Social Survey (ESS)." *Journal of Cross-Cultural Psychology* 42, no. 5: 759–776. https://doi.org/10.1177/0022022110362757.

Boer, D., K. Hanke, and J. He. 2018. "On Detecting Systematic Measurement Error in Cross-Cultural Research: A Review and Critical Reflection on Equivalence and Invariance Tests." *Journal of Cross-Cultural Psychology* 49, no. 5: 713–734. https://doi.org/10.1177/0022022117749042.

Bollen, K. A. 1989. *Structural Equations With Latent Variables*. Wiley. https://doi.org/10.1002/9781118619179.

Brosch, T., Y. Stussi, O. Desrichard, and D. Sander. 2018. "Not My Future? Core Values and the Neural Representation of Future Events." *Cognitive, Affective & Behavioral Neuroscience* 18, no. 3: 476–484. https://doi.org/10.3758/s13415-018-0581-9.

Browne, M. W., and R. Cudeck. 1992. "Alternative Ways of Assessing Model Fit." *Sociological Methods & Research* 21, no. 2: 230–258. https://doi.org/10.1177/0049124192021002005.

Byrne, B. M., R. J. Shavelson, and B. Muthén. 1989. "Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance." *Psychological Bulletin* 105, no. 3: 456–466. https://doi.org/10.1037/0033-2909.105.3.456.

Cheung, G. W., and R. B. Rensvold. 2000. "Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research Using Structural Equations Modeling." *Journal of Cross-Cultural Psychology* 31, no. 2: 187–212. https://doi.org/10.1177/0022022100031002003.

Cieciuch, J., E. Davidov, R. Algesheimer, and P. Schmidt. 2018. "Testing for Approximate Measurement Invariance of Human Values in the European Social Survey." *Sociological Methods & Research* 47, no. 4: 665–686. https://doi.org/10.1177/0049124117701478.

Coelho, G. L. H., P. H. P. Hanel, M. K. Johansen, and G. R. Maio. 2019. "Mapping the Structure of Human Values Through Conceptual Representations." *European Journal of Personality* 33, no. 1: 34–51. https://doi.org/10.1002/per.2170.

Daniel, E., A. Bardi, R. Fischer, M. Benish-Weisman, and J. A. Lee. 2022. "Changes in Personal Values in Pandemic Times." *Social Psychological and Personality Science* 13, no. 2: 572–582.

Davidov, E. 2008. "A Cross-Country and Cross-Time Comparison of the Human Values Measurements With the Second Round of the European Social Survey." *Survey Research Methods* 2: 33–46. https://doi.org/10.5167/UZH-95236.

Davidov, E. 2010. "Testing for Comparability of Human Values Across Countries and Time With the Third Round of the European Social Survey." *International Journal of Comparative Sociology* 51, no. 3: 171–191. https://doi.org/10.1177/0020715210363534.

Davidov, E., G. Datler, P. Schmidt, and S. H. Schwartz. 2018. "Testing the Invariance of Values in the Benelux Countries With the European Social Survey: Accounting for Ordinality." In *Cross-Cultural Analysis*. 2nd ed. Routledge.

Davidov, E., and A. De Beuckelaer. 2010. "How Harmful Are Survey Translations? A Test With Schwartz's Human Values Instrument." *International Journal of Public Opinion Research* 22, no. 4: 485–510. https://doi.org/10.1093/ijpor/edq030.

Davidov, E., and F. Depner. 2011. "Testing for Measurement Equivalence of Human Values Across Online and Paper-and-Pencil Surveys." *Quality & Quantity* 45, no. 2: 375–390. https://doi.org/10.1007/s11135-009-9297-9.

Davidov, E., P. Schmidt, and S. H. Schwartz. 2008. "Bringing Values Back in: The Adequacy of the European Social Survey to Measure Values in 20 Countries." *Public Opinion Quarterly* 72, no. 3: 420–445. https://doi.org/10.1093/poq/nfn035.

De Roover, K., J. K. Vermunt, and E. Ceulemans. 2022. "Mixture Multigroup Factor Analysis for Unraveling Factor Loading Non-Invariance

Across Many Groups." *Psychological Methods* 27, no. 3: 281–306. https://doi.org/10.1037/met0000355.

European Social Survey European Research Infrastructure (ESS ERIC). 2024. "European Social Survey (ESS), Round 11–2023." Sikt – Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ESS11-2023.

Felderer, B., L. Repke, W. Weber, J. Schweisthal, and L. Bothmann. 2024. *Predicting the Validity and Reliability of Survey Questions*. Open Science Framework. https://doi.org/10.31219/osf.io/hkngd.

Fischer, R. 2004. "Standardization to Account for Cross-Cultural Response Bias: A Classification of Score Adjustment Procedures and Review of Research in JCCP." *Journal of Cross-Cultural Psychology* 35, no. 3: 263–282.

Fischer, R. 2017. *Personality, Values, Culture: An Evolutionary Approach*. Cambridge University Press. https://doi.org/10.1017/9781316091944.

Fischer, R., and J. A. Karl. 2019. "A Primer to (Cross-Cultural) Multi-Group Invariance Testing Possibilities in R." *Frontiers in Psychology* 10: 1507. https://doi.org/10.3389/fpsyg.2019.01507.

Fischer, R., and J. A. Karl. 2023. "Niche Diversity Effects on Personality Measurement – Evidence From Large National Samples During the COVID-19 Pandemic." *Current Research in Ecological and Social Psychology* 4: 100116. https://doi.org/10.1016/j.cresp.2023.100116.

Fischer, R., J. A. Karl, J. R. Fontaine, and Y. H. Poortinga. 2021. "Evidence of Validity Does NOT Rule Out Systematic Bias: A Commentary on Nomological Noise and Cross-Cultural Invariance." *Sociological Methods & Research* 52: 1–18. https://doi.org/10.1177/00491241221091756.

Fischer, R., J. A. Karl, M. Luczak-Roesch, and L. Hartle. 2023. "Why We Need to Rethink Measurement Invariance: The Role of Measurement Invariance for Psychological Science." Preprint, PsyArXiv. https://doi.org/10.31234/osf.io/3f9ca.

Fischer, R., J. A. Karl, M. Luczak-Roesch, and L. Hartle. 2025. "Why We Need to Rethink Measurement Invariance: The Role of Measurement Invariance for Cross-Cultural Research." *Cross-Cultural Research* 59, no. 2: 147–179. https://doi.org/10.1177/10693971241312459.

Fischer, R., and M. Rudnev. 2024. "From MIsgivings to MIse-en-scène: The Role of Invariance in Personality Science." *European Journal of Personality* 39, no. 4: 662–673. https://doi.org/10.1177/08902070241283081.

Fontaine, J. R., Y. H. Poortinga, L. Delbeke, and S. H. Schwartz. 2008. "Structural Equivalence of the Values Domain Across Cultures: Distinguishing Sampling Fluctuations From Meaningful Variation." *Journal of Cross-Cultural Psychology* 39, no. 4: 345–365.

Fontaine, J. R. J. 2005. "Equivalence." In *Encyclopedia of Social Measurement*, edited by K. Kempf-Leonard, 803–813. Elsevier. https://doi.org/10.1016/B0-12-369398-5/00116-X.

Funder, D. C., and G. Gardiner. 2024. "MIsgivings About Measurement Invariance." *European Journal of Personality* 38, no. 6: 889–895. https://doi.org/10.1177/08902070241228338.

Gelfand, M. J., J. L. Raver, L. Nishii, et al. 2011. "Differences Between Tight and Loose Cultures: A 33-Nation Study." *Science* 332, no. 6033: 1100–1104.

Graeber, D., and D. Wengrow. 2021. *The Dawn of Everything: A New History of Humanity*. 1st ed. Farrar, Straus and Giroux.

Hamilton, W. L., J. Leskovec, and D. Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://doi.org/10.18653/v1/p16-1141.

Hofstede, G. 1980. *Culture's Consequences: International Differences in Work-Related Values*. SAGE Publications.

Hu, L., and P. M. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives." *Structural Equation Modeling: A Multidisciplinary Journal* 6, no. 1: 1–55. https://doi.org/10.1080/10705519909540118.

Ikizer, E. G., R. Fischer, J. R. Kunst, and J. F. Dovidio. 2024. "Cultural Tightness-Looseness and Individual Differences in Non-Normativeness Predict Stigmatization of Out-Groups: A Multilevel Cross-Cultural Study." *Personality and Social Psychology Bulletin*: 01461672241273285. https://doi.org/10.1177/01461672241273285.

Inglehart, R. 1997. *Modernization and Postmodernization*. Princeton University Press.

Inglehart, R. 2018. *Cultural Evolution: People's Motivations Are Changing, and Reshaping the World*. Cambridge University Press.

Karl, J., and R. Fischer. 2022. "More Than Yes and No: Predicting the Magnitude of Non-Invariance Between Countries From Systematic Features." In Xenophobia vs. Patriotism: Where Is My Home? *Proceedings From the 25th Congress of the International Association for Cross-Cultural Psychology*, edited by M. Klicperova-Baker and W. Friedlmeier, 1–21. International Association for Cross-Cultural Psychology. https://doi.org/10.4087/ELED5219.

Lee, J. A., and G. Soutar. 2010. "Is Schwartz's Value Survey an Interval Scale, and Does It Really Matter?" *Journal of Cross-Cultural Psychology* 41, no. 1: 76–86. https://doi.org/10.1177/0022022109348920.

Leitgöb, H., D. Seddig, T. Asparouhov, et al. 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.

Leszkowicz, E., D. E. J. Linden, G. R. Maio, and N. Ihssen. 2017. "Neural Evidence of Motivational Conflict Between Social Values." *Social Neuroscience* 12, no. 5: 494–505. https://doi.org/10.1080/17470919.2016.1183517.

Leszkowicz, E., G. R. Maio, D. E. J. Linden, and N. Ihssen. 2021. "Neural Coding of Human Values Is Underpinned by Brain Areas Representing the Core Self in the Cortical Midline Region." *Social Neuroscience* 16, no. 5: 486–499. https://doi.org/10.1080/17470919.2021.1953582.

Lilleoja, L., and W. E. Saris. 2015. "Does Correction for Measurement Error Have an Effect on the Structure of Basic Human Values?" *Survey Research Methods* 9: 169–187. https://doi.org/10.18148/SRM/2015.V9I3.6203.

Marsh, H. W., J. R. Balla, and R. P. McDonald. 1988. "Goodness-of-Fit Indexes in Confirmatory Factor Analysis: The Effect of Sample Size." *Psychological Bulletin* 103, no. 3: 391–410. https://doi.org/10.1037/0033-2909.103.3.391.

Meitinger, K. 2017. "Necessary but Insufficient: Why Measurement Invariance Tests Need Online Probing as a Complementary Tool." *Public Opinion Quarterly* 81, no. 2: 447–472. https://doi.org/10.1093/poq/nfx009.

Meredith, W. 1993. "Measurement Invariance, Factor Analysis and Factorial Invariance." *Psychometrika* 58, no. 4: 525–543. https://doi.org/10.1007/BF02294825.

Muthén, B., and T. Asparouhov. 2012. "Bayesian Structural Equation Modeling: A More Flexible Representation of Substantive Theory." *Psychological Methods* 17, no. 3: 313–335. https://doi.org/10.1037/a0026802.

Muthen, B. O. 1994. "Multilevel Covariance Structure Analysis." *Sociological Methods & Research* 22, no. 3: 376–398. https://doi.org/10.1177/0049124194022003006.

Oberski, D. L. 2014. "Evaluating Sensitivity of Parameters of Interest to Measurement Invariance in Latent Variable Models." *Political Analysis* 22, no. 1: 45–60. https://doi.org/10.1093/pan/mpt014.

Ollerenshaw, T. 2023. "Affective Polarization and the Destabilization of Core Political Values." *Political Science Research and Methods* 13: 1–9. https://doi.org/10.1017/psrm.2023.34.

Oort, F. J. 2005. "Using Structural Equation Modeling to Detect Response Shifts and True Change." *Quality of Life Research* 14, no. 3: 587–598. https://doi.org/10.1007/s11136-004-0830-y.

Parsons, T., and E. A. Shils. 1951. "PART 2. Values, Motives, and Systems of Action." In *Toward a General Theory of Action*, edited by T. Parsons and E. A. Shils, 45–276. Harvard University Press. https://doi.org/10.4159/harvard.9780674863507.c4.

Pirralha, A., and W. Weber. 2020. "Correction for Measurement Error in Invariance Testing: An Illustration Using SQP." *PLoS ONE* 15, no. 10: e0239421. https://doi.org/10.1371/journal.pone.0239421.

Poortinga, Y. H. 1989. "Equivalence of Cross-Cultural Data: An Overview of Basic Issues." *International Journal of Psychology* 24, no. 6: 737–756. https://doi.org/10.1080/00207598908247842.

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

Rapkin, B. D., and C. E. Schwartz. 2019. "Advancing Quality-of-Life Research by Deepening Our Understanding of Response Shift: A Unifying Theory of Appraisal." *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* 28, no. 10: 2623–2630. https://doi.org/10.1007/s11136-019-02248-z.

Revelle, W. 2024. "psych: Procedures for Psychological, Psychometric, and Personality Research." [R; R package version 2.4.6]. https://CRAN.R-project.org/package=psych.

Robitzsch, A., and O. Lüdtke. 2023. "Why Full, Partial, or Approximate Measurement Invariance Are Not a Prerequisite for Meaningful and Valid Group Comparisons." *Structural Equation Modeling: A Multidisciplinary Journal* 30, no. 6: 859–870. https://doi.org/10.1080/10705511.2023.2191292.

Rosseel, Y. 2012. "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software* 48, no. 2: 1–36. https://doi.org/10.18637/jss.v048.i02.

Rudnev, M. 2021. "Caveats of Non-Ipsatization of Basic Values: A Review of Issues and a Simulation Study." *Journal of Research in Personality* 93: 104118. https://doi.org/10.1016/j.jrp.2021.104118.

Rudnev, M. 2024. "Measurement Invariance Explorer." [R]. https://github.com/MaksimRudnev/MIE.package.

Rutkowski, L., and D. Svetina. 2014. "Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys." *Educational and Psychological Measurement* 74, no. 1: 31–57. https://doi.org/10.1177/0013164413498257.

Saris, W., and A. Satorra. 2018. "The Pooled Data Approach for the Estimation of Split-Ballot Multitrait–Multimethod Experiments." *Structural Equation Modeling: A Multidisciplinary Journal* 25, no. 5: 659–672. https://doi.org/10.1080/10705511.2018.1431543.

Saris, W. E., and I. Gallhofer. 2007. "Estimation of the Effects of Measurement Characteristics on the Quality of Survey Questions." *Survey Research Methods* 1, no. 1: 29–43. https://doi.org/10.18148/srm/2007.v1i1.49.

Saris, W. E., and M. Revilla. 2016. "Correction for Measurement Errors in Survey Research: Necessary and Possible." *Social Indicators Research* 127, no. 3: 1005–1020. https://doi.org/10.1007/s11205-015-1002-x.

Saris, W. E., A. Satorra, and W. M. van der Veld. 2009. "Testing Structural Equation Models or Detection of Misspecifications?" *Structural Equation Modeling: A Multidisciplinary Journal* 16, no. 4: 561–582. https://doi.org/10.1080/10705510903203433.

Schwartz, S. H., and W. Bilsky. 1987. "Toward a Universal Psychological Structure of Human Values." *Journal of Personality and Social Psychology* 53, no. 3: 550–562. https://doi.org/10.1037/0022-3514.53.3.550.

Schwartz, S. H. 1992. "Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries." In *Advances in Experimental Social Psychology*, edited by M. P. Zanna, Vol 25, 1–65. https://doi.org/10.1016/S0065-2601(08)60281-6.

Schwartz, S. H. 2003. "A Proposal for Measuring Value Orientations Across Nations." *Questionnaire Package of the European Social Survey* 259: 261–318.

Schwartz, S. H. 2006. "A Theory of Cultural Value Orientations: Explication and Applications." *Comparative Sociology* 5, no. 2–3: 137–182. https://doi.org/10.1163/156913306778667357.

Schwartz, S. H., and K. Boehnke. 2004. "Evaluating the Structure of Human Values With Confirmatory Factor Analysis." *Journal of Research in Personality* 38, no. 3: 230–255. https://doi.org/10.1016/S0092-6566(03)00069-2.

Schwartz, S. H., B. Breyer, and D. Danner. 2015. "Human Values Scale (ESS)." *ZIS—The Collection Items and Scales for the Social Sciences*, Mannheim, Germany. https://doi.org/10.6102/ZIS234.

Schwartz, S. H., and J. Cieciuch. 2022. "Measuring the Refined Theory of Individual Values in 49 Cultural Groups: Psychometrics of the Revised Portrait Value Questionnaire." *Assessment* 29, no. 5: 1005–1019. https://doi.org/10.1177/1073191121998760.

Schwartz, S. H., J. Cieciuch, M. Vecchione, et al. 2012. "Refining the Theory of Basic Individual Values." *Journal of Personality and Social Psychology* 103, no. 4: 663–688.

Schwartz, S. H., G. Melech, A. Lehmann, S. Burgess, M. Harris, and V. Owens. 2001. "Extending the Cross-Cultural Validity of the Theory of Basic Human Values With a Different Method of Measurement." *Journal of Cross-Cultural Psychology* 32, no. 5: 519–542. https://doi.org/10.1177/0022022101032005001.

Singh, J. 1995. "Measurement Issues in Cross-National Research." *Journal of International Business Studies* 26, no. 3: 597–619. https://doi.org/10.1057/palgrave.jibs.8490188.

Sivo, S. A., X. Fan, E. L. Witta, and J. T. Willse. 2006. "The Search for "Optimal" Cutoff Properties: Fit Index Criteria in Structural Equation Modeling." *Journal of Experimental Education* 74, no. 3: 267–288. https://doi.org/10.3200/JEXE.74.3.267-288.

Smith, P. B. 2004. "Acquiescent Response Bias as an Aspect of Cultural Communication Style." *Journal of Cross-Cultural Psychology* 35, no. 1: 50–61. https://doi.org/10.1177/0022022103260380.

Sneddon, J., E. Daniel, R. Fischer, and J. Lee. 2022. "The Impact of the COVID-19 Pandemic on Environmental Values." *Sustainability Science* 17, no. 5: 2155–2163. https://doi.org/10.1007/s11625-022-01151-w.

Steegen, S., F. Tuerlinckx, A. Gelman, and W. Vanpaemel. 2016. "Increasing Transparency Through a Multiverse Analysis." *Perspectives on Psychological Science* 11, no. 5: 702–712. https://doi.org/10.1177/1745691616658637.

Steenkamp, J.-B. E. M., and H. Baumgartner. 1998. "Assessing Measurement Invariance in Cross-National Consumer Research." *Journal of Consumer Research* 25, no. 1: 78–90. https://doi.org/10.1086/209528.

Sterner, P., F. Pargent, D. Deffner, and D. Goretzko. 2024. "A Causal Framework for the Comparability of Latent Variables." *Structural Equation Modeling: A Multidisciplinary Journal* 31, no. 5: 747–758. https://doi.org/10.1080/10705511.2024.2339396.

Stevens, S. S. 1946. "On the Theory of Scales of Measurement." *Science* 103, no. 2684: 677–680. https://doi.org/10.1126/science.103.2684.677.

Strack, M., and H. Dobewall. 2012. "The Value Structure in Socioeconomically Less Developed European Countries Still Remains an Ellipse." *Europe's Journal of Psychology* 8, no. 4: 587–602. https://doi.org/10.5964/ejop.v8i4.505.

Teed, A. R., J. Rakic, D. B. Mark, and D. C. Krawcyzk. 2020. "Relative Activation Patterns Associated With Self-Transcendent and Self-Enhancement Core Values: An fMRI Study of Basic Human Values Theory Concepts in Males." *Social Neuroscience* 15, no. 1: 1–14. https://doi.org/10.1080/17470919.2019.1598893.

Tucker, L. R., and C. Lewis. 1973. "A Reliability Coefficient for Maximum Likelihood Factor Analysis." *Psychometrika* 38, no. 1: 1–10. https://doi.org/10.1007/BF02291170.

Uz, I. 2018. "Cross-Validation of Cultural Tightness and Looseness Measures." *International Journal of Psychology* 53, no. 4: 287–294. https://doi.org/10.1002/ijop.12376.

Van Herk, H., and S. P. K. Goldman. 2022. "The Advancement of Measurement Invariance Testing in Cross-Cultural Research in the Period 1999–2020. Executing Rather Than Scrutinizing?" In *Measurement in*

Marketing, edited by H. Baumgartner and B. Weijters, 95–119. Emerald Publishing Limited. https://doi.org/10.1108/S1548-643520220000019005.

Vandenberg, R. J., and C. E. Lance. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 3, no. 1: 4–70. https://doi.org/10.1177/109442810031002.

van de Vijver, F. J. R., and K. Leung. 2000. "Methodological Issues in Psychological Research on Culture." *Journal of Cross-Cultural Psychology* 31: 33–51. https://doi.org/10.1177/0022022100031001004.

van de Vijver, F. J. R., and K. Leung. 2021. *Methods and Data Analysis for Cross-Cultural Research*, edited by V. H. Fetvadjiev, J. He, and J. R. J. Fontaine, 2nd ed., Cambridge University Press. https://doi.org/10.1017/9781107415188.

Van De Vijver, F. J. R., and Y. H. Poortinga. 1982. "Cross-Cultural Generalization and Universality." *Journal of Cross-Cultural Psychology* 13, no. 4: 387–408. https://doi.org/10.1177/0022002182013004001.

van de Vijver, F. J. R., and Y. H. Poortinga. 2002. "Structural Equivalence in Multilevel Research." *Journal of Cross-Cultural Psychology* 33, no. 2: 141–156. https://doi.org/10.1177/0022022102033002002.

van Erp, S., J. Mulder, and D. L. Oberski. 2018. "Prior Sensitivity Analysis in Default Bayesian Structural Equation Modeling." *Psychological Methods* 23, no. 2: 363–388. https://doi.org/10.1037/met0000162.

Vecchione, M., S. Schwartz, G. Alessandri, A. K. Döring, V. Castellani, and M. G. Caprara. 2016. "Stability and Change of Basic Personal Values in Early Adulthood: An 8-Year Longitudinal Study." *Journal of Research in Personality* 63: 111–122. https://doi.org/10.1016/j.jrp.2016.06.002.

Welkenhuysen-Gybels, J., F. van de Vijver, and B. Cambré. 2007. "A Comparison of Methods for the Evaluation of Construct Equivalence in a Multigroup Setting." In *Measuring Meaningful Data in Social Research*, edited by G. Loosveldt, and M. Swyngedouw, 357–371. Leuven/Voorburg.

Welzel, C. 2010. "How Selfish Are Self-Expression Values? A Civicness Test." *Journal of Cross-Cultural Psychology* 41, no. 2: 152–174. https://doi.org/10.1177/0022022109354378.

Welzel, C. 2014. *Freedom Rising: Human Empowerment and the Quest for Emancipation*. Cambridge University Press.

Welzel, C., L. Brunkert, S. Kruse, and R. F. Inglehart. 2023. "Non-Invariance? An Overstated Problem With Misconceived Causes." *Sociological Methods & Research* 52: 1368–1400. https://doi.org/10.1177/0049124121995521.

Ximénez, C., A. Maydeu-Olivares, D. Shi, and J. Revuelta. 2022. "Assessing Cutoff Values of SEM Fit Indices: Advantages of the Unbiased SRMR Index and Its Cutoff Criterion Based on Communality." *Structural Equation Modeling: A Multidisciplinary Journal* 29, no. 3: 368–380. https://doi.org/10.1080/10705511.2021.1992596.