Estimating the size of measurement errors of the Satisfaction With Democracy survey indicator for different scales, countries and languages

Carlos Poses
RECSM-UPF
carlos.gonzalezp@upf.edu

Melanie Revilla
RECSM-UPF
melanie.revilla@upf.edu

RECSM Working Paper Number 61

December 2020

Abstract

The Satisfaction With Democracy (SWD) indicator is often used in 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) is scarce. Nonetheless, measurement errors can affect the results and threaten comparisons across studies/countries/languages. Thus, in this paper, we estimated the measurement quality (complement of measurement errors) of the SWD indicator for seven response scales across 38 country-language groups, using 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. We also provide insights to improve questionnaires and evaluate the indicator's comparability across scales/countries/languages.

Keywords: Satisfaction with democracy (SWD), measurement errors, measurement quality, European Social Survey (ESS), multitrait-multimethod (MTMM) experiments, social indicators

Acknowledgements: We are thankful to Wiebke Weber for her comments and support for this paper, to Jorge Cimentada and Hannah Schwarz for their help with 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.

1 Introduction

Democracy is one of the most long-standing, fundamental topics within social sciences. Despite the concept of democracy itself being highly contested, both at the philosophical/normative level and regarding its empirical definitions (e.g., Coppedge *et al.*, 2011; Norris, 2011, Prezworksi, 1999), researchers have studied citizens' attitudes, opinions and orientations towards democracy (e.g., Almond and Verba, 1963; Linz and Stepan, 1996).

Political support (Easton, 1965; 1975) is considered key for the evolution of democracies. Scholars argue that democracies need its citizens' support to persist (Canache, Mondak and Seligson, 2001; Linde and Ekman, 2003) or that the legitimacy of democracies among its citizens is essential for citizens to abide to authorities' ruling (Thomassen and van Ham, 2017). Political support is a complex multi-dimensional concept. Originally, Easton distinguished between the concepts of 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). Later, other scholars developed more refined models of political support drawing on this conceptualization. Currently, the Norris' (Norris, 2011) five-fold model is the most common (Linde and Ekman, 2003; Norris, 2011; van Ham and Thomassen, 2017).

Many empirical studies of political support concentrated on a specific survey indicator: the Satisfaction With Democracy (SWD) indicator. Sometimes referred to as a measure of the third level in Norris' five-fold model of political support (Norris, 2011, 28; van Ham and Thomassen, 2017, 3), this indicator asks respondents how satisfied they are with the way democracy works in their country. This indicator is the focus of this paper.

The SWD indicator has been regularly included in major academic surveys (among others: Afrobarometer, Asian Barometer, Americas Barometer, Comparative Survey of Electoral Studies, Eurobarometer, European Social Survey, European Values Study, Latinobarometer), giving rise to a

huge literature about its determinants. In an updated overview, Dassonneville and McAllister (2020) classify the political determinants of the SWD indicator in those focusing on the role of political institutions (electoral systems effects, incumbency, turnout or the set of party choices) and the effects of regime performance (quality of democratic governance, corruption, economic performance or social policy). Additionally, the relationships of the SWD indicator with non-strictly political factors such as life satisfaction (Stadelmann-Steffen and Vatter, 2012) or the lockdown due to the COVID-19 pandemic (Bol *et al.*, 2020) were studied. However, some of the measurement properties of the SWD indicator are unclear.

First, scholars used the SWD indicator to measure different theoretical concepts (Ferrín, 2016; Quaranta, 2018). However, the extent to which the SWD indicator really measures these concepts (i.e. its content validity; Bollen, 1989, 185-186) has been an important source of discussion (Canache, Mondak and Seligson, 2001; Linde and Ekman, 2003; Ferrín, 2016).

Second, focusing on the simple concept "satisfaction with the way democracy works", the SWD indicator does not perfectly measure it. Slightly different requests for an answer (e.g. "Are you satisfied or dissatisfied with the way democracy works?" or "Please tell me how satisfied you are with the way democracy works") and scales (e.g. "a 4-point scale from "Very dissatisfied" to "Very satisfied", or a 11-point scale from "0-Extremely dissatisfied" to "10-Extremely satisfied") can be (and have been) used for the SWD indicator (see Section 5.2). However, each of these methods has its own level of measurement errors. Indeed, all measurements contain errors (Saris and Galhofer, 2007); especially survey questions measuring abstract, subjective concepts such as "satisfaction with the way democracy works".

Yet, some methods measure better the underlying concepts of interest than others. Besides, not accounting for these errors can affect the results (Saris and Revilla, 2016). Thus, it is important to estimate the size of measurement errors for different instruments and under different conditions (e.g.

across time, countries or languages). Instead of estimating the size of measurement errors, researchers often estimate their complement[1]: measurement quality (Saris and Gallhofer, 2007). Measurement quality is a statistical measure defined as the strength of the relationship between the latent concept of interest and the observed survey answers (see Section 4.1). It can be estimated using data from multitrait-multimethod (MTMM) experiments, which consist in repeating several questions using different methods (often different scales; see Section 4.2). Information about measurement quality can be used both to improve questionnaire design, by selecting the formulations and scales with lower size of measurement errors (Revilla, Zavala-Rojas and Saris, 2016), and to correct for the remaining measurement errors after the data is collected (Saris and Revilla, 2016).

To the best of our knowledge, there is no substantive paper regarding the SWD indicator that addresses or corrects for measurement errors. This has several implications: First, some of the substantive findings regarding the SWD indicator may be a by-product 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. Third, the questions' formulations and/or scales currently used might not be the ones with smaller size of measurement errors. Thus, the main goal of this paper is to start filling this gap by estimating the measurement quality of the SWD indicator for seven different response scales and across 38 country-language groups [2].

[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 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).

2 Background

2.1 What does the SWD indicator measure?

There is some debate about which theoretical concept the SWD indicator measures (i.e., its content validity). Canache, Mondak and Seligson (2001, 525) claimed that "for any given observation—be it individual-level or aggregate-level—we simply do not know what SWD measures". Contrarily, Linde and Ekman (2003, 405) argued that the indicator "taps the level of support for how the democratic regime works in practice" and "is far from a perfect indicator of support for the performance of a democratic regime". More recently, Norris (2011) proposed that the SWD indicator measures "evaluations of regime performance", the third level in her five-fold model of political support. However, this interpretation is not unique: Ferrín (2016, 4) and Quaranta (2018, 4) identified that SWD has been reported to measure at least 13 concepts, such as "overall satisfaction with the present democratic political system", "performance of the democratic political system", or "support for the democratic processes". The theoretical concepts that the SWD indicator measures has been assessed with theoretical arguments or statistical analyses.

Theoretical arguments.

To assess theoretically which concept the SWD indicator measures, it is useful to distinguish between direct and indirect measures (Saris and Gallhofer, 2007, Ch. 1). A direct measure of a concept is a question asking explicitly about that concept. Thus, the SWD indicator is a direct measure of the concept "satisfaction with the way democracy works". Indirect measures of a concept are derived "on the assumption that there is a strong relationship between the variable of interest and another variable that can more easily be measured" (Saris and Gallhofer, 2007, 21). In practice the SWD indicator is sometimes treated as a direct measure of other concepts, such as "satisfaction with the performance of democracy" (e.g., Norris, 2011, 28). Nonetheless, since there are interpretations

of “the way democracy works” that do not correspond to “performance of democracy”, the SWD indicator is not a direct measure of the second concept, but an indirect one.

Using the SWD as an indirect measure has the disadvantage that the strength of the relationships between the SWD indicator and the theoretical concepts indirectly measured is a priori unknown and may be weak. Thus, a statistical association which holds for the concept “satisfaction with the way democracy works” might not hold for indirectly measured concepts.

Statistical analyses.

The theoretical concept that the SWD indicator measures might be inferred based on its statistical relationship with other measures (e.g., another indicator of “system support”). If the correlation between the two indicators is high, both indicators may be considered to measure the same concept. However, these interpretations require both theoretical arguments and that other explanations for the correlations (e.g., spurious effects) are controlled for (e.g., using Structural Equation Modelling [SEM]). Canache, Mondak and Seligson (2001), Ferrín (2016) and Quaranta (2018) presented evidence on the statistical relationship of the SWD indicator with other measures. However, as long as they do not control for alternative explanations of the observed relationships, these cannot be discarded.

Thus, the extent to which the SWD indicator measures concepts different from “satisfaction with the way democracy works” is uncertain. High content validity is only guaranteed for the concept “satisfaction with the way democracy works”. However, this does not imply that the SWD indicator measures this concept perfectly: measurement errors can occur.

2.2 **Measurement errors**

2.2.1 Implications

Each measurement instrument has its own level of measurement errors. This can lead to different results. To illustrate this point, we use data from the European Social Survey (ESS) round 4 (United

Kingdom), 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 end-points slightly changed: "Extremely dis/satisfied" (fixed reference points) versus "Dis/satisfied" (not fixed reference points). A fixed reference point is a response option that all respondents understand without doubt in the same way, such as "completely satisfied" (DeCastellarnau, 2018).

The cross-distribution of the answers (see Online Appendix 1) 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 satisfied" with one scale, but "dis/satisfied" with the other. Similarly, correlations with other questions vary depending on the scale used (see Online Appendix 2). Thus, results for multivariate statistical analyses are also expected to change.

This illustrates that using different response scales can produce different results even in the same sample. This is linked to the sizes of measurement errors. Since previous substantive research used different scales for the SWD indicator (e.g. different numbers of answer categories, different labels), this could explain part of the differences in results across studies. However, since the true distributions or correlations are unknown, in order to determine which method is better, we need to estimate the measurement quality for each scale. The scale with the highest measurement quality (i.e., closer to 1, see Section 4.1) is the one with the smallest size of measurement errors.

2.2.2 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.

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

Source	Country	Mode of data collection	Scale characteristics		Measurement quality	
			No. answer categories	Labels	Lower estimate	Higher estimate
Revilla, 2010	Netherlands	Face-to-face, telephone, web	11	Extremely dis/satisfied	.83	.83
				Very dis/satisfied	.57	.63
Revilla and Saris, 2013a	Netherlands	Face-to-face, web	11	Extremely dis/satisfied	.67	.78
				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
				Dis/satisfied	.83	.83
			5	Strongly dis/agree	.46	.65

Revilla and Ochoa, 2015	Mexico, Colombia	Web	11	Completely dis/satisfied	.78	.88
				Dis/satisfied	.70	.80
			5	Strongly dis/Agree	.44	.57
DeCastellarnau and Revilla, 2017	Norway	Web	11	Extremely dis/satisfied	.85	.89
				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). 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).

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. In general, the 11-point item specific scales yield higher quality than the 5-point "Strongly dis/agree" scales, although differences exist across studies.

However, this previous research has some limitations. First, estimates are only available for few countries and scales. Second, the estimates differ across studies, but the reasons behind these

variations are unclear. For example, the same scale (11-point extremely "Dis/satisfied") yields the highest quality (.85) 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.

2.2.3 Determinants of measurement quality

Saris and Gallhofer (2007; 2014) proposed a list of characteristics expected to affect measurement quality, including formal, topic-based, linguistic, layout and mode of data collection characteristics. In this paper, we focus on differences in measurement quality across response scales, countries and languages. The topic (SWD) and mode of data collection (face-to-face using showcards) are fixed.

On the one hand, previous research found that response scales characteristics affect measurement quality (for an overview see DeCastellarnau, 2018, who identified up to 23 characteristics). For instance, item specific scales have been found of higher quality than dis/agree scales (Saris *et al.*, 2010). Similarly, scales with at least two fixed reference points have been found of higher quality (Revilla and Ochoa, 2015). Additionally, scales with higher number of answer categories (up to a certain level) are argued to have higher quality, although the evidence is mixed (DeCasterllarnau, 2018).

However, previous research has not studied sufficiently all the different scale characteristics that could affect measurement quality. Besides, it usually provides information only about the effect of one characteristic at a time, but characteristics often interact with each other. Finally, a given scale may have different measurement qualities when used to measure different concepts.

On the other hand, previous research 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 (Bosch and Revilla, *in press*): 1) social desirability (the tendency of respondents

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

to select answers that are more socially accepted); 2) centrality (or saliency) of the topic in respondents' mind; and 3) linguistic characteristics. Due to the first two points, one can expect different measurement qualities across countries, even when the language and measurement instrument are constant. Moreover, due to linguistic differences, one can expect different measurement qualities across and within countries, because languages have different inherent structures (Zavala-Rojas, 2016).

3 Contribution

Substantive literature has not addressed the size of measurement errors of the SWD indicator. However, previous literature suggests that measurement errors can be large, and vary across response scales, countries and languages.

Thus, the main goal of this paper is to provide estimates of the size of measurement errors for different scales, countries and languages. In doing so, we contribute to the literature in several ways. First, these estimates are useful because: 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 compared across groups if the 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.

Second, 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 analyze three MTMM experiments implemented in the ESS, providing estimates for seven response scales and 38 country-language groups.

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.

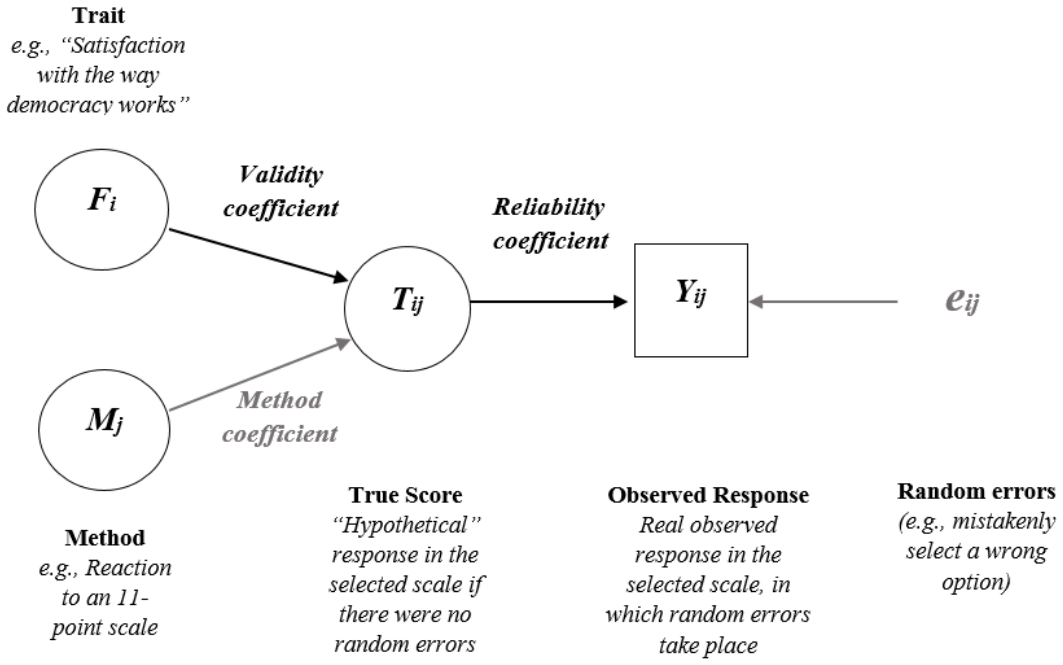
Fourth, previous estimates of the measurement quality of the SWD indicator are presented in papers in which its use was incidental and with a clear methodological focus (e.g., comparing modes of data collection). Thus, these estimates were not connected to the substantive literature and difficult to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

4 Method and data

4.1 True Score Model

In order to estimate it, measurement quality is defined with a SEM model (concretely, a specific Confirmatory Factor Analysis). While different models have been proposed to analyse data from MTMM experiments, we use the True Score model (Saris and Andrews, 1991) following Saris and Satorra (2018),. Figure 1 represents this model for the concept “satisfaction with the way democracy works”.

Figure 1. Path Diagram of the True Score model



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

$$T_{ij} = v_{ij} F_i + m_{ij} M_j \quad (1)$$

$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \quad (2)$$

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 or systematic component (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 random errors take place). 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 ($m_{ij}^2=1-v_{ij}^2$). 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 the answer).

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. Some of these assumptions can be relaxed in order to improve the fit of the model, leading to a Final Model from which the estimates are collected.

Measurement quality (q_{ij}^2), the strength of the relationship between the trait and the observed responses, is equal to the product of reliability and validity: $q_{ij}^2 = r_{ij}^2 * v_{ij}^2$. It represents the proportion of the variance in the observed responses explained by the variance in the underlying trait. It ranges from 0 (no relationship between the indicator and the trait) to 1 (perfect measurement). Measurement errors are defined as $1\text{-}q_{ij}^2$. Following DeCastellarnau and Revilla (2017), we use the following thresholds to interpret the estimates: the quality is classified as "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 $q^2 < .5$.

4.2 **Multitrait-multimethod approach**

In order to estimate measurement quality, the True Score model needs to be identified. This is the case when enough correlated traits (typically three) are repeated using enough methods (also typically three). However, this so-called MTMM approach (Campbell and Fiske, 1959; Andrews, 1984; Saris and Andrews, 1991) has several problems, particularly an increased respondent burden due to the extra answers required, and possible memory effects (Van Meurs and Saris, 1995; Schwarz, Revilla and Weber, 2020).

Thus, Saris, Satorra and Coenders (2004) proposed the Split-Ballot MTMM (SB-MTMM) approach, where the sample is randomly divided in different groups, each group answering a different

combination of two methods. However, the SB-MTMM approach frequently led to estimation problems, especially for a 2-group design (Revilla and Saris, 2013b). To overcome these problems, Saris and Satorra (2018) proposed the EUPD approach, that can be used when similar datasets are available. The general idea of the EUPD approach is to estimate a Pooled Data Model (PDM), store its estimates and use them to get an identified model in each country(-language) group (the unit of interest). 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 shows 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).

4.3 **Data**

We used data from three 2-group SB-MTMM experiments about Political Satisfaction implemented in the ESS rounds 1 (R1, 2002), 2 (R2, 2004) and 4 (R4, 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 is 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, 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 one 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 Schwarz 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, with the way the government is doing its job and with 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.

Table 2. Main characteristics of M1 and main differences of M2-M7 with respect to M1

Round	Method	Number of points	Labels of end-points	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

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.

4.4 **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 using pairwise deletion. We excluded the individuals who did not answer the supplementary questionnaire during the same day (some countries allowed it in the first rounds) because answering on a different day 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, which correspond to the weighted average of the matrices of all country(-language) groups analyzed from the same round, as well as the weighted means and standard deviations. The weights are 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. Matrices were created with R 3.6.1 (R Core Team, 2019).

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 theoretical expectation was that the reaction of respondents to a given scale 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). Hence, the method effects for these traits might differ and these parameters were freed in the PDM when JRule suggested it (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).

5 Results

5.1 Overview of all estimates

Table 3 presents the full list of estimates of measurement quality for the SWD indicator. As stated in Section 3, these estimates can be used for different purposes. First, they can be used to select the best methods to measure SWD in future surveys. Particularly, the results 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 exact 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 3. For instance, the best option seems an 11-point scale with 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.

Table 3. Measurement quality estimates (q^2) for each country(-language) group and method

Country-language	M1 * (11-point, Extremely)	M2 (4-point, Very)	M3 (6-point, Extremely)	M4 (11-point, explicit midpoint)	M5 (11-point, Very)	M6 (11-point, Dis/satisfied)	M7 (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-	.89	.24	.46				

Mixed (R1)							
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
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).*

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, 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 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 an 11-point scale with labels "Extremely dis/satisfied" (M1) for the SWD indicator, but .44 in another study that used a 4-point scale with labels "Very dis/satisfied" (M2). 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 On-line 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 correlation were in fact the same.

Finally, these estimates can also be used to perform correction for measurement errors. In the previous paragraph, we provided an example of correction for measurement errors showing that: 1) both studies underestimate the correlation of interest, and 2) the difference in results could be fully explained by the use of scales with different sizes of measurement error to measure the SWD indicator. For a detailed explanation of how to correct for measurement errors in different models, we refer to Saris and Gallhofer (2007) and DeCastellarnau and Saris (2014).

In order to further investigate the variations across response scales and countries, we also present the aggregated results averaging across all country-language groups (to study variations across response scales) and averaging across methods (to study variations across country-language groups).

5.2 **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 different response scales across all the country-(language) groups included in a given round[4].

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 end- points	Other characteristics	Round	Average measurement quality (q^2)	Standard deviation

[4] We also computed the average for the subgroup of 15 countries analyzed in all three rounds (17 country-language groups for R2 and R4). 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 response scales from different rounds cannot be attributed to different country(-language) groups participating in each round.

M1	11	Extremely dis/satisfied	Horizontal layout, 3 fixed reference points, medium correspondence numerical/verbal labels, bipolar	R1	**.74**	*.06*
				R2	**.69**	*.07*
				R4	**.70**	*.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*

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 (<.50). This is an important finding: 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). However, this suggests that 4-point scales are

not a good option. In contrast, the highest quality is found for the 11-point scales with an explicit midpoint (M4) and with labels "Very dis/satisfied" (M5) (q^2=.84 for both, classified as "good").

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 the rest of 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).

5.3 **Measurement quality across country(-language) groups.**

In Table 4, the standard deviations 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. 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 (so groups that received the same methods[5]).

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
	Germany	.71	.10
	Great Britain	.70	.16

[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.

	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
	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 by 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.

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 four as "poor". Differences across groups are influenced by the rounds in which they participated (which determines 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 parameter were fixed to the same values in all country(-language) groups, but then we had to free some of them in order to obtain an acceptable fit. Additionally, the fact that the overall standard deviation of the methods ranges from .05 to .12 suggests 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 (.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. This is possible in the seven countries within which different language groups were analysed. 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 groups of countries which participated in R2 and R4, suggesting than country-specific characteristics are more

important that linguistic differences 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 rarer.

6 Discussion/conclusions

6.1 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 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 seven scales and 38 country(-language) groups using data from three MTMM experiments implemented in the ESS. Our results provide useful information to choose 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 correct for measurement errors.

Additionally, we found that the average 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, "good" quality) and the 4-point scale (M2, labels "Very dis/satisfied") the worst (.46, "unacceptable" quality). 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 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 reasons for the differences between the 11-point scales (M1, M4, M5 and M6), which differed only in their labels, is unclear and further research is needed.

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.

6.2 **Limitations**

This paper has some 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.

6.3 **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 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 difference in results can be a result of 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, 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 countries/languages not analyzed here.

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 can be improved by performing correction for measurement errors (Saris and Revilla, 2016).

7 References

Almond, Gabriel, and Simon Verba. 1963. *The Civic Culture*. Princeton, NJ: Princeton University Press.

Andrews, Frank M. 1984. Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach. *Public Opinion Quarterly* 48 (2): 409–42. doi:10.1086/268840.

Bol, Damien, Marco Giani, André Blais, and Peter John Loewen. 2020. The Effect of COVID- 19 Lockdowns on Political Support: Some Good News for Democracy? *European Journal of Political Research*, May. doi:10.1111/1475-6765.12401.

Bollen, Kenneth. 1989. *Structural Equations with Latent Variables*. Wiley.

Bosch, Oriol, and Melanie Revilla. (*in press)* The Quality of Survey Questions in Spain: A Cross-National Comparison. *Revista Española de Investigaciones Sociológicas*.

Campbell, Donald T., and Donald W. Fiske. 1959. Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin* 56 (2): 81–105. doi:10.1037/h0046016.

Canache, Damarys, Jeffery J. Mondak, and Mitchell A. Seligson. 2001. Meaning and Measurement in Cross-National Research on Satisfaction with Democracy. *Public Opinion Quarterly* 65 (4): 506–28. doi:10.1086/323576.

Coppedge, Michael, John Gerring, David Altman, Michael Bernhard, Steven Fish, Allen Hicken, Matthew Kroenig, *et al.* 2011. Conceptualizing and Measuring Democracy: A New Approach. *Perspectives on Politics* 9 (2): 247–67. doi:10.1017/S1537592711000880.

Dassonneville, Ruth, and Ian McAllister. 2020. The Party Choice Set and Satisfaction with Democracy. *West European Politics* 43 (1). Routledge: 49–73. doi:10.1080/01402382.2019.1609286.

DeCastellarnau, Anna, and Willem 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/.

DeCastellarnau, Anna, and Melanie Revilla. 2017. Two Approaches to Evaluate Measurement Quality in Online Surveys: An Application Using the Norwegian Citizen Panel. *Survey Research Methods* 11 (4). European Survey Research Association: 415–33. doi:10.18148/srm/2017.v11i4.7226.

DeCastellarnau, Anna. 2018. A Classification of Response Scale Characteristics That Affect Data Quality: A Literature Review. *Quality and Quantity* 52 (4). Springer Netherlands: 1523–59. doi:10.1007/s11135-017-0533-4.

Easton, David. 1965. *A System Analysis of Political Life*. New York: Wiley.

Easton, David. 1975. A Re-Assessment of the Concept of Political Support. *British Journal of Political Science* 5: 435–37.

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. doi: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. doi: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. doi:10.21338/NSD-ESS4-2008.

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

Ferrin, Mónica. 2016. An Empirical Assessment of Satisfaction with Democracy. In *How Europeans View and Evaluate Democracy,* 283–306. Oxford University Press.

Ham, Carolien van, and Jacques Thomassen. 2017. The Myth of Legitimacy Decline. In *Myth and Reality of the Legitimacy Crisis: Explaining Trends and Cross-National Differences in Established Democracies*, edited by Carolien van Ham, Jacques Thomassen, Kees Aarts, and Andeweg. Oxford University Press. doi:10.1093/oso/9780198793717.003.0002.

Jöreskog, Karl, and Dag Sörbom (version 8.72). 2005. Lisrel 8. Uppsala, Sweden: Scientific Software International.

Linde, Jonas, and Joakim 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.

Linz, Juan and Alfred Stepan. 1996. *Problems of Democratic Transition and Consolidation*. Baltimore, MD: Johns Hopkins University.

Meurs, A Van, and Willem E. Saris. 1989. Memory Effects in MTMM Studies. In *Evaluation of Measurement Instruments by Meta-Analysis of Multitraitmultimethod Studies*, edited by A. Van Meurs and Willem E. Saris, 134–146. Amsterdam: North Holland.

Norris, Pippa. 2011. The Conceptual Framework. In *Democratic Deficit: Critical Citizens Revisited*, edited by Pipa Norris 19–37. Cambridge University Press.

Oberski, Daniel, and Albert Satorra. 2013. Measurement Error Models With Uncertainty About the Error Variance. *Structural Equation Modeling: A Multidisciplinary Journal* 20 (3): 409–28. doi:10.1080/10705511.2013.797820.

Oberski, Daniel, Willem E. Saris, and Jacques Hagenaars. 2007. Why Are There Differences in the Quality of Questions across Countries? In *Measuring Meaningful Data in Social Research*, edited by Geer Loosveldt, Marc Swyngedouw, and Bart Cambre, 281–299. Acco.

Przeworski, Adam. 1999. Minimalist Conception of Democracy: A Defense. *Democracy's Value* 23: 12–17.

Quaranta, Mario. 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. R Foundation for Statistical Computing. Vienna, Austria.

Revilla, Melanie. 2010. Quality in Unimode and Mixed-Mode Designs: A Multitrait-Multimethod Approach. *Survey Research Methods* 4 (3): 151–64. doi:10.18148/srm/2010.v4i3.4278.

Revilla, Melanie, and Carlos Ochoa. 2015. Quality of Different Scales in an Online Survey in Mexico and Colombia. *Journal of Politics in Latin America* 7 (3): 157–77.

Revilla, Melanie, and Willem 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): 242–53. doi:10.1093/ijpor/eds007.

Revilla, Melanie, and Willem 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, Melanie, Carlos Poses, Oriol Serra, Marc Asensio, Hannah Schwarz, and Wiebke 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, Melanie, Diana Zavala-Rojas, and Willem E. Saris. 2016. Creating a Good Question: How to Use Cumulative Experience. In *The SAGE-Handbook of Survey Methodology*, edited by Christof Wolf, Dominique Joye, Tom W Smith, and Yang-chih Fu, 236-254. Sage

Revilla, Melanie, Willem E. Saris, Germán Loewe, and Carlos 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.

Saris, Willem E., Albert Satorra, and Germà Coenders. 2004. A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design. *Sociological Methodology*, 311–47. doi:10.1111/j.0081-1750.2004.00155.x.

Saris, Willem E., Albert Satorra, and William van der Veld. 2009. Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal* 16 (4): 561–82. doi:10.1080/10705510903203433.

Saris, Willem E., and Albert Satorra. 2018. The Pooled Data Approach for the Estimation of Split-Ballot Multitrait–Multimethod Experiments. *Structural Equation Modeling* 25 (5): 659–72. doi:10.1080/10705511.2018.1431543.

Saris, Willem E., and Albert Satorra. 2019. Comparing BSEM and EUPD Estimates for Two-Group SB-MTMM Experiments. *Structural Equation Modeling* 26 (5). Routledge: 745–49. doi:10.1080/10705511.2019.1576046.

Saris, Willem E., and Frank Andrews. 1991. Evaluation of Measurement Instruments Using a Structural Modeling Approach. In *Measurement Errors in Surveys*, edited by P.P. Biemer, R.M Groves, L.E. Lyber, N.A. Mathiowetz, and S. Sudman, 575–98. New York: Joh Wiley & Sons, Inc.

Saris, Willem E., and Irmtraud Gallhofer. 2007. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Wiley.

Saris, Willem E., and Melanie Revilla. 2016. Correction for Measurement Errors in Survey Research: Necessary and Possible. *Social Indicators Research* 127 (3). Springer Netherlands: 1005–20. doi:10.1007/s11205-015-1002-x.

Saris, Willem E., Melanie Revilla, Jon A. Krosnick, and Eric M. Shaeffer. 2010. Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response Options. *Survey Research Methods* 4 (1): 61–79.

Schwarz, Hannah, Melanie Revilla, and Weber Wiebke. 2020. Memory Effects in Repeated Survey Questions: Reviving the Empirical Investigation of the Independent Measurements Assumption. *Survey Research Methods* 14 (3): 325–44. doi:https://doi.org/10.18148/srm/2020.v14i3.7579.

Stadelmann-Steffen, Isabelle, and Adrian Vatter. 2012. Does Satisfaction with Democracy Really Increase Happiness? Direct Democracy and Individual Satisfaction in Switzerland. *Political Behavior* 34 (3): 535–59. doi:10.1007/s11109-011-9164-y.

Thomassen, Jacques, and Carolien van Ham. 2017. *A Legitimacy Crisis of Representative Democracy?* In *Myth and Reality of the Legitimacy Crisis: Explaining Trends and Cross-*

National Differences in Established Democracies, edited by Carolien van Ham, Jacques Thomassen, Kees Aarts, and Rudy Andeweg. Oxford University Press.

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

Zavala-Rojas, Diana. 2016. "Measurement Equivalence in Multilingual Comparative Survey Research." PhD Diss., Universitat Pompeu Fabra.