Survey Research Methods (2021)
Vol. 15, No. 3, pp. 235-256
doi:10.18148/srm/2021.v15i3.7816
European Survey Research Association

Measurement quality of 67 common social sciences questions across countries and languages based on 28 Multitrait-Multimethod experiments implemented in the European Social Survey

Carlos Poses
RECSM—Universitat Pompeu Fabra
Barcelona, Spain

Melanie Revilla
RECSM—Universitat Pompeu Fabra
Barcelona, Spain

Marc Asensio
RECSM—Universitat Pompeu Fabra
Barcelona, Spain

Hannah Schwarz
RECSM—Universitat Pompeu Fabra
Barcelona, Spain

Wiebke Weber
RECSM—Universitat Pompeu Fabra
Barcelona, Spain

Survey data is used in many social science studies. The measurement quality of these data is crucial as it determines the accuracy of the information on which these studies are based. Besides, since these studies are used to provide insights to key political and social actors, it also determines the accuracy of the information on which crucial decisions are based. In this paper, we estimated the measurement quality (proportion of the variance of the observed survey responses explained by the latent trait of interest) of 67 common social sciences questions that were part of Multitrait-Multimethod experiments in the seven first rounds of the European Social Survey. These questions were asked using response scales with different characteristics and in up to 41 country-language groups. Our results show that measurement errors are omnipresent: the average measurement quality across all questions is 0.65. Thus, overall, on average 35% of the variance in the observed survey answers can be attributed to measurement errors. Furthermore, the size of errors varies across questions as well as across country-language groups. The questions' average measurement quality across all country-language groups ranges from 0.25 to 0.88, depending on the response scale and topic, and the country-language groups' average measurement quality across questions ranges from 0.52 to 0.76. Thus, the impact of measurement errors on applied research can be different depending on the exact question formulation and response scale used as well as on the country and language of interest. Consequently, in each study, researchers should consider assessing the size of the measurement errors of their variables and how this affects their results.

Keywords: European Social Survey (ESS); measurement quality; reliability; validity; Multitrait-Multimethod (MTMM) experiments; social sciences; data quality; questionnaire design

1 Introduction

While researchers are usually interested in studying theoretical concepts, they often need to operationalize them as survey questions in order to collect data and answer their research questions and/or test their hypotheses. Yet survey questions are not perfect measures of the concepts of interest because of measurement errors, for instance, errors due to mistakes in option selection by respondents or interviewers or due to systematic reactions of the respondents to a certain response scale.

These measurement errors can be quite large, particularly in social sciences surveys that tend to study subjective concepts. Alwin (2007) results suggest that, on average, around half of the variance in observed survey answers is due to measurement errors and not substantive differences. Moreover, the size of measurement errors varies depending, among others, on the exact question formulation, the re-

Contact information: Carlos Poses, Universitat Pompeu Fabra, Edifici Mercè Rodoreda, Despatx 24.406 Ramón Trías Fargas 25–27, 08005 Barcelona (E-mail: carlos.gonzalezp@upf.edu)

sponse scale characteristics, the country, language, and mode of data collection (Saris & Gallhofer, 2014). Thus, it is important to estimate the size of measurement errors in different contexts, both to design better questionnaires, by selecting formulations and scales that lead to less measurement errors (Revilla, Zavala-Rojas, & Saris, 2016; Weber, Gallhofer, & Saris, 2020) and to obtain the information to perform correction for measurement errors (Saris & Revilla, 2016).

Different approaches exist to estimate measurement errors of single questions (e.g. Tourangeau, 2020). One common approach to estimate them is the multitrait-multimethod (MTMM) approach (e.g. Saris & Gallhofer, 2014), which consists in repeatedly measuring several correlated latent concepts (called "traits"), each with the same question stem but using different methods (e.g. different response scales). Then, the resulting MTMM matrices can be analyzed using Confirmatory Factor Analysis. More precisely, in this paper, we use the True Score Model (Saris & Andrews, 1991, see Section 4.1 for details). This model allows to estimate the measurement quality, defined as the strength of the relationship between the latent variable of interest and the observed survey responses. Measurement quality also represents the proportion of the variance of the observed survey responses stemming from the latent variable of interest. Thus, measurement quality equals 1 minus measurement errors. The closer to 1 the measurement quality, the lower the level of measurement errors and the better do survey questions measure the concept of interest.

In this paper, we use 28 MTMM experiments implemented in the first seven rounds of the European Social Survey (ESS) to estimate the measurement quality of 67 common social sciences survey questions across up to 41 country-language groups[1]. Our main goal is to provide estimates of the measurement quality of a large set of questions as measured in a probability-based survey known for its high quality standards, the ESS. Our focus is on the quality of single questions. Stated differently, we do not consider latent variables measured with multiple indicators but only traits measured with a single indicator.

2 Background

2.1 Previous studies providing measurement quality estimates

A lot of previous published studies already provide estimates of measurement quality coming from MTMM experiments (Andrews, 1984; Bosch, Revilla, DeCastellarnau, & Weber, 2019; Coromina & Coenders, 2006; Költringer, 1995; Mingwei, 2015; Revilla & Ochoa, 2015; Revilla, Saris, & Krosnick, 2014; Revilla, Saris, Loewe, & Ochoa, 2015; Rodgers, Andrews, & Regula Herzog, 1992; Saris & Gallhofer, 2014; Saris, Revilla, Krosnick, & Shaeffer, 2010; A. Scherpenzeel, 2009; A. C. Scherpenzeel & Saris, 1997). Some of them use similar ESS data as we do in this paper. For instance, Saris et al. (2010) use data from two experiments of round 2 (Current job and Evaluation of doctors, see Table 1) and two experiments of round 3 (Evaluation of immigration and Immigration perceptions). Revilla et al. (2014) use data from all experiments of round 3. Furthermore, Saris and Gallhofer (2014) present mainly aggregated results based on data coming from experiments in rounds 1 to 3. ESS reports about the MTMM experiments of each round are also available through the ESS website[2].

However, for the vast majority of survey questions, no estimates of measurement quality are available. Implementing MTMM experiments does not only increase the costs of data collection but also increases respondents burden due to longer questionnaires and the repetition of questions. Additionally, long surveys are required in order to prevent memory effects, i.e., respondents recalling their first answer and using this information to answer the second time instead of going through the process of answering anew (Van Meurs & Saris, 1990; Schwarz, Revilla, & Weber, 2020). As a result, only few surveys provide the information needed to estimate measurement quality based on MTMM experiments, and, even when they do, they usually provide this information only for a relatively small subset of variables. For instance, the ESS data allows to estimate the measurement quality of three to 12 questions in each round out of a total of more than 100 questions per round.

Besides, many of the previously published MTMM studies suffer from at least one of the following issues:

1. They use a two-group split-ballot MTMM design (e.g. Saris et al., 2010). In this design, respondents are randomly assigned to two groups. Each group answers the same question twice, but using different scales (Saris, Satorra, & Coenders, 2004). While this design reduces the number of repetitions for each respondent, it led to important estimation problems when analysed on a country-by-country basis (Revilla & Saris, 2013). These problems can affect the results.

2. The time between the first answer and its repetition is shorter than 20 minutes for at least a non-negligible part of

[1] In the ESS, in multilingual countries, the respondents can choose in which language to answer. Thus, the analyses are done for groups of respondents within a country that answered in the same language (e.g., Belgium-Dutch or Belgium-French).

[2] https://www.europeansocialsurvey.org/search?q=results%20of%20split-ballot&rows=25&fq=doctype_facet:%22Methodology%22 Additionally, Saris, Satorra, and Van der Veld (2009) present quality estimates for the experiment Media use, but only in round 1 in Austria; Revilla and Saris (2013) present quality estimates for the round 4 experiments but only in the Netherlands, and Revilla and Ochoa (2015) present quality estimates for the round 4 experiments Political satisfaction and Political trust, but only in Spain. Lastly, forthcoming papers are expected, some currently under preparation, that will use part of the estimates of this paper as well as additional results.

the respondents (e.g. Bosch et al., 2019) so memory effects can be expected, biasing the quality estimates (Van Meurs & Saris, 1990).

3. The studies are based on non-probability samples (e.g. Revilla & Ochoa, 2015).

2.2 Predicting measurement quality: the Survey Quality Predictor (SQP) software

As previously mentioned, for the vast majority of survey questions, no MTMM estimates of measurement quality are available. Furthermore, previous research suggests that measurement quality varies depending on many factors, that can also interact with each other (Saris & Gallhofer, 2014). Thus, it is difficult to infer the measurement quality for a specific question in a given survey without collecting new data.

Back in 1984, Andrews proposed a solution to this issue: first, one can use the existing MTMM estimates of measurement quality and try to explain them by the questions' characteristics (Andrews, 1984, p. 436). Then, one can use the characteristics of new questions to predict their quality. This idea was implemented by Saris, van der Veld, and Gallhofer (2000) who launched the first SQP software in 2001. In 2012, it was further improved in a new version, SQP 2 (Saris et al., 2011; Saris, 2013), available at: sqp.upf.edu.

SQP 2 is based on a meta-analysis of more than 3,000 MTMM estimates obtained from multiple surveys, during more than two decades, and in more than 20 different countries and languages (Saris, 2013). It uses a random forest approach to provide predictions of the measurement quality of survey questions based on a detailed coding scheme containing up to 60 different formal and linguistic characteristics, such as: the number of response categories, the centrality of the question in respondents' minds, the position of the question in the survey, or the presence of an interviewer.

Therefore, SQP 2 is another source providing information about the measurement quality that can be expected under different conditions, complementary to MTMM estimates. Besides, SQP 2 uses estimates from long surveys (usually much longer than 20 minutes) and based on probability samples. Thus, it does not suffer from two of the main problems mentioned for some of the previous studies presenting measurement quality estimates. However, SQP 2 has some limitations. In particular, the quality of its predictions depends on the data included in the meta-analysis of MTMM studies on which it is based. SQP 2 mainly uses data from the first three ESS rounds. Therefore, there are still many topics, question formats, countries and languages that are not included in the current SQP database. Even if predictions can be obtained, the quality of these predictions might be dubious if there are no similar topics, formats, countries or languages in the meta-analysis database. For instance, SQP 2 does not include data for countries such as Italy, Lithuania, Croatia or Russia, and/or languages spoken in those countries. Also, SQP 2 does not consider the evaluative dimension of the scale (item specific versus agree-disagree) even if previous research suggests that this has an impact on measurement quality (see DeCastellarnau, 2018) for an overview and http://sqp.upf.edu/loadui/#limits for details on SQP 2 limitations).

2.3 Contribution

To sum up, there is already some research using MTMM experiments to estimate the measurement quality of survey questions. In addition, the SQP 2 software allows predicting the measurement quality of new questions by coding their characteristics. Nevertheless, both sources have limitations. As a consequence, more research estimating measurement quality is needed.

In this paper, we present estimates of measurement quality for 67 questions of the ESS, one of the most important social science surveys in Europe, and across 41 country-language groups. By doing so, we make contributions in several ways. First, we make it easier to have an overview over the estimates because

1. we present them together in one place, in contrast to previous studies that looked at three or four experiments per publication;

2. we estimate all of them with the same procedure, which makes them more comparable;

3. we present estimates for questions for which, to the best of our knowledge, no MTMM estimates of measurement quality have previously been published (e.g. questions that were part of the round 5 experiment Satisfaction with the police or round 7—Subjective competence, see Table 1);

4. we provide estimates for most country-language groups present in the ESS, including for countries for which estimates were not previously available (e.g., Croatia).

Besides, we try to reduce several problems identified in previous research:

1. Many previous studies providing measurement quality estimates use data from a two-group split ballot design and analyse these data one country at a time. However, Revilla and Saris (2013) showed that this way of analysing two-group split-ballot MTMM experiments frequently leads to non-convergence and improper solutions (e.g. negative variances). Therefore, we use an alternative procedure, the estimation using pooled data approach or EUPD (Saris & Satorra, 2018) to analyse our experiments, because previous research suggests that it performs better than country-by-country analysis (Revilla et al., 2021; Saris & Satorra, 2018, see also section 4.3.)

2. In the ESS, the time between repetitions of the same question is often around one hour. Therefore, memory bias should be limited in our analyses.

3. The ESS is based on probability samples.

Overall, our results can be used to investigate differences in measurement quality across questions, countries and languages. Besides, the measurement quality estimates provided can help making informed decisions for future questionnaire design - by selecting the response scales that maximize measurement quality (Revilla et al., 2016; Weber et al., 2020). They can also help with the interpretation of results from previous studies, mainly the ones using the ESS questions studied, and especially regarding effect sizes. Finally, they can be used to correct for measurement errors (Saris & Gallhofer, 2014; Saris & Revilla, 2016).

3 Data

The data comes from the ESS round 1 (ESS, 2003) to round 7 (ESS, 2015). Data was collected face-to-face at respondents' homes. Most questions included visual aids in the form of showcards. In each round, the survey was divided into a core module and several rotating modules. A supplementary questionnaire was presented to respondents at the end of the main interview. Although this supplementary questionnaire was usually also administered face-to-face, in some rounds and countries, it was self-administered (paper-and-pencil). It included, first, a human values scale, and second, repetitions of previous questions usually using different response scales. Repetitions were used to estimate the measurement quality of a set of questions through split-ballot MTMM experiments.

In this paper, we focused on the first seven rounds since the MTMMs implemented in later rounds used a different design. The MTMM experiments analysed were all implemented using a split-ballot design (see Section 4.2) in which the same method was proposed to all respondents in the main questionnaire and then, in the supplementary questionnaires, respondents were randomly assigned to different methods. There were six split-ballot MTMM experiments in rounds 1, 2 and 4; four in rounds 3 and 6; and three in rounds 5 and 7. However, we excluded two experiments due to problems in the data. Furthermore, we could not find a satisfactory solution for another two experiments. Thus, we obtained results for 28 MTMM experiments (see Appendix A1). We analysed the 34 countries[3] for which data was available in at least five split-ballot MTMM experiments across the seven rounds. We used the final released integrated datasets for each round (main questionnaire and supplementary questionnaires) available on the ESS website (https://www.europeansocialsurvey.org/data/round-index.html) at the 1st of July, 2019 (ESS, 2003; ESS, 2005; ESS, 2007; ESS, 2009; ESS, 2011; ESS, 2013; ESS, 2015). The names of the variables within each experiment, as used in the ESS database and questionnaires, can be retrieved from Appendix A1.

Because different levels of measurement quality are expected when different languages are used (Zavala-Rojas, 2016), we analysed the data separately for each language in multilingual countries, except for round 1, where no information to divide the respondents based on language was available. However, when the number of observations in secondary languages was too small to analyse them separately (less than 70 observations for a given split-ballot group), respondents who answered in such languages were excluded (see Appendix A2). Furthermore, we only considered respondents who answered to the main and supplementary questionnaires on the same day since answering to the supplementary questionnaire on a different day can impact the answers and their quality (Oberski, Saris, & Hagenaars, 2007).

Lastly, as we used the EUPD procedure, following Saris and Satorra (2018) recommendation, we excluded some country-language groups with a very different data structure. Concretely, we excluded country-language groups where the correlation of a given trait measured with different methods was of opposite sign (e.g. negative rather than positive) than for the majority of country-language groups (see Appendix A2). Such exclusions were rarely needed.

4 Method and analyses

4.1 The True Score Model

Different models have been proposed to analyse MTMM experiments. Following Saris and Satorra (2018), we use the True Score model as proposed by Saris and Andrews (1991). The model is summarized by the following equations:

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

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

where T_{ij} is the True Score or systematic component of the response, F_i is the i^{th} trait (e.g. trust in country's parliament), M_j is the j^{th} method (e.g. a unipolar four-point fully labeled response scale), Y_{ij} is the observed answer for the i^{th} trait and the j^{th} method, v_{ij} is the validity coefficient (when completely standardized), r_{ij} is the reliability coefficient (when completely standardized), and e_{ij} is the random error associated with Y_{ij}. The square of the validity coefficient, called validity (v_{ij}^2), represents the proportion of the variance of the True Score explained by the latent trait. The square of the reliability coefficient, called reliability (r_{ij}^2), represents the variance of the observed survey responses explained by the True Score.

The total measurement quality is computed as the product of the reliability and validity: $q_{ij}^2 = r_{ij}^2 * v_{ij}^2$. Measurement

[3] Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Latvia, Lithuania, Luxembourg, Netherlands, Norway, Poland, Portugal, Romania, Russian Federation, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine and United Kingdom.

quality represents the proportion of the variance in the observed survey responses explained by the underlying trait. Measurement errors are equal to $1 - q_{ij}^2$.

In addition, we initially (as a Base Model) assume that:

1. the random errors are uncorrelated with each other and with the independent variables in the different equations;
2. the traits are correlated;
3. the method factors are uncorrelated between them and with the traits;
4. the impact of the method factor on the traits measured with a common scale is the same;
5. the trait and method factors, as well as the random errors, are multivariate normally distributed;
6. the relations between the latent factors are linear and homoscedastic with mutual relationships.

4.2 The split-ballot MTMM

The usual way to get an identified True Score model requires measuring a minimum of three correlated traits, each using at least three different methods, which leads to a set of nine questions that the same respondent has to answer. However, in order to reduce both the burden due to the repetition of the same questions to the same respondents, data collection costs, and the risk of memory effects (Van Meurs & Saris, 1990), Saris et al. (2004) proposed the split-ballot MTMM approach. In this approach, respondents are randomly assigned to different groups, each group receiving a different combination of two methods instead of three, which normally leads to six questions per respondent instead of nine.

In order to implement split-ballot MTMM experiments, different designs can be used. The most common is a two-group design (e.g. group 1 getting methods 1 and 2 and group 2 methods 1 and 3), which has the advantage of all respondents receiving method 1 at time 1. However, this design often leads to non-convergence or improper solutions (e.g. negative variances) when using the classical multiple group Maximum Likelihood (ML) estimation on a country-by-country basis (Revilla & Saris, 2013).

4.3 The EUPD procedure

To overcome the estimation problems observed for the two-group design when analysing one country at a time, Saris and Satorra (2018) proposed the EUPD procedure that can be used when several similar datasets are available (e.g. multiple countries with similar experiments). The main idea is that an identified model can be achieved in each dataset by looking first at all datasets together and then using the resulting information to get an identified model in each separate dataset. Therefore, the procedure consists of two steps:

1. estimating a Pooled Data Model (PDM) using multiple group ML and
2. analysing each dataset separately using multiple group ML, initially fixing the trait loadings and possibly the method loadings (if different from 1) to the PDM estimated values.

Then, the model is tested and the misspecified parameters can be freed in each separate dataset. Previous research suggests that this procedure performs better than alternatives such as a simple country-by-country analysis, based on simulations showing a lower mean square error (Saris & Satorra, 2018), and simulations and empirical analyses showing higher levels of convergence and less improper solutions (Revilla et al., 2021; Saris & Satorra, 2018). It has also been shown to perform better than Bayesian estimation (Saris & Satorra, 2019).

4.4 Analyses

Pooled data stage. First, we used R 3.6.1 (R Core Team, 2019) to create all the matrices to be analysed. For each experiment and split-ballot group, the matrix for the pooled data analyses was created as the weighted average correlation matrix of all country-language groups. The weights were calculated by dividing the sample size of each country-language group in a given experiment and split-ballot group by the total sample size across all country-language groups for that experiment and split-ballot group. The weighted standard deviations and means were calculated in a similar way.

Then, we used the program LISREL 8.72 (Joreskog & Sorbom, 2005) to estimate the True Score MTMM model (Base Model defined in Section 4.1). We used ML estimation for multiple groups (the different split-ballot groups; see example of input in Appendix B).

In order to determine which modifications were necessary for each PDM, we tested for misspecifications using the JRule software (Van der Veld, Saris, & Satorra, 2008) based on the testing procedure developed by Saris et al. (2009), which has two main advantages:

1. it takes into account the power and
2. it tests at the parameter level, which helps to decide which corrections to introduce.

We combined the information from JRule with theoretical considerations based on several experts' judgments (prior to the statistical analyses) about which corrections were more likely to be needed. For more details about the testing and parameters that were introduced in each experiment, we refer to Revilla et al. (2021, especially Table 1).

Separate datasets stage. Once a final PDM was found for an experiment, we used the estimates of the trait loadings and method loadings (if they differed from 1) from the PDM to fix the corresponding parameters to these values in each of the country-language groups analysed separately. Then, we used the ML estimation in LISREL for multiple groups (the different split-ballot groups; see example of input in Appendix B).

Again, we tested the goodness-of-fit of all models using JRule. In this case, we mainly expected misspecifications regarding the size of the parameters that were fixed to the PDM values, although other changes in the model were sometimes required. For more details about the testing stage, we refer to Revilla et al. (2021, especially Tables 2 and 3).

Overall, the analyses led to more than 7,200 quality estimates. However, our focus is on the quality of the questions[4] asked in the ESS main questionnaire, which represent 2,135 estimates. The reliability, validity and measurement quality for each question in each country-language group can be found in Online Appendix 1[5]. Within this manuscript, we summarize the results by presenting the average measurement quality:

1. across all questions and country-language groups, to get an overall idea of the measurement quality;
2. across country-language groups, to see how measurement quality varies across questions, and
3. across questions, to see how measurement quality varies across country-language groups.

We also present the 95% confidence interval for the mean, assuming normality. Since this is an average of estimates, each one with its own uncertainty, we are aware of the limitations of this confidence interval. However, confidence intervals for the original estimates are not available in LISREL 8.72.

Following DeCastellarnau and Revilla (2017), we use similar thresholds as the ones proposed for the Cronbach's alpha (Bland & Altman, 1997; Santos, 1999) to interpret the estimates of measurement quality. Thus, the quality is classified as: "unacceptable" if $q^2 < 0.5$ (more variance due to errors than due to the underlying concept to be measured), "poor" if $0.5 \leq q^2 < 0.6$, "questionable" if $0.6 \leq q^2 < 0.7$, "acceptable" if $0.7 \leq q^2 < 0.8$, "good" if $0.8 \leq q^2 < 0.9$ and "excellent" if $q^2 \geq 0.9$. However, one should keep in mind that each proposed threshold is partially arbitrary. The specific assessment of any level of measurement quality may depend on several aspects, such as the topic, the measurement instrument (e.g. survey questions versus other measurement instruments), the feasibility of attaining higher measurement qualities and/or the researchers' aims and objectives.

5 Results

5.1 Overall average measurement quality

The average measurement quality of the 67 questions is 0.65, with a standard deviation of 0.14. This means that, overall and on average, 65% of the variance of the observed responses is due to the underlying concepts of interest whereas 35% is due to measurement errors. This 65% of variance due to the underlying concept is somehow higher than what Alwin (2007) suggested (around 47% to 53% due to the underlying concept), based on the analyses of 500 survey measures from studies conducted at the University of Michigan. Thus, the measurement quality of these ESS questions could be slightly better than what is observed for other samples, surveys, and questions. Nevertheless, using similar thresholds as DeCastellarnau and Revilla (2017), this overall average would be classified as "questionable". Considered separately, the average validity is 0.88 and the average reliability is 0.74. Besides, the large standard deviation suggests that this average hides a lot of variability across questions, countries, and languages. A histogram of the distribution of all measurement quality estimates can be found in Appendix C.

5.2 Measurement quality per question: average across all country-language groups

Table 1 shows the average measurement quality for each question within each experiment, across all country-language groups, with 95% confidence intervals in brackets. The first column of Table 1 corresponds to the MTMM experiment name, a generic name aimed at encompassing the overall topic of the questions used in that experiment[6], as well as the round in which the experiment was fielded. Within each experiment, columns 2, 4 and 6 show the trait that each question aims to measure (as stated in the ESS documentation) and column 8 the corresponding response scale. The whole wording of each question can be retrieved from the ESS webpage and open documentation, using the questions' names provided in Online Appendix 1. Columns 3, 5 and 7 present the average measurement quality for each question, across country-language groups.

First, measurement quality varies a lot across questions. The average measurement quality ranges from 0.25 (Trait 1, "Doctors keep whole truth from patients", Evaluation of doctors, round 2) to 0.88 (Trait 3, "Trust in the police", Political trust, round 1). Second, some variability is also present between the different questions within a given experiment. More precisely, the difference in measurement quality for the questions within each experiment ranges from 0 (e.g. Social trust, round 4) to 0.27 (between Trait 1 "Doctors keep whole truth from patients" and Trait 2 "Regular general practitioner/doctor treat patients as equals", Evaluation of doctors, round 2; or between Trait 2 "Men should take as much responsibility as women for home and children" and Trait 3 "Men should have more right to job than women when jobs are scarce", Gender inequality, round 2). The differences between questions that are part of a given experiment cannot

[4] By "question" we refer to a given request for an answer and response scale (that are used to measure a trait), as can be seen in Table 1.

[5] Online appendix 1 can be found at the UPF Repository by searching the name of the article, or the appendix DOI. Online appendix 1 has the following DOI: 10.34810/data122

[6] This is not a latent variable or concept and does not form part of the model.

Table 1
Measurement quality (q^2) per question: average across all country-language groups, with 95% confidence intervals [in brackets]

Round-Experiment	Trait 1	$q^2_{(T1)}$	CI	Trait 2	$q^2_{(T2)}$	CI	Trait 3	$q^2_{(T3)}$	CI	Response Scale
R1—Media use	TV watching, total time on average weekday	.81	[.76,.85]	Radio listening, total time on average weekday	.80	[.75,.85]	Newspaper reading, total time on average weekday	.68	[.62,.74]	8-points, 0 No time at all/ 7 More than 3 hours[b]
R1—Political efficacy	Politics too complicated to understand	.59	[.56,.62]	Could take an active role in a group involved with political issues	.69	[.63,.74]	Making mind up about political issues	.60	[.56,.65]	5-points, Never/Frequently[b,c]
R1—Politicial orientation	The less government intervenes in economy, the better for country	.51	[.47,.56]	Government should reduce differences in income levels	.58	[.55,.60]	Employees need strong trade unions to protect work conditions/wages	.68	[.64,.72]	5-points, Agree strongly/Disagree strongly[b]
R1—Politicial satisfaction	How satisfied with present state of economy in country	.61	[.57,.65]	How satisfied with the national government	.72	[.69,.76]	How satisfied with the way democracy works in country	.74	[.71,.77]	11-points, 0-Extremely Dissatisfied/10 Extremely Satisfied[a]
R1—Politicial trust	Trust in country's parliament	.78	[.75,.82]	Trust in the legal system	.81	[.77,.84]	Trust in the police	.88	[.86,.90]	11-points, No trust at all/Complete trust[a]
R1—Social trust	Most people can be trusted or you can't be too careful	.61	[.59,.64]	Most people try to take advantage of you, or try to be fair	.58	[.54,.61]	Most of the time people helpful or mostly looking out for themselves	.54	[.50,.59]	11-points, You can't be too careful/Most people can be trusted[a,c]
R2—Current job	Current job: Variety in work	.76	[.71,.80]	Current job: Job is secure	.66	[.60,.71]	Current job: health/safety at risk because of work	.63	[.59,.66]	4-points, Not at all true/Very true[b]
R2—Evaluation of doctors	Doctors keep whole truth from patients	.25	[.19,.31]	Regular general practitioner/doctor treat patients as equals	.52	[.47,.56]	Doctors discuss treatment with patient before they decide	.47	[.42,.51]	5-points, 1-Never or almost never/5 Always or almost always[b]

Continues on next page

Continued from last page

Round-Experiment	Trait 1	$q^2_{(T1)}$	CI	Trait 2	$q^2_{(T2)}$	CI	Trait 3	$q^2_{(T3)}$	CI	Response Scale
R2—Gender inequalities	Women should be prepared to cut down on paid work for sake of family	.53	[.50,.57]	Men should take as much responsibility as women for home and children	.34	[.29,.38]	Men should have more right to job than women when jobs are scarce	.61	[.57,.64]	5-points, Agree strongly/Disagree strongly[b]
R2—Political satisfaction	How satisfied with present state of economy in country	.60	[.58,.63]	How satisfied with the national government	.69	[.66,.72]	How satisfied with the way democracy works in country	.69	[.66,.71]	11-points, 0-Extremely Dissatisfied/10 Extremely Satisfied[a]
R2—Political trust	Trust in country's parliament	.76	[.73,.79]	Trust in the legal system	.77	[.75,.79]	Trust in politicians	.77	[.75,.79]	11-points, No trust at all/Complete trust[a]
R3—Evaluation of immigration	Immigration bad or good for country's economy	.58	[.53,.63]	Country's cultural life undermined or enriched by immigrants	.60	[.55,.65]	Immigrants make country worse or better place to live	.57	[.53,.61]	11-points, Bad for the economy/Good for the economy[a,c]
R3—Immigration perceptions	Allow many/few immigrants of same race/ethnic group as majority	.76	[.74,.79]	Allow many/few immigrants of different race/ethnic group from majority	.80	[.78,.83]	Allow many/few immigrants from poorer countries outside Europe	.81	[.78,.84]	4-points, Allow many to come and live here/Allow none[a]
R3—Life satisfaction	Feel what I do in life is valuable and worthwhile	.50	[.47,.54]	There are people in my life who care about me	.52	[.47,.56]	Feel close to the people in local area	.60	[.56,.64]	5-points, Agree strongly/Disagree strongly[b]
R3—Well-being	Love learning new things	.51	[.48,.53]	Feel accomplishment from what I do	.41	[.38,.44]	Like planning and preparing for future	.61	[.58,.64]	5-points, Agree strongly/Disagree strongly[b]
R4—Political orientation	Government should reduce differences in income levels	.62	[.58,.66]	Gays and lesbians free to live life as they wish	.75	[.72,.78]	-	-	-	5-points, Agree strongly/Disagree strongly[b]
R4—Political satisfaction	How satisfied with present state of economy in country	.56	[.52,.59]	How satisfied with the national government	.69	[.65,.73]	How satisfied with the way democracy works in country	.70	[.67,.74]	11-points, 0-Extremely Dissatisfied/10 Extremely Satisfied[a]

Continues on next page

Continued from last page

Round-Experiment	Trait 1	$q^2_{(T1)}$	CI	Trait 2	$q^2_{(T2)}$	CI	Trait 3	$q^2_{(T3)}$	CI	Response Scale
R4—Political trust	Trust in country's parliament	.72	[.69,.75]	Trust in the legal system	.73	[.71,.76]	Trust in the police	.79	[.77,.82]	11-points, No trust at all/Complete trust[a]
R4—Social trust	Most people can be trusted or you can't be too careful	.60	[.56,.63]	Most people try to take advantage of you, or try to be fair	.60	[.57,.63]	-	-	-	11-points, *You can't be too careful/Most people can be trusted[a]
R5—Effectiveness of the police	How likely be caught if made exaggerated or false insurance claim	.76	[.72,.81]	How likely to be caught if bought something that might be stolen	.76	[.71,.80]	How likely to be caught if committed traffic offence	.77	[.72,.82]	4-points, Not at all likely/Very likely[b]
R5—Satisfaction with the police	How successful police are at preventing crimes in country	.63	[.59,.67]	How successful police are at catching house burglars in country	.72	[.69,.75]	How quickly would police arrive at a violent crime scene near to where you live	.77	[.72,.82]	11-point, Extremely unsuccessful/Extremely successful[a,c]
R6—Evaluation of democracy	In country opposition parties are free to criticise the government	.65	[.61,.69]	In country the media are free to criticise the government	.68	[.65,.71]	In country the media provide citizens with reliable information to judge the gov	.62	[.59,.64]	11-point, Does not apply at all/Applies completely[a]
R6—Evaluation of immigration	Immigration bad or good for country's economy	.73	[.71,.76]	Country's cultural life undermined or enriched by immigrants	.77	[.74,.80]	Immigrants make country worse or better place to live	.78	[.75,.81]	11-points, Bad for the economy/Good for the economy[a,c]
R6—Everyday life engagement	Interested in what you are doing, how much of the time	.57	[.54,.60]	Absorbed in what you are doing, how much of the time	.59	[.57,.61]	Enthusiastic about what you are doing, how much of the time	.63	[.60,.65]	11-point, None of the time/All of the time[a]
R6—Feelings past week	Felt depressed, how often past week	.57	[.54,.60]	Sleep was restless, how often past week	.60	[.56,.65]	Felt lonely, how often past week	.53	[.50,.56]	4-point, None or almost one of the time/All or almost of the time[b]

Continues on next page

Continued from last page

Round-Experiment	Trait 1	$q^2_{(T1)}$	CI	Trait 2	$q^2_{(T2)}$	CI	Trait 3	$q^2_{(T3)}$	CI	Response Scale
R7—Importance to immigration	Qualification for immigration: speak country's official language	.67	[.63,.70]	Qualification for immigration: be white	.66	[.62,.70]	Qualification for immigration: committed to way of life in country	.69	[.64,.73]	11-point, Extremely unimportant/Extremely important[a]
R7—Subjective competence	Able to take active role in political group	.66	[.64,.68]	Confident in own ability to participate in politics	.70	[.67,.73]	Easy to take part in politics	.59	[.57,.62]	11-point, Not at all able/Completely able[a,c]
R7—System responsiveness	Political system allows people to have a say in what government does	.56	[.54,.59]	Political system allows people to have influence on politics	.60	[.58,.62]	Politicians care what people think	.64	[.61,.67]	11-point, Not at all/Completely[a]

Table 1 should be read as follows: in the experiment R1—Political trust, the question measuring Trait 2 "Trust in the legal system" using an 11-point, partially labelled scale, has a measurement quality of 0.81. This means that, on average across country-language groups, 81% of the variance of the observed survey answers to this question is explained by the latent trait "Trust in the legal system" while 19% correspond to measurement errors.

[a] Partially labelled [b] Fully labelled, [c] Verbal labels for the response scales vary across questions because they are tailored to the question stem, e.g. in the experiment R7-Subjective competence the response scale is "not at all able/completely able" for the first question but "not at all confident/completely confident" for the second question.

be attributed to differences in the response scales, since they use the same ones. Overall, the absolute average difference in quality between questions within the same experiment is 0.07. This suggests that some traits were more difficult to measure with accuracy than others, even if the same method is used.

The average measurement quality for each question can also be classified according to the thresholds defined in Section 4.4: the average quality is "good" ($0.8 \leq q^2 < 0.9$) for 7% of the questions, "acceptable" ($0.7 \leq q^2 < 0.8$) for 27%, "questionable" ($0.6 \leq q^2 < 0.7$) for 33%, "poor" ($0.5 \leq q^2 < 0.6$) for 28%, and "unacceptable" ($q^2 < 0.5$) for 5%. None can be classified as "excellent" ($q^2 > 0.9$). The four average measurement qualities classified as "unacceptable" correspond to questions within the experiments Gender inequalities and Evaluation of doctors, from round 2 (two questions in each experiment). Hence, our recommendation is for researchers who use these questions to be extremely careful in their conclusions, due to serious concerns with their measurement quality. On the other extreme, questions within the experiments Political trust or Evaluation of immigration and Immigration perceptions have, in general, the highest measurement quality. Generally, the lower the average quality, the more careful researchers should be with their conclusions, especially regarding effect sizes.

Some of the traits were measured in several rounds (all the traits from the Political satisfaction and Evaluation of immigration experiments, and some of the traits of the Social trust, Political trust and Left-Right orientation experiments), providing some information about the evolution of measurement quality across time. Overall, the quality does not seem to change much across rounds: the differences between qualities of the same traits asked in different rounds are small (around 0.05 or less) in most cases. The main exception is the Evaluation of immigration experiments, where differences between the same traits measured in rounds 3 and 6 are around 0.15 to 0.21. However, some caution is needed in interpreting these estimates, since they might be related to differences in the countries or the other methods analysed in each round, rather than only related to time.

5.3 Measurement quality per country-language group: average across all questions

Next, Table 2 shows the average measurement quality across all questions, per country-language group.

Average qualities across all rounds range from 0.52 (Hungary; classified as "poor") to 0.76 (Cyprus; classified as "acceptable"). Thus, on average, for the analyzed experiments, the level of measurement errors varies across country-language groups (with a maximum difference of 0.24). However, most country-language groups do not diverge much from the overall mean. This may suggest that country-language groups are overall generally comparable, although the situation may differ when looking at specific questions or rounds. So, for cross-country-language group comparisons, researchers should not only test for measurement equivalence but also consider accounting for measurement error before testing (Pirralha & Weber, 2020). Information from Online Appendix 1 can be used for this purpose. Generally, the lower the measurement quality, the more careful researchers need to be in their conclusions for a given country-language group, since higher levels of measurement errors are more likely to disturb the results.

However, one problem with these results is that some countries participated in fewer rounds than others. Thus, fewer questions (and hence, fewer methods) were analysed for those countries. Therefore, we also compared the average quality for each country-language group within each round in which it participated (193 estimates in total). The difference between the country-language group with the smallest and the one with the largest measurement quality in a given round varies from 0.15 (round 1, between Belgium/Greece and France/Finland and round 7, between Portugal and Norway/Germany/Lithuania) to 0.39 (round 4, between Bulgaria and Israel-Arabic). More specifically, the average measurement quality of 61% of the country-language groups falls within the same category as the overall mean ($0.6 \leq q^2 < 0.7$; classified as "questionable"), while 23% are above (19% are "acceptable", i.e., $0.7 \leq q^2 < 0.8$; and 4% are "good"; i.e., $0.8 \leq q^2 < 0.9$) and 17% below the overall mean (16% are classified as "poor", i.e., $0.5 \leq q^2 < 0.6$, while only 1% is classified as unacceptable, $q^2 < 0.5$). This suggests that it might be better to look at each round to assess the comparability of different countries.

Lastly, differences in measurement quality due to the use of different languages seem to be small, on average. In the seven countries in which we could analyse two separate languages, the differences in measurement quality across languages range from 0 in Switzerland to 0.10 in Luxembourg. However, they are bigger in some rounds (e.g. in round 5, there is a difference of 0.12 points between Belgium-Dutch and Belgium-French). This suggests that differences in measurement quality across languages may interact with the specific methods and/or topics used, although it is unclear to what extent they also simply reflect estimation uncertainty.

6 Conclusions

The main goal of this paper was to provide an overview of the measurement quality (defined as the proportion of the variance of the observed survey answers explained by the latent trait of interest) of 67 questions included in the main questionnaire of the ESS. To do so, we analysed 28 MTMM experiments from the seven first rounds in up to 41 country-language groups, using the EUPD procedure (Saris & Satorra, 2018). Thus, we used data from a large academic probability-based survey and applied a new estimation pro-

Table 2
Measurement quality (q2) per country-language group: average across all questions, with 95% confidence intervals [in brackets]

Country-Language	Round 1		Round 2		Round 3		Round 4		Round 5		Round 6		Round 7		Average	
	q^2	CI	q^2	CI	q^2	CI	q^2	CI	q^2	CI	q^2	CI	q^2	CI	q^2	CI
Cyprus-Greek	-	-	-	-	.72	[.60,.84]	.77	[.69,.86]	.83	[.77,.90]	.75	[.70,.81]	-	-	.76	[.73,.80]
Greece-Greek	.77	[.69,.84]	.70	[.59,.81]	-	-	.72	[.64,.79]	.85	[.81,.89]	-	-	-	-	.75	[.71,.79]
Bulgaria-Bulgarian	-	-	-	-	.62	[.53,.72]	.89	[.86,.93]	.74	[.67,.80]	.70	[.64,.75]	-	-	.73	[.68,.78]
Croatia-Croatia	-	-	-	-	-	-	.64	[.53,.75]	.80	[.75,.86]	-	-	-	-	.70	[.62,.78]
Sweden-Swedish	.63	[.59,.68]	-	-	-	-	.73	[.66,.79]	.82	[.75,.89]	.70	[.61,.78]	.68	[.60,.77]	.69	[.66,.72]
Iceland-Icelandic	-	-	-	-	-	-	-	-	-	-	.69	[.63,.76]	-	-	.69	[.63,.76]
Portugal-Portugal	.75	[.69,.80]	.66	[.57,.75]	.61	[.52,.71]	.70	[.64,.76]	-	-	.69	[.64,.73]	.55	[.47,.62]	.67	[.64,.70]
Spain-Spanish	.72	[.66,.77]	.60	[.52,.67]	.67	[.57,.77]	.70	[.63,.76]	.80	[.65,.94]	.65	[.59,.71]	.61	[.55,.67]	.67	[.64,.70]
Luxembourg-French	-	-	.67	[.55,.78]	-	-	-	-	-	-	-	-	-	-	.67	[.55,.78]
Austria-German	.71	[.64,.78]	.64	[.57,.71]	.63	[.55,.72]	-	-	-	-	-	-	.65	[.59,.71]	.66	[.63,.70]
Denmark-Danish	.66	[.60,.73]	.62	[.51,.73]	.66	[.57,.76]	.64	[.56,.72]	.75	[.63,.87]	.68	[.59,.77]	.67	[.63,.72]	.66	[.63,.69]
Finland-Finish	.62	[.57,.68]	.60	[.49,.72]	.65	[.57,.74]	.68	[.58,.71]	.79	[.71,.87]	.69	[.63,.76]	.67	[.61,.73]	.66	[.63,.69]
Germany-German	.66	[.60,.72]	.64	[.56,.73]	.62	[.49,.75]	.65	[.58,.71]	.79	[.66,.92]	.64	[.55,.73]	.70	[.64,.76]	.66	[.63,.69]
Great Britain-English	.67	[.61,.73]	.60	[.50,.71]	.65	[.59,.72]	.66	[.60,.72]	.74	[.61,.87]	.67	[.60,.73]	.64	[.58,.69]	.65	[.63,.68]
Israel-Hebrew	.68[a]	[.60,.76]	-	-	-	-	.68	[.62,.74]	.55	[.51,.59]	.67	[.62,.71]	.65	[.60,.70]	.65	[.62,.68]
Norway-Norwegian	.63	[.56,.70]	.58	[.45,.70]	.64	[.54,.75]	.64	[.54,.74]	.78	[.67,.89]	.67	[.61,.72]	.70	[.62,.78]	.65	[.62,.68]
Belgium-Dutch	.77[a]	[.69,.84]	.61	[.52,.70]	.61	[.51,.70]	.66	[.57,.74]	.76	[.68,.85]	.65	[.60,.69]	.67	[.62,.72]	.65	[.62,.68]
Russia-Russian	-	-	-	-	.57	[.49,.66]	.70	[.63,.77]	.77	[.64,.90]	.61	[.53,.70]	-	-	.65	[.60,.69]
Switzerland-French	.68[a]	[.61,.75]	.60	[.46,.73]	.65	[.52,.78]	.62	[.55,.69]	.67	[.56,.79]	.64	[.58,.71]	.67	[.60,.75]	.64	[.61,.68]
Switzerland-German	.68[a]	[.61,.75]	.61	[.51,.71]	.66	[.56,.76]	.67	[.61,.73]	.69	[.57,.82]	.64	[.55,.72]	.63	[.59,.66]	.64	[.61,.68]
Czechia-Czech	.69	[.62,.75]	.56	[.47,.65]	-	-	.68	[.65,.72]	.64	[.55,.72]	.64	[.56,.72]	.61	[.57,.65]	.64	[.61,.67]
France-French	.62	[.62,.75]	.58	[.50,.65]	.67	[.57,.77]	.69	[.61,.77]	.76	[.63,.89]	.63	[.56,.71]	.63	[.57,.68]	.64	[.61,.67]
Netherlands-Dutch	.69	[.64,.74]	.60	[.47,.72]	.57	[.49,.64]	.67	[.60,.74]	.68	[.52,.84]	.66	[.59,.74]	.63	[.54,.71]	.64	[.61,.67]

Continues on next page

Continued from last page

Country-Language	Round 1		Round 2		Round 3		Round 4		Round 5		Round 6		Round 7		Average	
	q^2	CI	q^2	CI	q^2	CI	q^2	CI	q^2	CI	q^2	CI	q^2	CI	q^2	CI
Israel-Arabic	.68[a]	[.60,.76]	-	-	-	-	.50	[.42,.59]	.86	[.77,.96]	.61	[.52,.70]	.67	[.61,.74]	.64	[.59,.70]
Latvia-Russian	-	-	-	-	.58	[.50,.66]	.71	[.64,.79]	-	-	-	-	-	-	.64	[.58,.70]
Poland-Poland	.70	[.64,.77]	.57	[.48,.66]	.58	[.47,.68]	.65	[.56,.73]	.75	[.66,.84]	.61	[.54,.68]	.61	[.55,.67]	.63	[.60,.67]
Estonia-Estonian	-	-	.62	[.53,.71]	.62	[.56,.68]	.66	[.58,.73]	-	-	.63	[.57,.70]	.65	[.61,.68]	.63	[.60,.66]
Slovenia-Slovenian	.65	[.59,.71]	.58	[.49,.68]	.56	[.46,.66]	.67	[.61,.74]	.75	[.66,.84]	.63	[.58,.68]	.63	[.54,.72]	.63	[.60,.66]
Lithuania-Lithuanian	-	-	-	-	-	-	-	-	.52	[.45,.59]	.64	[.59,.70]	.70	[.64,.75]	.63	[.59,.67]
Belgium-French	.77[a]	[.69,.84]	.60	[.50,.70]	.61	[.52,.69]	.64	[.54,.74]	.64	[.47,.81]	.69	[.63,.76]	.61	[.54,.69]	.63	[.59,.67]
Ukraine-Ukrainian	-	-	.55	[.44,.65]	.62	[.53,.71]	.72	[.64,.80]	.79	[.73,.85]	.56	[.46,.67]	-	-	.63	[.58,.65]
Estonia-Russian	-	-	.68	[.58,.77]	.53	[.44,.62]	.65	[.57,.74]	-	-	.60	[.54,.65]	.60	[.53,.67]	.61	[.58,.65]
Turkey-Turkish	-	-	.61	[.51,.71]	-	-	.62	[.52,.72]	-	-	-	-	-	-	.61	[.55,.68]
Italy-Italian	-	-	.57	[.47,.67]	-	-	-	-	-	-	.65	[.56,.75]	-	-	.61	[.54,.67]
Slovakia-Slovakian	-	-	.56	[.46,.67]	.54	[.45,.62]	.72	[.67,.77]	.69	[.64,.73]	.58	[.50,.66]	-	-	.60	[.56,.64]
Ireland-English	.64	[.55,.73]	.55	[.43,.67]	.53	[.44,.62]	-	-	.60	[.46,.72]	.66	[.59,.74]	.62	[.58,.66]	.60	[.56,.63]
Latvia-Latvia	-	-	-	-	.51	[.40,.62]	.71	[.66,.77]	-	-	-	-	-	-	.60	[.53,.68]
Ukraine-Russian	-	-	.54	[.42,.66]	.55	[.45,.64]	.66	[.60,.72]	.70	[.66,.74]	.61	[.52,.70]	-	-	.59	[.55,.64]
Romania-Romanian	-	-	-	-	.59	[.48,.70]	.56	[.47,.64]	-	-	-	-	-	-	.58	[.51,.65]
Luxembourg-Luxembourgish	-	-	.57	[.45,.69]	-	-	-	-	-	-	-	-	-	-	.57	[.45,.55]
Hungary-Hungarian	-	-	-	-	-	-	-	-	-	-	.49	[.44,.54]	.56	[.51,.61]	.52	[.48,.55]
Mean quality		.68		.60		.61		.68		.74		.64		.64		-
Maximum quality		.77		.70		.72		.89		.86		.75		.70		-
Minimum quality		.62		.54		.51		.50		.52		.49		.55		-
Maximum − Minimum		.15		.16		.21		.39		.34		.26		.15		-

[a] Country could not be split by language in round 1. Therefore, for round 1, the results of the analyses with mixed languages are shown (same for both language groups). For these country-language groups, results for round 1 are not taken into account when calculating country-language average across rounds. Additionally, results in round 1 include Catalan for Spain, Swedish for Finland and Russian for Israel.

cedure to reduce the occurrence of non-convergent and improper solutions.

6.1 Results

Overall, we found that the data from the 67 questions from the ESS main questionnaire has an average quality of 0.65. The quality varies across questions and country-language groups. Questions within the experiments Gender inequalities and Evaluation of doctors present the lowest measurement quality, whereas questions within the experiments Political trust, Evaluation of immigration and Immigration perceptions generally have the highest measurement quality. In terms of country-language groups, the lowest average measurement quality was found for Hungary, while the highest was found for Cyprus. Generally, the lower the average quality, the more careful researchers should be with their conclusions, especially regarding parameter sizes.

6.2 Limitations

This study estimates measurement quality using the True Score model. As such, the acceptance of its results depends on the acceptability of the theoretical model itself and its assumptions, which might not always hold. Some of the assumptions of the Base Model (see Section 4.1) can be relaxed when testing the model (e.g. allow a different effect for a given method on the different traits), while for others, this is not the case. However, there are situations where we can expect violations of these assumptions. For instance, the independence of within-individual observations might be violated if there are memory effects. While Van Meurs and Saris (1990) results suggest that memory effects are not present anymore after 20-minutes, other authors (e.g. Schwarz et al., 2020; Revilla & Höhne, 2021) still found memory effects after 20 minutes. Others (e.g. Alwin, 2011; Krosnick, 2011) even argue that they cannot be completely ruled out even with much longer time periods. In addition, some systematic sources of measurement error cannot be detected due to the design of the experiments (Cernat & Oberski, 2019), as order effects (due to the fixed order in which the methods were implemented), learning or fatigue effects (Batista-Foguet, Revilla, Saris, Boyatzis, & Serlavós, 2014; Krosnick, 2011), or systematic errors that are constant across methods (i.e., if social desirability or acquiescence is constant across methods, these are undetected; however, both social desirability and acquiescence can, at least theoretically, occasionally vary with the methods). Finally, the treatment of variables as interval-measurement instead of ordinal/categorical affects quality estimates, especially for the response scales with smaller numbers of categories, which may have implications also concerning the assumptions of linearity of the relationships of the model (for details, see Coenders & Saris, 1995; Oberski, Saris, & Hagenaars, 2010).

Additionally, it is important to note that due to the high number of estimates, we had to aggregate the results to discuss them. This means that often some aspects varied across our comparison groups (both for questions and rounds, e.g. the methods and the number of countries changed), making it hard to interpret some of the observed differences in measurement quality. However, interested readers can use the data in Online Appendix 1 to achieve additional comparisons.

Lastly, although we followed the more recent recommendations both for the estimation and the testing procedure, there is always some unavoidable subjectivity in these procedures. Moreover, while the EUPD works better than alternative procedures,

1. we still found some non-convergent and improper solutions and
2. there is still uncertainty in the measurement quality estimates, among other reasons due to random sampling variability, which is difficult to account for.

All the previously mentioned issues are not only specific to the present study but they could affect the results.

Furthermore, it is worth noting that this subset of questions is not a random subset of all ESS questions, but only those selected for the MTMM experiments (e.g. there are no estimates for sociodemographic or background variables). We do not know to what extent these results hold for other questions within the ESS, and more generally, we do not know if these results hold for other surveys, countries, or modes of data collection. More sophisticated techniques, such as meta-analyses, would be required to make inferences to other questions.

6.3 Discussion and implications

Our results show that measurement errors can explain a large proportion of the variance of single questions. Even if this is slightly higher than what could be expected based on previous research (e.g. Alwin, 2007), this points towards the general difficulties of collecting data of high quality in surveys, particularly for opinions and attitudes, even when abiding to the highest methodological standards. Furthermore, in line with previous research, our results suggest that measurement errors are not homogenous but vary depending on several aspects.

Differences across questions could be linked to several features, including the response scales, the question wording and/or topic. There is abundant evidence of the impact of formal characteristics and/or response scales on measurement quality (see e.g. the literature review by DeCastellarnau, 2018). Regarding topics, one possible reason for the lower quality of some questions, even when keeping the methods constant, is that these questions ask about topics that are less central in respondents' minds. Therefore, at least some respondents may have either weak or non-existent opinions or

attitudes. This could increase both random and systematic measurement errors.

Differences across countries, even when keeping the method and/or topic constant, could also be linked to several aspects. A proposed explanation for cross-country differences especially in response rates lies in differences in "survey climate" across countries (Smith, 2007, p. 48). Similarly, some of the differences in measurement quality observed across countries could be linked to the "survey climate" (e.g. survey fatigue because some populations might be "oversurveyed", distrust in surveys, higher concerns about confidentiality), but also to cultural reasons (e.g. varying degrees of topic centrality across countries, weaker opinions, lower willingness to disclose personal opinions to strangers) or to survey procedures (e.g. differences in the interview process across countries that may remain even after standardization; Smith (2007, p.48), among other factors. Additionally, these differences could be linked to linguistic/translation issues (e.g. the questions being longer, less natural or more difficult to understand in a given language, or translations that are not functionally equivalent[7](see e.g. Oberski et al., 2007).

Further research using more sophisticated statistical analyses is needed both to allow the possibility of more general inferences and to help understand the exact reasons behind variations in the size of measurement errors across questions and country-language groups, such as the ones outlined above. Particularly, it would be interesting to study how different aspects of the survey items that researchers can control (e.g. the question wording, the response scale, or the layout) affect measurement quality. This could help finding the best measurement instruments to reduce measurement errors in future surveys.

Overall, the results highlight the omnipresence of measurement errors: on average 35% of the variance in observed answers is due to measurement errors. The size of the reliability and validity suggests that, of those, around one third are due to method effects but the remaining two thirds are random error. Generally, random errors decrease the observed relationships between variables while systematic errors can either decrease or increase the observed relationships between variables. The results also show that the size of these errors varies across traits, methods, and country-language groups. Thus, even when measured with the same scale, comparisons across traits and country-language groups are not always straightforward. Only when the size of measurement errors is similar between the groups one wants to compare, standardized relationships (e.g. correlations) can be compared. Furthermore, the information provided in this study can be used to correct for measurement errors (Saris & Gallhofer, 2014; Saris & Revilla, 2016). With a proper application of the information of this and other studies (e.g. avoiding response scales that resulted in lower quality), we believe that reductions in the size of measurement errors can be achieved.

7 Acknowledgements

We want to thank Oriol Serra for his help with data analyses and Jorge Cimentada for his help with data preparation. We thank Willem Saris and the ESS ERIC CST for their continuous support for this line of research. Finally, we thank the editor and the reviewer for their very helpful comments on a previous draft of this paper.

References

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. Hoboken: John Wiley & Sons.

Alwin, D. F. (2011). Evaluating the reliability and validity of survey interview data using the MTMM approach. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), *Question evaluation methods: Contributing to the science of data quality* (pp. 263–293). Hoboken: Wiley Online Library.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public opinion quarterly*, *48*(2), 409–442.

Batista-Foguet, J. M., Revilla, M., Saris, W. E., Boyatzis, R., & Serlavós, R. (2014). Reassessing the effect of survey characteristics on common method bias in emotional and social intelligence competencies assessment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 596–607.

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *Bmj*, *314*(7080), 572.

Bosch, O. J., Revilla, M., DeCastellarnau, A., & Weber, W. (2019). Measurement reliability, validity, and quality of slider versus radio button scales in an online probability-based panel in Norway. *Social Science Computer Review*, *37*(1), 119–132.

Cernat, A., & Oberski, D. (2019). Extending the within-persons experimental design: The multitrait-multierror (MTME) approach. In P. Lavrakas, M. Traugott, C. Kennedy, A. Holbrook, E. de Leeuw, & B. West (Eds.), (pp. 481–500). Hoboken: Wiley Online Library.

Coenders, G., & Saris, W. E. (1995). Categorization and measurement quality. The choice between Pearson and polychoric correlations. *The Multitrait-Multimethod approach to evaluate measurement instruments*, 125–144.

[7]Some translation problems were reported, at least in the first rounds of the ESS, in Saris et al. (2011).

Coromina, L., & Coenders, G. (2006). Reliability and validity of egocentered network data collected via web: A meta-analysis of multilevel multitrait multimethod studies. *Social networks*, *28*(3), 209–231.

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality & quantity*, *52*(4), 1523–1559.

DeCastellarnau, A., & Revilla, M. (2017). Two approaches to evaluate measurement quality in online surveys: An application using the Norwegian Citizen Panel. *Survey Research Methods*, *11*(4), 415–433.

ESS. (2003). European Social Survey (ESS), Round 1—2002. NSD—Norwegian Centre for Research Data. doi:10.21338/nsd-ess1-2002

ESS. (2005). European Social Survey (ESS), Round 2—2004. NSD—Norwegian Centre for Research Data. doi:10.21338/nsd-ess2-2004

ESS. (2007). European Social Survey (ESS), Round 3—2006. NSD—Norwegian Centre for Research Data. doi:10.21338/nsd-ess3-2006

ESS. (2009). European Social Survey (ESS), Round 4—2008. NSD—Norwegian Centre for Research Data. doi:10.21338/nsd-ess4-2008

ESS. (2011). European Social Survey (ESS), Round 5—2010. NSD—Norwegian Centre for Research Data. doi:10.21338/nsd-ess5-2010

ESS. (2013). European Social Survey (ESS), Round 6—2012. NSD—Norwegian Centre for Research Data. doi:10.21338/nsd-ess6-2012

ESS. (2015). European Social Survey (ESS), Round 7—2014. doi:10.21338/nsd-ess7-2014

Joreskog, K., & Sorbom, D. (2005). *LISREL 8.72*. Chicago: Scientific Software International.

Költringer, R. (1995). Measurement quality in Austrian personal interview surveys. In *The multitrait-multimethod approach to evaluate measurement instruments* (pp. 207–225). Budapest: Eötvös Univ. Press.

Krosnick, J. A. (2011). Experiments for evaluating survey questions. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), *Question evaluation methods: Contributing to the science of data quality* (pp. 213–238). doi:10.1002/9781118037003.ch14

Mingwei, P. (2015). Rating scale validation: An MTMM approach. In *Nonverbal delivery in speaking assessment. From an argument to a rating scale formulation and validation* (pp. 119–214). doi:10.1007/978-981-10-0170-3_7

Oberski, D., Saris, W. E., & Hagenaars, J. A. (2010). Categorization errors and differences in the quality of questions across countries. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, … T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 435–453). Wiley.

Oberski, D., Saris, W. E., & Hagenaars, J. A. (2007). Why are there differences in the quality of questions across countries? In G. Loosveldt, M. Swyngedouw, & B. Cambre (Eds.), *Measuring meaningful data in social research* (pp. 281–299). doi:10.1002/9780470609927.ch23

Pirralha, A., & Weber, W. (2020). Correction for measurement error in invariance testing: An illustration using SQP. *Plos one*, *15*(10), e0239421.

R Core Team. (2019). R: A language and environment for statistical computing (version 3.6. 1)[computer software]. R Foundation for Statistical Computing.

Revilla, M., & Höhne, J. K. (2021). Repeatedly measuring political interest: Can we reduce respondent's recall ability and memory effects in surveys using memory interference tasks? *International Journal of Public Opinion Research*, *33*(3), 678–689.

Revilla, M., & 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., Poses, C., Serra, O., Asensio, M., Schwarz, H., & Weber, W. (2021). 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*, *28*(3), 463–474.

Revilla, M., & Saris, W. E. (2013). The split-ballot multitrait-multimethod approach: Implementation and problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 27–46.

Revilla, M., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree-disagree scales. *Sociological Methods & Research*, *43*(1), 73–97.

Revilla, M., Saris, W. E., Loewe, G., & Ochoa, C. (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.

Revilla, M., Zavala-Rojas, D., & Saris, W. E. (2016). Creating a good question: How to use cumulative experience. In C. Wolf, D. J. Joye, T. W. Smith, & Y.-C. Fu (Eds.), *The sage-handbook of survey methodology* (pp. 236–254). Sage.

Rodgers, W. L., Andrews, F. M., & Regula Herzog, A. (1992). Quality of survey measures: A structural modeling approach. *Journal of Official Statistics*, *8*, 251–251.

Santos, J. R. A. (1999). Cronbach's alpha: A tool for assessing the reliability of scales. *Journal of extension*, *37*(2), 1–5.

Saris, W. E. (2013). The prediction of question quality: The SQP 2.0 software. In B. Kleiner, I. Renschler, B. Wernli, P. Farago, & D. Joye (Eds.), *Understanding research infrastructures in the social sciences* (pp. 135–144). Zurich: Seismo Press.

Saris, W. E., & Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. S. (Eds.), *Measurement errors in surveys* (pp. 575–597). doi:10.1002/9781118150382.ch28

Saris, W. E., & Gallhofer, I. N. (2014). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken: John Wiley & Sons.

Saris, W. E., Oberski, D., Revilla, M., Zavala-Rojas, D., Lilleoja, L., Gallhofer, I. N., & Gruner, T. (2011). The development of the program SQP 2.0 for the prediction of the quality of survey questions. RECSM Working paper 24.

Saris, W. E., & Revilla, M. (2016). Correction for measurement errors in survey research: Necessary and possible. *Social Indicators Research*, *127*(3), 1005–1020.

Saris, W. E., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, *4*(1), 61–79. doi:10.18148/srm/2010.v4i1.2682

Saris, W. E., & Satorra, A. (2018). The pooled data approach for the estimation of split-ballot multitrait–multimethod experiments. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 659–672.

Saris, W. E., & Satorra, A. (2019). Comparing BSEM and EUPD estimates for two-group sb-mtmm experiments. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 745–749.

Saris, W. E., Satorra, A., & Coenders, G. (2004). A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design. *Sociological methodology*, *34*(1), 311–347.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*(4), 561–582.

Saris, W. E., van der Veld, W., & Gallhofer, I. (2000). *A program for prediction of the quality of survey measurement*. Paper resented in October 2000 at the methodology conference in Köln, Germany.

Scherpenzeel, A. (2009). *Online interviews and data quality: A multitrait-multimethod study*. Arbeitspapier, CentERdata, Tilburg University, Niederlande.

Scherpenzeel, A. C., & Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods & Research*, *25*(3), 341–383.

Schwarz, H., Revilla, M., & Weber, W. (2020). Memory effects in repeated survey questions: Reviving the empirical investigation of the independent measurements assumption. *Survey Research Methods*, *14*(3), 325–344.

Smith, T. W. (2007). Survey non-response procedures in cross-national perspective: The 2005 ISSP non-response survey. *Survey Research Methods*, *1*(1), 45–54.

Tourangeau, R. (2020). Survey reliability: Models, methods, and findings. *Journal of Survey Statistics and Methodology*. doi:10.1093/jssam/smaa021

Van der Veld, W. M., Saris, W. E., & Satorra, A. (2008). Judgement rule aid for structural equation models. (version 3.0.4 beta). Computer Software. doi:10.13140/RG.2.2.32873.75362

Van Meurs, A., & Saris, W. E. (1990). Memory effects in MTMM studies. In A. Munnich & W. E. Saris (Eds.), *Multitrait multimethod approach to evaluate measurement instruments* (pp. 134–146). Budapest: Eotvos University Press.

Weber, W., Gallhofer, I. N., & Saris, W. E. (2020). Survey questions, design of. In P. Atkinson, S. Delamont, A. Cernat, J. Sakshaug, & R. Williams (Eds.), *Sage research methods foundations*. doi:/10.4135/9781526421036889614

Zavala-Rojas, D. (2016). Measurement equivalence in multilingual comparative survey research. PhD thesis Universitat Pompeu Fabra. Retrieved from http://hdl.handle.net/10803/399146

Appendix A
Tables

(Appendix tables follow on next page)

Table A1
List of experiments and name of all items in these experiments.

Round-Experiment	Trait 1	Trait 2	Trait 3
R1 - Political orientation	The less government intervenes in economy, the better for country (ginveco, test16, test34) [B43, H16, H34]	Government should reduce differences in income levels (gincdif, test17, test35) [B44, H17, H35]	Employees need strong trade unions to protect work conditions/wages (needtru, test18, test36) [B45, H18, H36]
R1 - Media use	TV watching, total time on average weekday (tvtot, test1, test19) [A1, H1, H19]	Radio listening, total time on average weekday (rdtot, test2, test20) [A3, H2, H20]	Newspaper reading, total time on average weekday (nwsptot, test3, test21) [A5, H3, H21]
R1 - Political efficacy	Politics too complicated to understand (polcmpl, test4, test22) [B2, H4, H22]	Could take an active role in a group involved with political issues (polactiv, test5, test23) [B3, H5, H23]	Making mind up about political issues (poldcs, test6, test24) [B4, H6, H24]
R1 - Political satisfaction	How satisfied with present state of economy in country (stfeco, test7, test25) [B30, H7, H25]	How satisfied with the national government (stfgov, test8, test26) [B31, H8, H26]	How satisfied with the way democracy works in country (stfdem, test9, test27) [B32, H9, H27]
R1 - Political trust	Trust in country's parliament (trstprl, test13, test31) [B7, H13, H31]	Trust in the legal system (trstlgl, test14, test32) [B8, H14, H32]	Trust in the police (trstplc, test15, test33) [B9, H15, H33]
R1 - Social trust	Most people can be trusted or you can't be too careful (ppltrst, test10, test28) [A8, H10, H28]	Most people try to take advantage of you, or try to be fair (pplfair, test11, test29) [A9, H11, H29]	Most of the time people helpful or mostly looking out for themselves (pplhlp, test12, test30) [A10, H12, H30]
R2 - Current job	Current job: Variety in work (vrtywrk, testa19, testa32) [G64, I19, I32]	Current job: Job is secure (jbscr, testa20, testa33) [G66, I20, I33]	Current job: health/safety at risk because of work (hlthrwk, testa21, testa34) [G70, I21, I34]
R2 - Evaluation of doctors	Doctors keep whole truth from patients (dckptrt, testa5, testa28) [D25, I5, I28]	Regular general practitioner/-doctor treat patients as equals (dctreql, testa6, testa29) [D26, I6, I29]	Doctors discuss treatment with patient before they decide (dcdisc, testa7, testa30) [D27, I7, I30]
R2 - Gender inequalities	Women should be prepared to cut down on paid work for sake of family (wmcpwrk, testa8, testa22) [G6, I8, I22]	Men should take as much responsibility as women for home and children (mnrsphm, testa9, testa23) [G7, I9, I23]	Men should have more right to job than women when jobs are scarce (mnrgtjb, testa10, testa24) [G8, I10, I24]
R2 - Political satisfaction	How satisfied with present state of economy in country (stfeco, testa11, testa35) [B25, I11, I35]	How satisfied with the national government (stfgov, testa12, testa36) [B26, I12, I36]	How satisfied with the way democracy works in country (stfdem, testa13, testa37) [B27, I13, I37]
R2 - Political trust	Trust in country's parliament (trstprl, testa25, testa38) [B4, I25, I38]	Trust in the legal system (trstlgl, testa26, testa39) [B5, I26, I39]	Trust in politicians (trstplt, testa27, testa40) [B7, I27, I40]
R3 - Well-being	Love learning new things (lrnnew, testb7, testb19, testb31) [E26, H7, H19, H31]	Feel accomplishment from what I do (accdng, testb8, testb20, testb32) [E27, H8, H20, H32]	Like planning and preparing for future (plprftr, testb9, testb21, testb33) [E28, H9, H21, H33]
R3 - Evaluation of immigration	Immigration bad or good for country's economy (imbgeco, testb4, testb16, testb28) [B38, H4, H16, H28]	Country's cultural life undermined or enriched by immigrants (imueclt, testb5, testb17, testb29) [B39, H5, H17, H29]	Immigrants make country worse or better place to live (imwbcnt, testb6, testb18, testb30) [B40, H6, H18, H30]
R3 - Immigrations perception	Allow many/few immigrants of same race/ethnic group as majority (imsmetn, testb1, testb13, testb25) [B35, H1, H13, H25]	Allow many/few immigrants of different race/ethnic group from majority (imdfetn, testb2, testb14, testb26) [B36, H2, H14, H26]	Allow many/few immigrants from poorer countries outside Europe (impcntr, testb3, testb15, testb27) [B37, H3, H15, H27]

Continues on next page

Continued from last page

R3 - Life satisfaction	Feel what I do in life is valuable and worthwhile (dngval, testb10, testb22, testb34) [E40, H10, H22, H34]	There are people in my life who care about me (ppllfcr, testb11, testb23, testb35) [E43, H11, H23, H35]	Feel close to the people in local area (flclpla, testb12, testb24, testb36) [E45, H12, H24, H36]
R4 - Political orientation	Government should reduce differences in income levels (gincdif, testc10, testc31) [B30, H10, H31]	Gays and lesbians free to live life as they wish (freehms, testc11, testc32) [B31, H11, H32]	-
R4 - Political satisfaction	How satisfied with present state of economy in country (stfeco, testc7, testc19) [B25, H7, H19]	How satisfied with the national government (stfgov, testc8, testc20) [B26, H8, H20]	How satisfied with the way democracy works in country (stfdem, testc9, testc21) [B27, H9, H21]
R4 - Political trust	Trust in country's parliament (trstprl, testc16, testc28) [B4, H16, H28]	Trust in the legal system (trstlgl, testc17, testc29) [B5, H17, H29]	Trust in the police (trstplc, testc18, testc30) [B6, H18, H30]
R4 - Social trust	Most people can be trusted or you can't be too careful (ppltrst, testc4, testc25) [A8, H4, H25]	Most people try to take advantage of you, or try to be fair (pplfair, testc5, testc26) [A9, H5, H26]	-
R5 - Effectiveness of the police	How likely be caught if made exaggerated or false insurance claim (insclct, testd10, testd19) [D4, I10, I19]	How likely to be caught if bought something that might be stolen (bystlct, testd11, testd20) [D5, I11, I20]	How likely to be caught if committed traffic offence (trfoct, testd12, testd21) [D6, I12, I21]
R5 - Satisfaction with the police	How successful police are at preventing crimes in country (plcpvcr, testd4, testd13) [D12, I4, I13]	How successful police are at catching house burglars in country (plccbrg, testd5, testd14) [D13, I5, I14]	How quickly would police arrive at a violent crime scene near to where you live (plcarcr, testd6, testd15) [D14, I6, I15]
R6 - Evaluation of democracy	In country opposition parties are free to criticise the government (oppcrgvc, teste7, teste16) [E20, I7, I16]	In country the media are free to criticise the government (medcrgvc, teste8, teste17) [E21, I8, I17]	In country the media provide citizens with reliable information to judge the gov (meprinfc, teste9, teste18) [E22, I9, I18]
R6 - Evaluation of immigration	Immigration bad or good for country's economy (imbgeco, teste19, teste28) [B32, I19, I28]	Country's cultural life undermined or enriched by immigrants (imueclt, teste20, teste29) [B33, I20, I29]	Immigrants make country worse or better place to live (imwbcnt, teste21, teste30) [B34, I21, I30]
R6 - Everyday of life engagement	Interested in what you are doing, how much of the time (tmimdng, teste1, teste10, teste22, teste31) [D31, I1, I10, I22, I31]	Absorbed in what you are doing, how much of the time (tmabdng, teste2, teste11, teste23, teste32) [D32, I2, I11, I23, I32]	Enthusiastic about what you are doing, how much of the time (tmendng, teste3, teste12, teste24, teste33) [D33, I3, I12, I24, I33]
R6 - Feelings past week	Felt depressed, how often past week (fltdpr, teste4, teste13, teste25, teste34) [D5, I4, I13, I25, I34]	Sleep was restless, how often past week (slprl, teste5, teste14, teste26, teste35) [D7, I5, I14, I26, I35]	Felt lonely, how often past week (fltlnl, teste6, teste15, teste27, teste36) [D9, I6, I15, I27, I36]
R7 - Importance to immigration	Qualification for immigration: speak country's official language (qfimlng, testf1, testf10) [D2, I1, I10]	Qualification for immigration: be white (qfimwht, testf2, testf11) [D4, I2, I11]	Qualification for immigration: committed to way of life in country (qfimcmt, testf3, testf12) [D6, I3, I12]
R7 - Subjective competence	Able to take active role in political group (actrolg, testf7, testf16) [B1b, I7, I16]	Confident in own ability to participate in politics (cptppol, testf8, testf17) [B1d, I8, I17]	Easy to take part in politics (etapapl, testf9, testf18) [B1f, I9, I18]
R7 - System responsiveness	Political system allows people to have a say in what government does (psppsgv, testf4, testf13) [B1a, I4, I13]	Political system allows people to have influence on politics (psppipl, testf5, testf14) [B1c, I5, I14]	Politicians care what people think (ptcpplt, testf6, testf15) [B1e, I6, I15]

Note: Name of the variable in the ESS databases displayed in parentheses "()". Name of the variables in the ESS questionnaires displayed in brackets "[]". Only results from the main questionnaire are reported in the paper.

Table A2
Languages excluded from analysis because of having less than 70 observations for a given group

Round	Country	Language excluded
2, 3, 4, 5, 6, 7	Switzerland	Italian
2, 3, 4, 5, 6, 7	Spain	Catalan
2, 3, 4, 5, 6, 7	Finland	Swedish
2	Luxembourg	English, German and Portuguese
2, 3, 4, 5, 6	Slovakia	Hungarian
4, 5, 6, 7	Israel	Russian
4	Turkey	Kurdish
5, 6, 7	Lithuania	Russian
5, 6, 7	Norway	English
7	Finland	English

Table A3
Country-language groups excluded in a given round and experiment because of having correlations of opposite sign

Round	Experiment	Country-language groups excluded
1	Political orientation	Austria, Portugal, France
2	Evaluation of doctors	Portugal, Ukraine-Russian, Slovenia-Slovene, Ukraine-Ukrainian, Turkey, Switzerland-French, Belgium-French, France, Italy
2	Gender inequalities	Portugal
4	Left-right placement	Portugal
4	Political orientation	Israel-Hebrew, Latvia-Latvian, Slovenia, Turkey, Latvia-Russian, Norway, Romania, Switzerland-French, Portugal, Sweden, Denmark, Spain, Greece, Estonia-Russian, Russian Federation, Cyprus, Switzerland-German, Finland, Belgium-Dutch, Great Britain, France, Netherlands, Israel-Arabic
7	Importance to immigration	Switzerland-French, Netherlands, Israel-Arabic

Note: The focus was on the correlations for the same trait measured with different methods

Appendix B
Code example

Example of LISREL input for the pooled data analyses in the case of a split-ballot two-group design with three traits and three methods

```
!Pooled data Group 1
da ng=2 ni=9 no=18696 ma=cm
km file=group1.corr
mean file=group1.mean
sd  file=group1.sd

model ny=9 ne=9 nk=6 ly=fu,fi te=di,fi ps=di,fi be=fu,fi
     ga=fu,fi ph=sy,fi
va 1 ly 1 1 ly 2 2 ly 3 3 ly 4 4 ly 5 5 ly 6 6
fr te 1 1 te 2 2 te 3 3 te 4 4 te 5 5 te 6 6
va 1 te 7 7 te 8 8 te 9 9
va 0 ly 7 7 ly 8 8 ly 9 9
va 1 ga 1 1 ga 2 2 ga 3 3
fr ga 4 1 ga 5 2 ga 6 3 ga 7 1 ga 8 2 ga 9 3
va 1 ga 1 4 ga 4 5 ga 7 6 ga 2 4 ga 5 5 ga 8 6 ga 3 4 ga
     6 5 ga 9 6
fr ph 1 1 ph 2 2 ph 3 3 ph 2 1 ph 3 1 ph 3 2 ph 4 4 ph 5
     5 ph 6 6
out iter =2000 ns adm=off all sc mi

!Pooled data Group 2
da ni=9 no=17983 ma=cm
km file=group2.corr
mean file=group2.mean
sd file=group2.sd

model  ny=9 ne=9 nk=6 ly=fu,fi te=di,fi ps=in be=in ga=in
      ph=in
value 1 ly 1 1  ly 2 2  ly 3 3  ly 7 7  ly 8 8  ly 9 9
eq  te 1 1 1  te 1 1
eq  te 1 2 2  te 2 2
eq  te 1 3 3  te 3 3
fr  te 7 7  te 8 8  te 9 9
va 1 te 4 4  te 5 5  te 6 6
va 0 ly 4 4 \ ly 5 5 \ ly 6 6
pd
out  iter =2000 ns adm=off  all sc mi
```

Example of LISREL input for a country-language group analysis in the case of a split-ballot two-group design with three traits and three methods

```
!Country 1 group 1
Data ng=2  ni=9 no=1225 ma=cm
km file=group1.corr
mean file=group1.mean
sd file=group1.sd

model ny=9 ne=9 nk=6  ly=fu,fi te=di,fi ps=sy,fi be=fu,fi
      ga=fu,fi ph=sy,fi

va 1 ly 1 1 ly 2 2 ly 3 3 ly 4 4 ly 5 5 ly 6 6
fr te 1 1 te 2 2 te 3 3 te 4 4 te 5 5 te 6 6
va 1 te 7 7 te 8 8 te 9 9
va 0 ly 7 7 ly 8 8 ly 9 9
va 1 ga 1 1 ga 2 2 ga 3 3
va 1 ga 1 4 ga 2 4  ga 3 4 ga 4 5 ga 5 5  ga 6 5 ga 7 6
     ga 8 6 ga 9 6
fr ph 1 1 ph 2 2 ph 3 3 ph 2 1 ph 3 1 ph 3 2 ph 4 4  ph 5
      5 ph 6 6

!fix gammas traits using pooled data estimates
va .25 ga 4 1 ga 5 2
va .22 ga 6 3
va .45 ga 7 1
va .43 ga 8 2
va .41 ga 9 3

out iter= 2000 adm=off sc ec mi

!Country 1 group 2
Data ni=9 no=920 ma=cm
km file=group2.corr
mean file=group2.mean
sd file=group2.sd

model ny=9 ne=9 nk=6  ly=fu,fi te=di,fi ps=in be=in ga=in
      ph=in

va 1  ly 1 1 ly 2 2 ly 3 3 ly 7 7 ly 8 8 ly 9 9
eq te 1 1 1 te 1 1
eq te 1 2 2 te 2 2
eq te 1 3 3 te 3 3
fr te 7 7 te 8 8 te 9 9
va 1 te 4 4 te 5 5 te 6 6
va 0 ly 4 4 ly 5 5 ly 6 6
pd

out iter= 2000 adm=off sc ec mi
```

Appendix C
Figures

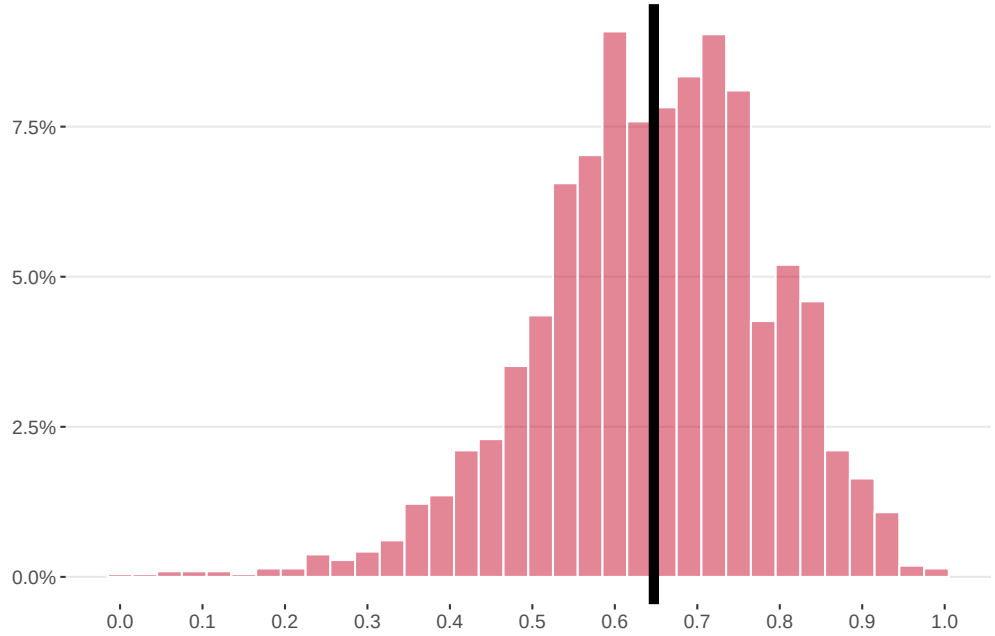

Figure C1. Histogram of measurement quality estimates. Mean indicated by the black vertical line.