European Political Science Review (2022), **14**, 18–35
doi:10.1017/S1755773921000266

EPSR

RESEARCH ARTICLE

Measuring satisfaction with democracy: how good are different scales across countries and languages?

Carlos Poses* and Melanie Revilla

RECSM – Universitat Pompeu Fabra
*E-mail: carlos.gonzalezp@upf.edu

(Received 03 February 2021; revised 21 July 2021; accepted 29 July 2021; first published online 13 September 2021)

Abstract

The Satisfaction With Democracy (SWD) indicator is very often used in social sciences' research. However, while there is debate about which concept it measures, the discussion about the size of its measurement errors (how well it measures the underlying concept 'satisfaction with the way democracy works') is scarce. Nonetheless, measurement errors can affect the results and threaten comparisons across studies, countries and languages. Thus, in this paper, we estimated the measurement quality (complement of measurement errors) of the SWD indicator for 7 response scales across 38 country-language groups, using three multitrait-multimethod experiments from the European Social Survey. Results show that measurement errors explain from 16% (11-point scale) to 54% (4-point scale) of the variance in the observed responses. Additionally, we provide insights to improve questionnaire design and evaluate the indicator's comparability across scales, countries and languages.

Keywords: Satisfaction With Democracy (SWD); measurement errors; measurement quality; European Social Survey (ESS); multitrait-multimethod (MTMM) experiments

Introduction

Political support (Easton, 1965; 1975) is considered key for the evolution of democracies. Originally, Easton distinguished between a) specific support: essentially, support based on short-term utility and rather immediate performance; and b) diffuse support: a more stable, long-term attachment to the democratic regime (Thomassen and Van Ham, 2017). Drawing on this conceptualization, other scholars developed more refined models of political support, for example, Norris' fivefold model (Norris, 2011).

Many empirical studies of political support concentrated on a specific survey indicator: the Satisfaction With Democracy (SWD) indicator, which asks respondents how satisfied they are with the way democracy works in their country. This indicator is the focus of this paper.

There is a huge literature about the determinants of this indicator (for an overview, see, e.g., Dassonneville and McAllister, 2020). Moreover, many scholars used this indicator as a measure of the third level in Norris' fivefold model of political support (Norris, 2011, p. 28; van Ham and Thomassen, 2017, p. 3). Nevertheless, Ferrín (2016) and Quaranta (2018) pointed out that the SWD indicator has been reported to measure at least 13 different concepts. Besides, the extent to which the SWD indicator really measures these concepts (i.e., its content validity; Bollen, 1989, p. 185-186) has been (and still is) an important source of discussion (Linde & Ekman, 2003; Canache, Mondak & Selignson, 2001; Norris, 2011; Ferrín, 2016). While a high content validity seems guaranteed for the simple concept 'satisfaction with the way democracy works', more theoretical arguments and empirical evidence are needed for the other concepts.

The SWD indicator has been regularly included in major academic surveys, such as Afrobarometer, Asian Barometer, Americas Barometer, Comparative Study of Electoral Systems (CSES), Eurobarometer, European Social Survey (ESS), European Values Study (EVS), or Latinobarometer. However, these surveys measure the SWD indicator in different ways: in particular, they use different answer scales. Almost all surveys use 4-point scales, but these scales vary on other characteristics: for instance, the Eurobarometer uses a unipolar fully labelled scale ('very satisfied', 'fairly satisfied', 'not very satisfied', and 'not at all satisfied'), whereas the Americas Barometer uses a bipolar fully labelled scale ('very satisfied', 'satisfied', 'dissatisfied', 'very dissatisfied'). Additionally, the ESS main questionnaire uses a 11-point bipolar scale with verbal labels only for the endpoints ('0 – Extremely dissatisfied' to '10 – Extremely satisfied'). Based on previous research (e.g., Saris and Gallhofer, 2007; Saris and Revilla, 2016; but also research that considered directly the measurement quality of the SWD indicator, see Table 1), we expect that: 1) none of these scales will lead to a perfect measurement; 2) some of these scales will be better than others (since each response scale has its own level of measurement errors) and 3) not accounting for these measurement errors could affect the results.

This has some implications for the research using the SWD indicator, which has, however, received little attention in the literature. First, some of the substantive findings regarding the SWD indicator may be a byproduct of imperfect measurement instruments not accounted for. Second, mixed results may be linked to different levels of measurement errors across measurement instruments, countries, and languages. For instance, Christmann (2018) summarizes the mixed evidence presented by cross-sectional studies regarding the effects of economics variables on the SWD indicator. Among many others, a potential reason for these differences may be that some of these studies use different scales (e.g., van der Meer and Hakhverdian, 2017, use a 4-point scale with labels 'very satisfied', 'rather satisfied', 'not very satisfied', 'not at all satisfied', whereas Schäfer, 2012 uses the ESS 11-point scale). Lastly, the questions' formulations and/or the scales currently used might not be the ones with the smaller size of measurement errors. Thus, it is important to estimate the size of measurement errors for different scales, and under different conditions (e.g., across time, countries or languages).

Researchers commonly estimate the size of measurement errors by estimating its complement[1], measurement quality. Measurement quality (see also Section 4.1) is a statistical measure ranging from 0 to 1, defined as the strength of the relationship between the latent concept of interest (here the simple concept 'satisfaction with the way democracy works', which is what we really want to assess) and the observed survey answers (here, the answers to the SWD indicator, asked using a given scale). The higher the measurement quality, the better the SWD indicator measures the concept 'satisfaction with the way democracy works'. Information about measurement quality can be used both to improve questionnaire design, by selecting the formulations and scales with a lower size of measurement errors (Revilla, Zavala-Rojas and Saris, 2016), and to correct for the remaining measurement errors after the data are collected (Saris and Revilla, 2016).

The main goal of this paper is to provide estimates of the measurement quality of the SWD indicator using ESS data for 7 different response scales and across 38 country-language groups[2]. By making this information easily accessible, our objectives are to make researchers aware of the presence of measurement errors and to provide specific insights about the best scales to use and their comparability across countries and languages. Besides, we also discuss the potential bias that measurement errors introduce and suggest considering correction for measurement errors.

[1] We use the term complement because measurement errors + measurement quality = 1. Complement refers to the associated counterpart.

[2] For this count, we considered country-language groups, that is, combinations of one country and one language (e.g., Spain-Spanish, Belgium-Dutch), but not the groups in round 1 of the ESS that combine one country and several languages (e.g., Spain-Spanish/Catalan and Belgium-Dutch/French).

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press

Table 1. Previous studies providing estimates of measurement quality for the SWD indicator

			Scale characteristics		Measurement quality	
Source	Country	Mode of data collection	No. answer categories	Labels	Lower estimate	Higher estimate
Revilla, 2010	The Netherlands	Face-to-face, telephone, web	11	Extremely dis/satisfied	.66	.85
			11	Very dis/satisfied	.55	.63
Revilla and Saris, 2013a	The Netherlands	Face-to-face, web	11	Extremely dis/satisfied	.67	.78
			11	Very dis/satisfied	.78	.85
			5	Strongly dis/agree	.57	.60
Revilla *et al.*, 2015	Spain	Face-to-face, web	11	Completely dis/satisfied	.75	.81
			11	Dis/satisfied	.83	.83
			5	Strongly dis/agree	.46	.65
Revilla and Ochoa, 2015	Mexico, Colombia	Web	11	Completely dis/satisfied	.78	.88
			11	Dis/satisfied	.70	.80
			5	Strongly dis/agree	.44	.57
DeCastellarnau and Revilla, 2017	Norway	Web	11	Extremely dis/satisfied	.85	.89
			11	Very dis/satisfied	.63	.63
			5	Very satisfied - Not satisfied at all	.74	.74

Note: Measurement quality ranges from 0 (only measurement errors) to 1 (no measurement errors; for more information, see Section 4.1). The table shows higher and lower estimates across modes (Revilla, 2010; Revilla and Saris, 2013a; Revilla et. al, 2015), different timing (DeCastellarnau and Revilla, 2017), or countries (Revilla and Ochoa, 2015).

Background

Evidence from previous literature

Previous research provides some estimates of the measurement quality of the SWD indicator under different conditions. Table 1 summarizes the existing knowledge.

Overall, the measurement quality ranges from .44 (in Colombia, 5-point 'Strongly dis/agree' scale) to .89 (in Norway, 11-point extremely 'Dis/satisfied' scale). This means that between 44% and 89% of the variance of the observed survey responses is due to variations in the latent trait 'satisfaction with the way democracy works', whereas between 11% and 56% come from measurement errors (for details, see Section 4.1 and 4.2). In general, the 11-point-item-specific scales yield a higher quality than the 5-point 'Strongly dis/agree' scales, although differences exist across studies.

Overall, this previous research confirms that: 1) the measurement quality of the SWD indicator is far from being perfect. Said differently, the SWD indicator does not measure perfectly the simple concept 'satisfaction with the way democracy works' and 2) the measurement quality of the SWD indicator varies across scales, modes of data collection and countries. Thus, measurement quality needs to be estimated under different conditions, to provide information allowing to select the best scales possible, assess comparability across studies or groups, and/or correct for measurement errors.

Determinants of measurement quality

In order to understand the reasons behind the variations in measurement quality observed in previous research about the measurement quality of the SWD indicator, we use the list of characteristics expected to affect measurement quality proposed by Saris and Gallhofer (2007). This list includes formal, topic-based, linguistic, layout and mode of data collection characteristics.

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press

In this paper, we focus on differences in measurement quality for the SWD indicator across response scales, countries and languages with the mode of data collection (face to face using showcards) being fixed.

On the one hand, previous research has found that scales' characteristics affect measurement quality (for an overview, see DeCastellarnau, 2018). In particular, item-specific scales have been found of higher quality than dis/agree scales (Saris et al, 2010). Possible explanations include that dis/agree scales are prone to acquiescence bias and that the response process is more complex for such scales (one extra cognitive step, see Saris *et al.*, 2010). Furthermore, scales with at least two fixed reference points have been found to be of higher quality (Revilla and Ochoa, 2015). A fixed reference point is a response option that all respondents understand without doubt in the same way, such as 'completely satisfied' (DeCastellarnau, 2018). They arguably increase quality by making the understanding of the scale clearer and unequivocal. Additionally, scales with a higher number of answer categories (up to a certain level) are argued to have higher quality, although the evidence is mixed (DeCastellarnau, 2018). However, many scales' characteristics and their possible interactions are currently understudied.

On the other hand, previous research has found differences in measurement quality across countries (e.g., Saris *et al.*, 2010, Revilla and Ochoa, 2015). There are mainly three types of characteristics proposed by Saris and Gallhofer (2007) that are expected to vary across countries and thus might lead to cross-national variations in measurement quality (Bosch and Revilla, 2021): 1) social desirability: if a topic is considered as more sensitive in a given country, the tendency of respondents to select answers that are more socially accepted could be higher, leading to a lower measurement quality; 2) centrality (or saliency) of the topic in respondents' minds: if a topic is less central in a given country, respondents are likely to have less formed or consistent opinions, leading to a lower measurement quality; and 3) linguistic characteristics, because languages have different inherent structures (Zavala-Rojas, 2016). For instance, a given question may be longer or more complex to understand in one language compared to another, even when following the highest translation standards, leading to a lower measurement quality. Languages can lead to different qualities across and within countries.

Implications

The variations in measurement quality observed in previous research about the measurement quality of the SWD indicator have some practical implications. In particular, they can lead to different results. To illustrate this point, we use data from the ESS Round 4 (UK), where the same respondents (n = 725) answered the SWD indicator twice: once at the beginning and once at the end of the survey. The wording of the question was the same in both cases: 'And on the whole, how satisfied are you with the way democracy works in the UK?' In both cases, an 11-point scale was used. However, the labels of the endpoints changed: 'Extremely dis/satisfied' (fixed reference points) versus 'Dis/satisfied' (not fixed reference points).

The cross-distribution of the answers (see Table 2) shows that only 33% of the respondents selected the same numerical option in both scales. Moreover, with the first scale, 44% of respondents are classified as 'dissatisfied' (answers 0 to 4), 18% as 'neither dissatisfied nor satisfied' (answer 5), and 38% as 'satisfied' (answers 6 to 10), whereas with the second scale these proportions are, respectively, 35%, 22%, and 43%. Hence, the first scale gives a more negative view of the satisfaction with the way democracy works of the same sample. Additionally, 8% of the respondents are classified as 'satisfied' with one scale, but 'dissatisfied' with the other and 23% are classified as 'neither satisfied nor dissatisfied' with one scale, but 'dis/satisfied' with the other. Similarly, correlations with other questions vary depending on the scale used (see Online Appendix 1). Thus, results for multivariate statistical analyses are also expected to change.

This illustrates that using different scales can produce different results. Population inferences based on the exact same sample may change depending on which scale was used. This is linked to

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press

Table 2. Contingency table of responses with both scales (UK, Round 4, same respondents answer with both scales). Question: *And on the whole, how satisfied are you with the way democracy works in the UK?*

Methods		Method at the end											
		Dissatisfied										Satisfied	Percent
	Options	0	1	2	3	4	5	6	7	8	9	10	column
Method at beginning	Extremely dissatisfied – 0	24	7	6	1	4	1	2	0	0	0	0	
	1	2	3	13	8	2	6	1	0	1	0	0	
	2	3	4	11	18	11	9	2	1	1	1	0	44%
	3	1	3	9	18	21	22	6	2	0	0	0	
	4	0	1	0	19	20	37	12	6	2	1	0	
	5	0	1	2	7	15	59	26	11	5	1	0	18%
	6	0	0	2	5	4	10	22	19	5	1	0	
	7	1	0	0	1	1	11	22	42	17	2	0	
	8	0	1	0	0	1	4	10	21	30	5	3	38%
	9	0	0	0	0	1	2	0	5	14	4	1	
	Extremely satisfied – 10	1	0	0	0	0	0	0	0	2	2	4	
	Percent row		35%				22%			43%			100%

Note: Light grey cells: Combinations of answers in which respondents are classified as 'dissatisfied' (options 0–4) with one scale, but 'satisfied' (options 6–10) with the other. This represents 8% of respondents.
Dark grey cells: Combinations in which respondents select the same numerical option with both scales. This represents 33% of respondents. Percent row and percent column: Percentages shown correspond to the grouping of satisfied (0–4), dissatisfied (6–10), and neutral (5) options with the method asked at the end (percent row) and at the beginning (percent column). Differences across dissatisfied and neutral are statistically significant ($P < .05$), non-significant for 'satisfied' ($P = 0.06$). A Smirnov–Kolmorogov test also confirms that the distributions' differences are statistically significant ($P < 0.01$).

the sizes of measurement errors and could provoke differences across studies. However, the true distributions or correlations are unknown and cannot be inferred only with this information. In order to determine which method is better, we need to estimate the measurement quality for each scale.

Contribution

Even though there is some research about its content validity, substantive literature has not addressed the size of measurement errors of the SWD indicator. However, several methodological papers provide estimates of the measurement quality of the SWD indicator (Section 2.1). Their results suggest that measurement errors can be large and vary across response scales, countries/languages and modes of data collection. However, this previous research suffers from several limitations. First, estimates are only available for a few scales and countries. Second, the estimates are provided in a way that does not lead to specific insights for substantive researchers. Third, the estimates differ across studies, but the reasons behind these variations are unclear. For example, the same scale (11-point very 'Dis/satisfied') yields the highest quality (around .82) in the study of Revilla and Saris (2013a) but generally the lowest one (around .58[3]) in an earlier study of Revilla (2010), even if both studies took place in the same country. This difference may be related to differences in the survey modes or in the model specification, although the reasons are not clear.

Thus, the main goal of this paper is to provide estimates of the size of measurement errors for different scales, countries, and languages. In particular, we contribute to the scarce literature in the following ways.

First, compared to previous studies looking at the measurement errors of the SWD indicator, we use a much larger and richer amount of data (more countries and methods). Particularly, we

[3] In this study the quality of the same method is computed for different modes of data collection: .58 is the average across modes.

analyze three multitrait-multimethod (MTMM) experiments implemented in the ESS, providing estimates for 7 response scales and 38 country-language groups.

Second, we use a unique estimation method for all the MTMM analyses, whereas previous research has used different ones. This makes our estimates more easily comparable.

Third, the MTMM analyses are performed following the recently developed Estimation Using Pooled Data (EUPD) approach (Saris and Satorra, 2018) that reduces the estimation problems observed in the past (see Section 4.2) and hence is expected to provide more accurate results.

Fourth, previous estimates of the measurement quality of the SWD indicator are presented in papers in which its use was incidental and with a clear methodological focus (e.g., comparing modes of data collection). Thus, these estimates were not connected to the substantive literature and are difficult to find for applied researchers. In contrast, this paper makes estimates of the measurement quality of the SWD indicator easily available to applied researchers, with the aim of raising awareness regarding the presence of measurement errors in surveys and their implications for substantive research.

Finally, these estimates are useful for several reasons: 1) they allow selecting the best instruments for future surveys, since they indicate how well different instruments measure the same concept; 2) they inform about the comparability of the indicator across groups (e.g., countries and languages). Indeed, standardized relationships can only be directly compared across groups if the measurement quality is the same in these groups; 3) they can help to disentangle which differences in results between studies/countries/languages may come from measurement errors; and 4) they are needed to correct for remaining measurement errors in applied research.

Method and data

Measurement quality

Measurement quality (q_{ij}^2) is defined as the strength of the relationship between the latent concept one wants to measure and the observed responses to a specific survey question asked to measure this latent concept (Saris and Andrews, 1991). It represents the proportion of the variance in the observed responses explained by the variance in the underlying latent concept of interest. It ranges from 0 (no relationship between the indicator and the latent concept) to 1 (perfect measurement). Measurement errors are defined as 1-q_{ij}^2. Following DeCastellarnau and Revilla (2017), we consider that the quality is 'excellent' if $q^2 \geq .9$; 'good' if $.9 > q^2 \geq .8$; 'acceptable' if $.8 > q^2 \geq .7$; 'questionable' if $.7 > q^2 \geq .6$; 'poor' if $.6 > q^2 \geq .5$; and 'unacceptable' if $q^2 < .5$.

In order to estimate measurement quality, we use the True Score model (Saris and Andrews, 1991). In Online Appendix 2, Figure 1 represents this model for the concept 'satisfaction with the way democracy works'.

Alternatively, the model can be summarized by the following system of equations:

$$\mathrm{T_{ij} = v_{ij}F_i + m_{ij}M_j} \tag{1}$$

$$\mathrm{Y_{ij} = r_{ij}T_{ij} + e_{ij}} \tag{2}$$

where F_i is the i^{th} trait (e.g., the concept 'satisfaction with the way democracy works'), M_j is the j^{th} method (each of the response scales), T_{ij} is the True Score (i.e., the hypothetical response of a person in a given scale corrected for random errors), and Y_{ij} is the observed response (i.e., the answer actually selected). When standardized, v_{ij}, m_{ij}, and r_{ij} are respectively the validity, method and reliability coefficients. The validity (square of the validity coefficient; ${v_{ij}}^2$) measures the strength of the relationship between the trait and the True Score. The method effects represent respondents' systematic reaction to a given method and are the complement of the validity (${m_{ij}}^2 = 1\text{-}{v_{ij}}^2$). The reliability (square of the reliability coefficient; ${r_{ij}}^2$) measures the strength of the relationship between the True Score and the observed responses. Finally, e_{ij} represents the random errors (e.g., selecting the wrong option by accident or interviewers' errors in recording

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press

the answer). Measurement quality can be computed as the product of reliability and validity: $q_{ij}^2 = r_{ij}^2 * v_{ij}^2$

This model (from now on 'Base Model') assumes that: a) random errors are uncorrelated with each other or with the trait and method factors; b) the traits are correlated; c) the method factors are uncorrelated between them or with the traits; and d) the impact of the method factor on the traits measured with a common scale is the same.

In order to estimate this model, structural equation modelling (SEM) is used (concretely, confirmatory factor analysis). Specifically, in order to identify the model in a SEM framework, it is necessary to consider several traits, each one measured using several methods. It is also possible to test the fit of the model, and some of the assumptions of the Base Model can be relaxed in order to improve this fit, leading to a Final Model from which the estimates are collected.

MTMM approach

The idea of repeating several traits, each measured with several methods, comes from Campbell and Fiske (1959), who were the first to introduce the MTMM approach. In this approach, researchers look at the correlations between the observed answers of questions asking for several (usually at least three) correlated traits using several (usually also at least three) methods. Andrews (1984) proposed to analyze such matrices of correlations (sometimes called MTMM matrices) through SEM. Saris and Andrews (1991) proposed to do this more specifically using the True Score model to reproduce the matrix of correlations.

However, asking at least three times the same questions to the same respondents using different methods increases respondent burden and could generate memory effects (Van Meurs and Saris, 1990). To reduce such problems, Saris, Satorra and Coenders (2004) proposed to randomly divide the respondents into different groups, each group answering a different combination of only two methods. Nevertheless, this Split-Ballot MTMM (SB-MTMM) approach frequently led to estimation problems (Revilla and Saris, 2013b). Thus, Saris and Satorra (2018) proposed, when similar datasets are available, to estimate a Pooled Data Model (PDM) with all the datasets, store its estimates, and use them to get an identified model in each dataset (here the country(-language) groups are the unit of interest). The rationale of this EUPD approach is that higher sample sizes lead to lower identification issues. The approach works under the assumption that especially trait effects, but also method effects, are expected to be quite similar across each group. We followed this approach since previous research suggests that it performs better than other alternatives, such as Bayesian SEM (Saris and Satorra, 2019) or estimation on a country-by-country basis (Revilla et al, 2020).

Data

We used data from three SB-MTMM experiments about Political Satisfaction implemented in the ESS rounds 1 (ESS Round 1: European Social Survey Round 1 Data, 2002; ESS Round 1: Test variables from Supplementary questionnaire, 2002), 2 (ESS Round 2: European Social Survey Round 2 Data, 2004; ESS Round 2: Test variables from Supplementary questionnaire, 2004), and 4 (ESS Round 4: European Social Survey Round 4 Data, 2008; ESS Round 4: Test variables from Supplementary questionnaire, 2008). The ESS is a biannual cross-national survey aimed at tracking the attitudes, opinions, and behaviours of citizens in most European countries.

A slightly different set of countries participated in each round. Thus, the number of countries analyzed are, respectively, 18 (R1), 22 (R2), and 27 (R4). Moreover, from R2 onwards, information about the language in which the survey was fielded is available. Therefore, whereas in R1 the analyses were done by country (sometimes with mixed languages, e.g., Switzerland), in R2 and R4 they were done by country-language group (e.g., Switzerland-French and Switzerland-German). However, languages with less than 70 observations in a given SB group were excluded. Thus,

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press

Table 3. Main characteristics of M1 and main differences of M2–M7 with respect to M1

Round	Method	Number of points	Labels of endpoints	Other characteristics
R1, R2, and R4	M1	11	Extremely dis/satisfied	Horizontal layout, three fixed reference points, medium correspondence numerical/verbal labels, bipolar
R1	M2	4	Very dis/satisfied	Fully labelled, vertical layout, no fixed reference point
	M3	6	Extremely dis/satisfied	No midpoint
R2	M4	11	Extremely dis/satisfied	Explicit midpoint
	M5	11	Very dis/satisfied	One fixed reference point
R4	M6	11	Dis/satisfied	One fixed reference point
	M7	5	Dis/agree strongly	Dis/agree scale, fully labelled, vertical layout

we analyzed 28 country-language groups in R2 and 33 in R4. For more information about the country(-language) groups and their sample sizes, we refer to Online Appendix 3.

In each round, the survey is implemented face to face and lasts around 1 hour. The main questionnaire consists in core modules repeated in each round and rotating modules addressing different topics. In the first seven rounds, it is followed by a supplementary questionnaire including a short version of the Schwartz Human Values scale and some repeated questions (usually varying the scale), part of several MTMM experiments.

In each round, the Political Satisfaction experiment asks about the same three traits: satisfaction with the present state of the economy, the way the government is doing its job, and the way democracy works. The requests for an answer for these traits are:

- Trait 1: On the whole how satisfied are you with the present state of the economy in [country]?
- Trait 2: Now thinking about the [country] government, how satisfied are you with the way it is doing its job?
- Trait 3: And on the whole, how satisfied are you with the way democracy works in [country]?

Moreover, three methods (i.e., response scales) are used in each round. One method (M1) is asked in the main questionnaire and is the same in the three rounds. It is a bipolar, item-specific, 11-point scale, with three fixed reference points, two verbal labels at the extremes (extremely dis/satisfied), horizontal layout and medium correspondence between verbal and numerical labels. The other two methods are asked in the supplementary questionnaire and differ across rounds. Table 3 presents the main characteristics of M1 and summarizes the main differences of the other methods with respect to M1. Showcards of all methods are available in Online Appendix 4.

Due to the SB design, respondents in each round get the method from the main questionnaire (M1) and then are randomly assigned to one of the two methods from the supplementary questionnaire.

Analyses and testing

The analyses are done for each round separately. First, for each SB group within a country(-language) group, the correlation matrices, standard deviations, and means were created with R 3.6.1 (R Core Team, 2019) using pairwise deletion. We excluded the individuals who did not answer the supplementary questionnaire during the same day because this has an impact on answers' quality (Oberski, Saris and Hagenaars, 2007), as well as a few individuals who did not follow the experimental procedure. Then, we used these matrices to create the pooled data matrices/standard deviations/means, which correspond to the weighted average of the matrices/standard deviations/means of all country(-language) groups analyzed from the same round. The weights are

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press

the sample size of each SB group within each country(-language) group divided by the total sample size across all country(-language) groups for that SB group.

Second, the True Score PDM (Base Model described in Section 4.1) was estimated using Lisrel 8.72 (Jöreskog and Sörbom, 2005) multiple-group maximum likelihood estimation (examples of inputs in Online Appendix 5).

Third, we tested the fit of the Base Model for each round using the JRule software (van der Veld, Saris and Satorra, 2008), based on the procedure developed by Saris, Satorra and van der Veld (2009). This procedure has the advantages of 1) testing at the parameter level and not at the global level and 2) considering the statistical power. Besides the indications of JRule, deviations from the Base Model were decided based on theoretical grounds. Our expectation was that the reaction of respondents to a given scale (method effect) might differ for either SWD (because the government and the economy are more specific and connected between them than with the democracy) or satisfaction with the way the government is doing its job (since citizens have more control on government than on the economy or democracy; see Online Appendix 6 for final PDMs).

For the country(-language) group analyses, we started again by estimating the Base Model using multiple-group maximum likelihood. However, in this case, the value of the parameters of the trait and method effects were previously fixed to the PDM values for the same round. This model was corrected using JRule in each group until reaching a Final Model (see Online Appendix 7). The priority was freeing the parameters fixed to the PDM (different value of the parameters, but same model specification), but other changes were often required (mainly freeing other method effects that were fixed to 1 in the PDM).

Results

Measurement quality of the SWD indicator across response scales

Table 4 shows the average measurement quality and standard deviation of the SWD indicator for seven response scales across all the country-(language) groups included in a given round[4].

First, the measurement quality of M1 (main questionnaire's 11-point scale with labels 'Extremely dis/satisfied') is on average similar across the three rounds, even if different individuals and countries participated in each round. This quality can be qualified as 'acceptable': around 70% of the variance in the observed survey responses can be attributed to variations in the underlying concept of interest and around 30% to measurement errors.

Second, the average measurement quality clearly varies across response scales. The lowest quality is found for the 4-point scale with labels 'Very dis/satisfied' (M2), with $q^2 = .46$, meaning that on average only 46% of the variance in observed responses is due to variations in the underlying concept of interest, while 54% is due to measurement errors. This scale is the only one with 'unacceptable' quality (<.5). This is an important finding since most regular surveys (all the ones mentioned in Section 1, except the ESS) currently use 4-point scales for the SWD indicator, although generally different among them (e.g., different labels). This suggests that 4-point scales are not a good option. In contrast, the highest quality ($q^2 = .84$ for both, classified as 'good') is found for the 11-point scales with an explicit midpoint (M4) and with labels 'Very dis/satisfied' (M5).

Moreover, the measurement quality for the 4-point scale with labels 'Very dis/satisfied' (M2; $q^2 = .46$) is lower than for the 6-point scale with labels 'Extremely dis/satisfied' (M3; $q^2 = .60$), which is lower than for the 11-point scale with labels 'Extremely dis/satisfied' (M1; $q^2 \approx .70$) and for the rest

[4] We also computed the average for the subgroup of 15 countries analyzed in all 3 rounds (17 country-language groups for R4 and 16 country-language groups for R2, since we could not obtain a satisfactory solution for Greece). The changes were minimal (the quality for M1-R2 was .68, for M1-R4, .72, and for M4, .83; the rest did not change). Thus, observed differences in average quality across scales from different rounds cannot be attributed to different country(-language) groups participating in each round.

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press

Table 4. Measurement quality (q^2) of the SWD indicator for seven response scales: average and standard deviation across country(-language) groups

Response scale	Number of points	Labels of endpoints	Other characteristics	Round	Average measurement quality (q^2)	Standard deviation
M1*	11	Extremely dis/satisfied	Horizontal layout, three fixed reference points, medium correspondence numerical/ verbal labels, bipolar	R1 R2 R4	.74 .69 .70	.06 .07 .09
M2	4	Very dis/satisfied	Fully labelled, vertical layout, no fixed reference points	R1	.46	.10
M3	6	Extremely dis/satisfied	No midpoint	R1	.60	.08
M4	11	Extremely dis/satisfied	Explicit midpoint	R2	.84	.05
M5	11	Very dis/satisfied	One fixed reference point	R2	.84	.08
M6	11	Dis/satisfied	One fixed reference point	R4	.74	.09
M7	5	Dis/agree strongly	Fully labelled, vertical layout	R4	.51	.12

*The same estimates for M1 are also presented in Poses *et al.* (2021).

of the 11-point scales (M4, M5, and M6; respectively .84, .84 and .74). This suggests that using more answer categories (up to 11) reduces measurement errors. Also, the 5-point 'dis/agree strongly' scale (M7) displays the second worst quality (.51, classified as 'poor'), consistent with previous research on the low quality of dis/agree scales (Saris *et al.*, 2010). Lastly, previous research suggests that using at least two fixed reference points is preferable, but our results do not support this: for instance, the 11-point scale with labels 'Very dis/satisfied' (M5; one fixed reference point) has a higher measurement quality than the one with three fixed reference points (M1).

Measurement quality across country(-language) groups

Besides variations across methods, our data indicate that, for a given method, there are variations in the estimated quality across country(-language) groups. This could be due to systematic or random fluctuations in the estimation. To further study them, Table 5 shows, for each country(-language) group, the average measurement quality and its standard deviation across methods. The country(-language) groups are divided according to the rounds in which they participated. They should only be compared with other groups participating in the same rounds (i.e., groups that received the same methods[5]).

Overall, one country-language group has an average measurement quality that can be classified as 'good' (Bulgaria, R4), 17 as 'acceptable', 22 as 'questionable', and 4 as 'poor'. Differences across groups are influenced by the rounds in which they participated (which determine the methods received). Average quality is .60 for R1, .79 for R2, and .65 for R4. Despite that, Table 5 suggests that some differences across countries do exist. This is further supported by the fact that in the analyses, in the Base Model, some parameters that were initially constrained to equality in all country(-language) groups had to be freed to obtain an acceptable fit. Additionally, the overall standard deviation of the methods ranges from .05 to .12, suggesting that systematic differences across country(-language) groups may be more pronounced for some methods.

Nevertheless, comparing the countries which participated in all three rounds, the average measurement quality of the SWD indicator varies from .63 in Denmark to .71 in Germany. In 14 out of 15 countries, average qualities fall within the interval .67–.71. Comparing countries that participated only in R2 and R4, differences tend to be larger (.13 difference between the higher and lower estimates). Comparing countries that participated only in R4, differences are even larger

[5] For any of these comparisons, 95% confidence intervals are always overlapping. This is not unexpected considering the small sample of methods within each country.

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press

Table 5. Measurement quality (q^2) across country(-language) groups: average and standard deviation across methods

Round	Country(-language) group	Average measurement quality (q^2)	Standard deviation
R1, R2, and R4	Germany	.71	.10
	Great Britain	.70	.16
	Norway	.69	.13
	Portugal	.69	.18
	Belgium*	.69	.18
	Spain**	.69	.15
	Finland**	.68	.16
	Greece	.68	.13
	France	.68	.11
	Czech Republic	.68	.14
	Slovenia	.68	.15
	Switzerland*	.67	.18
	The Netherlands	.67	.16
	Poland	.67	.14
	Denmark	.63	.17
R2 and R4	Estonia Russian	.79	.14
	Estonia Estonian	.77	.10
	Slovakia	.75	.11
	Belgium French	.74	.13
	Switzerland German	.73	.09
	Belgium Dutch	.70	.23
	Turkey	.70	.21
	Ukraine Ukrainian	.69	.20
	Switzerland French	.68	.14
	Ukraine Russian	.66	.25
R1 and R4	Sweden	.64	.15
	Israel*	.64	.11
R1 and R2	Austria	.72	.15
	Ireland	.71	.10
R4	Bulgaria	.86	.09
	Latvia Russian	.72	.09
	Latvia Latvian	.72	.11
	Cyprus	.71	.19
	Israel Hebrew	.68	.15
	Russia	.64	.10
	Israel Arabian	.63	.07
	Croatia	.54	.11
	Romania	.51	.07
R2	Luxembourg Luxembourgish	.78	.13
	Luxembourg French	.73	.14
	Italy	.71	.18
R1	Israel-Mixed	.62	.11
	Belgium-Mixed	.59	.15
	Switzerland-Mixed	.53	.33
	Overall	.69	.15

*Belgium, Switzerland, and Israel participated in R1 but could only be split by language for R2 and/or R4. Hence, we included them twice: considering their average across languages/rounds in which they participated, and by separate languages.
**Spain included Catalan and Spanish in R1; Finland included Finnish and Swedish in R1.

(.25 difference between the higher and lower estimates). Higher differences may be related to the fact that less methods (and different combinations) are included in the average.

In many cases, we cannot separate country from language effects. However, there are seven countries where different language groups were analyzed. In these cases, differences range from 0 within Latvia languages to .05 within Israel, Switzerland, and Luxembourg languages. Additionally, two country-language groups with the same language (Ukraine-Russian and Estonia-Russian) have, respectively, the maximum (.79) and minimum (.66) qualities for the

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press

groups of countries which participated in R2 and R4, suggesting that country-specific characteristics are more important than the use of different languages in explaining quality differences across groups. Finally, each country-language group presents standard deviation across methods oscillating around the average standard deviation of .15. These results suggest that variations in quality due to the method occur in all countries, while variations in quality for a given method due to country(-language) characteristics are less noticeable.

Overview of all estimates and how to use them

Table 6 presents the full list of estimates of measurement quality for the SWD indicator. These estimates can be used for different purposes. First, they can be used to select the best methods to measure 'satisfaction with the way democracy works' in future surveys. Particularly, the results (see also Table 4) indicate that in the majority of countries, an 11-point scale with explicit midpoint (M4) and/or an 11-point scale with labels 'Very dis/satisfied' (M5) are the best options. Nevertheless, since the scale with higher quality may depend on the country-(language) group(s) of interest (and the methods analyzed for each country), researchers can tailor this general recommendation to the specific countries of their interest using Table 6. For instance, the best option seems to be an 11-point scale with an explicit midpoint (M4) in Finland, but an 11-point scale with labels 'Dis/satisfied' (M6) in France. For cross-national surveys, the scale that is the best in most countries of interest can be selected.

Second, a necessary condition for comparing standardized relationships between satisfaction with the way democracy works and other variables across groups is to have a similar measurement quality in each group. Readers can compare the measurement quality of the groups they are interested in to assess if this condition is met for different methods. For instance, since the quality for M1 (11-point, labels 'Extremely dis/satisfied') is .52 in Romania but .90 in Bulgaria, our results suggest that, without correction, one cannot compare standardized relationships (e.g., correlations and standardized regression coefficients) between the SWD indicator (M1) and another variable across Romania and Bulgaria.

Third, these estimates can help to disentangle which differences in results between studies/countries/languages may come from measurement errors. In general, the lower the estimate of measurement quality estimate, the lower the observed correlation compared to the real one, unless there is common method variance (Saris and Revilla, 2016). To illustrate this point, let assume that in the Netherlands the observed correlation between the SWD indicator and another variable (measured without errors) was .60 in a study that used M1 (11-point scale with labels 'Extremely dis/satisfied') for the SWD indicator, but .44 in another study that used M2 (4-point scale with labels 'Very dis/satisfied'). While the results of both studies may seem inconsistent, once we take into account the difference in qualities, the corrected correlation would be the same (.7) in both cases (see Online Appendix 8 for details). Generally, when there are large differences in measurement quality, we can expect that observed correlations will differ across studies even if the true correlations were in fact the same.

Finally, these estimates can also be used to perform correction for measurement errors. To illustrate this point, we replicated part of the regression analyses of a study by Vlachová (2019) about the determinants of the SWD indicator (Table 2, p. 233). We selected this example because it is based on ESS data, so the estimates of measurement quality presented in Table 6 can be used to do the correction in the case of the SWD indicator. However, such estimates are only available for R2. Thus, we simplified the example by focusing only on R2 data[6]. Moreover, to keep

[6] For the measurement quality of the variables 'satisfaction with the economy' and 'trust in parliament', we used estimates from Poses *et al.* (2021) for the ESS R2: such estimates are based on similar analyses as the ones in this paper. For the remaining sociodemographic and factual variables, we assume there are no measurement errors, since no measurement error estimates are available.

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press

Table 6. Measurement quality estimates (q^2) for each country(-language) group and method

	M1*	M2	M3	M4	M5	M6	M7
Country-language	(11-point, Extremely)	(4-point, Very)	(6-point, Extremely)	(11-point, explicit midpoint)	(11-point, Very)	(11-point, Dis/satisfied)	(5-point, AD)
Austria	.72	.48	.65	.86	.87		
Belgium-Dutch	.79			.85	.88	.56	.31
Belgium-French	.78			.88	.76	.78	.49
Belgium-Mixed (R1)	.73	.43	.60				
Bulgaria	.90					.92	.76
Switzerland-French	.67			.85	.80	.64	.45
Switzerland-German	.69			.71	.79	.88	.64
Switzerland-Mixed (R1)	.89	.24	.46				
Cyprus	.69					.90	.53
Czech Republic	.69	.66	.70	.76	.87	.72	.37
Germany	.72	.50	.64	.83	.83	.75	.65
Denmark	.62	.34	.57	.87	.88	.68	.46
Estonia-Estonian	.73			.88	.90	.69	.65
Estonia-Russian	.83			.89	.94	.72	.56
Spain	.68	.36	.61	.80	.81	.86	.72
Finland	.74	.40	.56	.90	.80	.75	.49
France	.72	.58	.62	.73	.75	.83	.48
Great Britain	.69	.45	.57	.92	.92	.78	.59
Greece	.76	.62	.70			.76	.44
Croatia	.57					.63	.42
Ireland	.60			.73	.80		
Israel-Arabian	.59					.71	.58
Israel-Hebrew	.70					.83	.52
Israel-Mixed	.73	.50	.62				
Italy	.50			.83	.80		
Luxembourg-French	.76			.85	.57		
Luxembourg-Luxembourgish	.64			.83	.88		
Latvia-Latvian	.78					.78	.60
Latvia-Russian	.80					.73	.62
The Netherlands	.73	.40	.52	.83	.85	.73	.48
Norway	.70	.43	.57	.85	.86	.72	.67
Poland	.70	.55	.60	.81	.88	.66	.44
Portugal	.79	.43	.49	.80	.88	.81	.46
Romania	.52					.57	.43
Russia	.67					.72	.53
Sweden	.67	.44	.61			.89	.54
Slovenia	.66	.46	.78	.86	.88	.63	.51
Slovakia	.70			.83	.90	.75	.60
Turkey	.66			.90	.94	.64	.37
Ukraine-Russian	.67			.83	.89	.73	.19
Ukraine-Ukrainian	.68			.87	.81	.78	.31
Total general	.71	.46	.60	.84	.84	.74	.51

*For M1, the table shows the average for all rounds in which a country participated (M1 appeared in R1, R2, and R4). These estimates for M1 can also be computed from the Online Appendix of Poses *et al.* (2021).

the example concise, we focus on two countries: Czech Republic (n = 3,026) and Slovakia (n = 1,512). We use the COSME package (Cimentada & Weber, 2020) in R 4.0.4 (R Core Team, 2021) to automatically[7] correct for measurement errors the correlation matrix on which the regression analyses are based (for more details, see: https://sociometricresearch.github.io/cosme/index.html and Online appendix 9) and the lavaan package (Rosseel, 2012) for estimation. For a more general explanation of how to correct for measurement errors in different models, we

[7] The user only needs to provide estimates of quality, validity, and reliability.

Table 7. Standardized coefficients of regressions (dependent variable: the SWD indicator), with 95 % confidence intervals in brackets

	Czech Republic		Slovakia	
	Not corrected	Corrected	Not corrected	Corrected
Satisfaction with the economy	**.399**	**.551**	**.425**	**.667**
	(.365, .433)	**(.530, .571)**	**(.379, .471)**	**(.648, .685)**
Trust in parliament	**.255**	**.270**	**.221**	**.181**
	(.220, .291)	**(.248, .292)**	**(.173, .270)**	**(.160, .201)**
Age	−.085	−.089	−.068	−.050
	(−.119, −.050)	(−.107, −.071)	(−.115, −.021)	(−.067, −.033)
Voted last elections	.020	.012	−.009	−.028
	(−.015, .055)	(−.006, .03)	(−.056, .038)	(−.045, −.011)
Income	.043	.013	.052	.002
	(.008, .077)	(−.05, .032)	(.004, .099)	(−.015, .020)
Woman	−.005	−.007	−.006	−.007
	(−.038, .028)	(−.024, .011)	(−.051, .040)	(−.024, .009)
Tertiary education	.016	.011	.039	.011
	(−.018, .049)	(−.007, .029)	(−.008, .085)	(−.006, .028)
R^2	.336	.589	.342	.637

Note: in bold, variables where we corrected for measurement errors. More information about the variables included in the regressions in Online appendix 10.

refer to Saris and Gallhofer (2007) and DeCastellarnau and Saris (2014). Table 7 presents the standardized regression coefficients with and without correction for measurement errors.

All estimates change once correction for measurement error is implemented. The change is, as expected, especially pronounced for the variables for which we correct for measurement errors directly. For instance, the standardized coefficient of the effect of satisfaction with the economy on the SWD indicator increases by around 30% in the Czech Republic and 50% in Slovakia. Besides, one of the conclusions of Vlachová (2019) is that ‘satisfaction with the present state of the economy is a stronger predictor of SWD than trust in parliament’ (p. 232). We can see that this conclusion still holds after correction for measurement errors. However, the difference is now much stronger. Furthermore, while the effect of satisfaction with the present state of the economy is quite similar in both countries before correction, once applying the correction, this effect is larger in Slovakia. Besides, some coefficients are statistically different from 0 with correction, but not without correction (e.g., income in Czech Republic). This illustrates that comparisons across countries might be affected as well. Finally, the R^2 sharply increases.

Discussion/conclusions

Main results

While there has been some debate about which concepts – beyond ‘satisfaction with the way democracy works’ – the SWD indicator measures, the size of the measurement errors of this indicator has been ignored in substantive literature. In this paper, we started to fill this gap by providing estimates of the measurement quality of the SWD indicator for 7 scales and 38 country(-language) groups using data from 3 MTMM experiments implemented in the ESS. Our results provide useful information for the choice of better scales in future surveys, help to check if the necessary condition for comparing standardized relationships (equal quality) across groups is met, help to disentangle differences in results due to measurement errors, and can be used to both assess the effect of measurement errors in a single study and correct for them.

Additionally, we found that the average measurement qualities vary systematically across response scales. On average, two 11-point scales (M4, with an explicit midpoint, and M5, with labels ‘Very dis/satisfied’) present the highest quality (.84) and the 4-point scale (M2, labels

'Very dis/satisfied') the worst (.46). The response scale from the ESS main questionnaire (M1) displayed an acceptable quality (around .70). All 11-point scales (M1, M4, M5, and M6) present a higher quality than the 4-point scale (M2, .46), the 6-point scale (M3, .60), and the 5-point dis/agree scale (M7, .51). The reason for the differences between the 11-point scales (M1, M4, M5, and M6), which differed only in their labels, is unclear. Further research is needed to disentangle this.

Moreover, we found that systematic differences across country-language groups are often (very) small. However, they are bigger in some cases (especially when less methods are included in the average). Most differences between languages are also small.

Limitations

First, not all methods were asked at the same time. Hence, differences in quality between methods in the main (M1) versus supplementary questionnaires (M2–M7) could be explained both by the timing and the variations in response scales, while differences between the methods of the supplementary questionnaires are not affected by the timing. Also, M2–M7 are asked as a repetition of the same question that would not occur in normal surveys and may affect respondents' answers (e.g., memory effects).

Second, confidence intervals of the quality estimates are not easily retrievable. Thus, it is difficult to know which differences between estimates are a product of estimation uncertainty (Oberski and Satorra, 2013). However, the results' consistency across groups and rounds and the large sample sizes may partially account for these problems, especially regarding average estimates across methods.

Third, there were still some problems of improper solutions and to a lesser extent non-convergence. Fourth, the testing procedure involves some non-avoidable subjectivity. The last two issues might affect the values of the estimates. Future research in the broader field of SEM shedding light on these problems would be desirable.

Finally, the results are obtained for a face-to-face survey using showcards. Further research that explores whether these results hold for different modes of data collection (e.g., telephone, web surveys), as well as including more scales/countries, is needed.

Practical implications

Based on our results, we derive some general guidelines/recommendations for the SWD indicator.

First, in general, we recommend using 11-point scales, particularly with an explicit midpoint (as M4), at least for face-to-face surveys. Currently, most regular surveys use different variations of 4-point scales for the SWD indicator. In our study, M2 is the best approximation for the quality of these scales because it also has four points. Based on our results, 4-point scales do not seem to be a good option: measurement errors explain more than half of the variance of the observed responses.

Second, comparing studies that use different methods, it is likely that differences in results can be due to differences in the size of measurement errors if these methods have different qualities. Particularly, differences in results between studies that use 4-point versus 11-point scales can be expected if no correction is implemented.

Third, differences in quality across country-language groups for the SWD indicator are on average small for many country-language groups. Thus, when comparing countries that use the same method, differences in results across countries are not very likely to be due to different sizes of measurement errors. However, this cannot be ruled out for all groups, especially for those countries/languages not analyzed here. Besides, this does not imply that these estimates are unbiased: that all are equally affected by the size of measurement errors allows comparing them but does not reduce the size of the bias.

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press

Lastly, these findings suggest, in line with previous research, that standardized relationships between different concepts based on survey measures may not be well estimated because of the presence of measurement errors, potentially affecting substantive results. They may be infra-estimated because of random errors or over-estimated because of the presence of common method variance. Researchers should correctly tackle this issue. Particularly, this situation could be improved by performing correction for measurement errors (Saris and Revilla, 2016).

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/S1755773921000266.

Acknowledgements. We are thankful to Wiebke Weber for her comments and support for this paper, to Jorge Cimentada for his help with the data preparation and analysis, to Hannah Schwarz for her help with the data preparation, and to Andrea Noferini for his support to the initial project. We also thank the ESS CST for their continuous support of this line of research.

Competing interests. The author(s) declare none.

References

Andrews, F.M. (1984), 'Construct validity and error components of survey measures: a structural modeling approach', *Public Opinion Quarterly* **48**(2): 409–442. doi: 10.1086/268840.

Bollen, K. (1989), *Structural Equations with Latent Variables*, New York: Wiley.

Bosch, O. and M. Revilla (2021), 'The quality of survey questions in Spain: a cross-national comparison', *Revista Española de Investigaciones Sociológicas*, **175**: 3–26. doi: 10.5477/cis/reis.175.3

Campbell, D.T. and D.W. Fiske (1959), 'Convergent and discriminant validation by the multitrait-multimethod matrix'. *Psychological Bulletin* **56**(2): 81–105. doi: 10.1037/h0046016.

Canache, D., J.J. Mondak, and M.A. Seligson (2001), 'Meaning and measurement in cross-national research on satisfaction with democracy', *Public Opinion Quarterly* **65**(4): 506–528. doi: 10.1086/323576.

Christmann, P. (2018), 'Economic performance, quality of democracy and satisfaction with democracy', *Electoral Studies*, **53**: 79–89. doi: 10.1016/j.electstud.2018.04.004

Cimentada, J. and W. Weber (2020), 'Cosme: A flexible tool to correct for survey measurement errors', sociometricresearch/cosme: Package release (Version v0.0.1). Zenodo. http://doi.org/10.5281/zenodo.4316599

Dassonneville, R. and I. McAllister (2020), 'The party choice set and satisfaction with democracy', *West European Politics* **43**(1): 49–73. Routledge. doi: 10.1080/01402382.2019.1609286.

DeCastellarnau, A. (2018). 'A classification of response scale characteristics that affect data quality: a literature review', *Quality and Quantity* **52**(4): 1523–1559. Springer Netherlands,. doi: 10.1007/s11135-017-0533-4.

DeCastellarnau, A., and M. Revilla (2017), 'Two approaches to evaluate measurement quality in online surveys: an application using the Norwegian citizen panel', *Survey Research Methods* **11**(4): 415–433. European Survey Research Association. doi: 10.18148/srm/2017.v11i4.7226.

DeCastellarnau, A. and W.E. Saris (2014), 'A simple way to correct for measurement errors', European Social Survey Education Net (ESS EduNet). http://essedunet.nsd.uib.no/cms/topics/measurement/.

Easton, D. (1965), *A System Analysis of Political Life*, New York: Wiley.

Easton, D. (1975). 'A re-assessment of the concept of political support', *British Journal of Political Science* **5**: 435–437.

ESS Round 1: European Social Survey Round 1 Data (2002), Data file edition 6.6. NSD - Data Archive and distributor of ESS data for ESS ERIC. http://doi.org/10.21338/NSD-ESS1-2002

ESS Round 1: Test variables from Supplementary questionnaire (2002), Data file edition 1.1. NSD - Data Archive and distributor of ESS data for ESS ERIC.

ESS Round 2: European Social Survey Round 2 Data (2004), Data file edition 3.6. NSD - Data Archive and distributor of ESS data for ESS ERIC. http://doi.org/10.21338/NSD-ESS2-2004

ESS Round 2: Test variables from Supplementary questionnaire (2004), Data file edition 3.2. NSD - Data Archive and distributor of ESS data for ESS ERIC.

ESS Round 4: European Social Survey Round 4 Data (2008), Data file edition 4.5. NSD - Data Archive and distributor of ESS data for ESS ERIC. http://doi.org/10.21338/NSD-ESS4-2008.

ESS Round 4: Test variables from Supplementary questionnaire (2008), Data file edition 1.0. NSD - Data Archive and distributor of ESS data for ESS ERIC.

Ferrin, M. (2016), An Empirical Assessment of Satisfaction with Democracy', in M. Ferrín and H. Kriesi (eds), *How Europeans View and Evaluate Democracy*, Great Britain: Oxford University Press, pp. 283–306.

Jöreskog, K. and D. Sörbom (version 8.72) (2005), *Lisrel 8*. Uppsala, Sweden: Scientific Software International.

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press

Linde, J. and J. Ekman (2003), 'Satisfaction with democracy: a note on a frequently used indicator in comparative politics', *European Journal of Political Research* **42**(3): 391–408. doi: 10.1111/1475-6765.00089.

Norris, P. (2011), 'The conceptual framework', in P. Norris (ed.), *Democratic Deficit: Critical Citizens Revisited*, United States of America: Cambridge University Press, pp. 19–37.

Oberski, D., W.E. Saris, and J. Hagenaars (2007), 'Why are there differences in the quality of questions across countries?', in G. Loosveldt, M. Swyngedouw, and B. Cambre (eds), *Measuring Meaningful Data in Social Research*, Leuven: Acco, pp. 281–299.

Oberski, D. and A. Satorra (2013), 'Measurement error models with uncertainty about the error variance', *Structural Equation Modeling: A Multidisciplinary Journal* **20**(3): 409–428. doi: 10.1080/10705511.2013.797820.

Poses, C., M. Revilla, M. Asensio, H. Schwarz, and W. Weber (2021), 'Measurement quality of 67 common social sciences questions across countries and languages based on 28 Multitrait-Multimethod experiments implemented in the European Social Survey', *Survey Research Methods*.

Quaranta, M. (2018), 'How citizens evaluate democracy: an assessment using the European social survey'. *European Political Science Review* **10**(2): 191–217. doi: 10.1017/S1755773917000054.

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

Revilla, M. (2010), 'Quality in unimode and mixed-mode designs: a multitrait-multimethod approach', *Survey Research Methods* **4**(3): 151–164. doi: 10.18148/srm/2010.v4i3.4278.

Revilla, M. and Ochoa, C. (2015), 'Quality of different scales in an online survey in Mexico and Colombia', *Journal of Politics in Latin America* **7**(3): 157–177.

Revilla, M., C. Poses, O. Serra, M. Asensio, H. Schwarz, and W. Weber (2020), 'Applying the estimation using pooled data approach to the multitrait-multimethod experiments of the European social survey (Rounds 1 to 7)', *Structural Equation Modeling: A Multidisciplinary Journal*, September, 1–12. doi: 10.1080/10705511.2020.1807988.

Revilla, M. and W.E. Saris (2013a), 'A comparison of the quality of questions in a face-to-face and a web survey', *International Journal of Public Opinion Research* **25**(2): 242–253. doi: 10.1093/ijpor/eds007.

Revilla, M. and W.E. Saris (2013b), 'The split-ballot multitrait-multimethod approach: implementation and problems', *Structural Equation Modeling* **20**(1): 27–46. doi: 10.1080/10705511.2013.742379.

Revilla, M., W.E. Saris, G. Loewe, and C. Ochoa (2015), 'Can a non-probabilistic online panel achieve question quality similar to that of the European social survey?', *International Journal of Market Research* **57**(3): 395–412. doi: 10.2501/IJMR-2015-034.

Revilla, M., D. Zavala-Rojas, and W.E. Saris (2016), 'Creating a good question: how to use cumulative experience', in (eds), C. Wolf, D. Joye, T.W. Smith, and Y. Yang-chih Fu, *The SAGE-Handbook of Survey Methodology*, pp. 236–254. United Kingdom: Sage.

Rosseel, Y. (2012), 'Lavaan: an R package for structural equation modeling and more', *Journal of Statistical Software*, **48**(2): 1–36.

Saris, W.E. and F. Andrews (1991), 'Evaluation of measurement instruments using a structural modeling approach', in P.P. Biemer, R.M Groves, L.E. Lyber, N.A. Mathiowetz, and S. Sudman (eds), *Measurement Errors in Surveys*, pp. 575–598, New York: John Wiley & Sons, Inc.

Saris, W.E. and I. Gallhofer (2007), *Design, evaluation, and analysis of questionnaires for survey research*. New York: Wiley.

Saris, W.E. and M. Revilla (2016), 'Correction for measurement errors in survey research: necessary and possible', *Social Indicators Research* **127**(3): 1005–1020. Springer Netherlands. doi: 10.1007/s11205-015-1002-x.

Saris, W.E., M. Revilla, J.A. Krosnick, and E.M. Shaeffer (2010), 'Comparing questions with agree/disagree response options to questions with item-specific response options', *Survey Research Methods* **4**(1): 61–79.

Saris, W.E. and A. Satorra (2018), 'The pooled data approach for the estimation of split-ballot multitrait–multimethod experiments', *Structural Equation Modeling* **25**(5), 659–672. doi: 10.1080/10705511.2018.1431543.

Saris, W.E. and A. Satorra (2019), 'Comparing BSEM and EUPD estimates for two-group SB-MTMM experiments', *Structural Equation Modeling* **26**(5): 745–749. Routledge. doi: 10.1080/10705511.2019.1576046.

Saris, W.E., A. Satorra, and G. Coenders (2004), 'A new approach to evaluating the quality of measurement instruments: the split-ballot MTMM design', *Sociological Methodology*, 311–347. doi: 10.1111/j.0081-1750.2004.00155.x.

Saris, W.E., A. Satorra, and W. van der Veld (2009), 'Testing structural equation models or detection of misspecifications?', *Structural Equation Modeling: A Multidisciplinary Journal* **16**(4): 561–582. doi: 10.1080/10705510903203433.

Schäfer, A. (2012), 'Consequences of social inequality for democracy in Western Europe', *Zeitschrift Für Vergleichende Politikwissenschaft* **6**(S2): 23–45. doi: 10.1007/s12286-010-0086-6

Thomassen, J. and C. van Ham (2017), 'A Legitimacy Crisis of Representative Democracy?', in C. van Ham, J. Thomassen, K. Aars, and R. Andeweg (eds), *Myth and Reality of the Legitimacy Crisis: Explaining Trends and Cross-National Differences in Established Democracies*, New York: Oxford University Press.

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press

van der Meer, T. and A. Hakhverdian (2017), 'Political trust as the evaluation of process and performance: a cross-national study of 42 European countries', *Political Studies* **65**(1): 81–102. doi:10.1177/0032321715607514

Van der Veld, W., W.E. Saris, and A. Satorra (Version 3.0.4 Beta) (2008), Judgement Rule Aid for Structural Equation Models.

van Ham, C. and J. Thomassen (2017), 'The myth of legitimacy decline', in C. van Ham, J. Thomassen, K. Aars, and R. Andeweg (eds), *Myth and Reality of the Legitimacy Crisis: Explaining Trends and Cross-National Differences in Established Democracies*, New York: Oxford University Press. doi: 10.1093/oso/9780198793717.003.0002.

Van Meurs, A.V. and W.E. Saris (1990), 'Memory effects in MTMM studies', in A.V. Van Meurs and W.E. Saris, *Evaluation of Measurement Instruments by Meta-Analysis of Multitraitmultimethod Studies*, Amsterdam: North Holland, pp. 134–146.

Vlachová, K. (2019), 'Lost in transition, found in recession? Satisfaction with democracy in Central Europe before and after economic crises', *Communist and Post-Communist Studies* **52**(3): 227–234. doi: 10.1016/j.postcomstud.2019.07.007

Zavala-Rojas, D. (2016), 'Measurement Equivalence in Multilingual Comparative Survey Research,' PhD Thesis. Barcelona: Universitat Pompeu Fabra.

Cite this article: Poses C and Revilla M (2022). Measuring satisfaction with democracy: how good are different scales across countries and languages?. *European Political Science Review* **14**, 18–35. https://doi.org/10.1017/S1755773921000266

https://doi.org/10.1017/S1755773921000266 Published online by Cambridge University Press