doi:10.5477/cis/reis.175.3

The Quality of Survey Questions in Spain: A Cross-National Comparison

La calidad de las preguntas de encuesta en España: una comparación transnacional

Oriol J. Bosch and Melanie Revilla

Key words

Data Quality
- Measurement Errors
- MultiTrait-MultiMethod Experiment
- Cross-National Research
- Survey Methodology

Abstract

Most social research collects data about abstract concepts (e.g., attitudes) using survey questions. However, survey data suffer from measurement errors that affect substantive conclusions. When measurement errors differ across countries, cross-national comparisons of standardized relationships can result in incorrect substantive conclusions. However, no research has analysed the measurement quality of survey questions in Spain in a comparative perspective. Using a Split-Ballot Multitrait-Multimethod experiment conducted in the European Social Survey round 8, we compare the quality of questions in Spain with their quality in other participating countries. The average measurement quality in Spain is higher than the overall average for all ESS countries. In addition, when comparing Spain with other countries, substantive conclusions can be incorrect if differences in the size of measurement errors are not taken into account.

Palabras clave

Calidad de los datos
- Errores de medición
- Experimento MultiRasgo-MultiMétodo
- Investigación transnacional
- Metodología de encuestas

Resumen

La mayoría de la investigación social estudia conceptos abstractos (p. ej., actitudes) mediante preguntas de encuestas. No obstante, las encuestas adolecen de errores de medición que afectan a las conclusiones sustantivas. Cuando dichos errores difieren entre países, comparar relaciones estadísticas estandarizadas entre países puede resultar en conclusiones incorrectas. Sin embargo, la calidad de medición de las preguntas de encuestas en España no ha sido investigada de forma comparada. Utilizando un experimento MultiRasgo-MultiMétodo, realizado en la Encuesta Social Europea (ESS), comparamos la calidad de las preguntas en España con la de otros países. En general, la calidad de medición en España es superior a la mayoría de países participantes. Además, si no se tienen en cuenta los errores de medición al comparar España con otros países, las conclusiones sustantivas pueden ser erróneas.

Citation

Bosch, Oriol J. and Revilla, Melanie (2021). "The Quality of Survey Questions in Spain: A Cross-National Comparison". *Revista Española de Investigaciones Sociológicas*, 175: 3-26. (http://dx.doi.org/10.5477/cis/reis.175.3)

Oriol J. Bosch: The London School of Economics and Political Science and Research and Expertise Centre for Survey Methodology (RECSM) - Universitat Pompeu Fabra | o.bosch-jover@lse.ac.uk

Melanie Revilla: Research and Expertise Centre for Survey Methodology (RECSM) - Universitat Pompeu Fabra | melanie.revilla@upf.edu

Introduction[1]

Most social research requires collecting data about abstract concepts such as attitudes, feelings or opinions. These concepts, corresponding to mental representations that are not directly observable, are usually operationalized by specifying empirical indicators, the most common ones being survey questions (Saris and Gallhofer, 2014).

Properly operationalizing these concepts involves designing questions that maximize the strength of the relationship between the latent concept researchers want to measure (for example, happiness, *F*) and the observed indicators (responses to the questions, *Y*). The strength of the relationship between *F* and *Y* when standardized is referred to as measurement quality (q^2) and can be computed as the product of reliability (r^2) and validity (v^2) (Saris and Andrews, 1991). Reliability represents the strength of the relationship between the observed responses (*Y*) and the true score (*T*), i.e., the true value for a survey question and scale if no random errors would have occurred when answering. Validity represents the strength of the relationship between the latent concept of interest (*F*) and the true score (*T*) of a given question. Measurement quality take values from 0 to 1.

Ideally, measurement quality should be equal to one (the question measures the concept of interest perfectly). However, in practice, survey data suffer from random and systematic measurement errors, which are a counterpart to measurement quality and thus can be computed as $1 - q^2$.

Alwin (2007) suggests that 50% of the variance (i.e., the spread or variability of the distribution) of the observed variables in surveys is due to measurement errors. Thus, there are large differences between the variable researchers want to measure (*F*) and the one that is actually measured by the question (*Y*).

The size of measurement errors depends on how survey questions are designed (e.g., exact formulation or response scales), the language and country where the survey is being administered (Liao, Saris and Zavala-Rojas, 2019), the mode of data collection and, for online surveys, the type of device used to answer (Bosch *et al.*, 2019). This, in turn, can have serious impact on the conclusions drawn from the research. Saris and Gallhofer (2007) illustrated this point using data from an experiment conducted in the European Social Survey (ESS) round 1 in Great Britain: while the correlation between interpersonal trust and trust in the parliament measured using a four-point scale was negative and significant (–0.15), when using an 11-point scale the same correlation was positive and significant (0.29). In addition, both correlations contain measurement errors. In order to know what the true correlation between interpersonal trust and trust in the parliament is, it is necessary to obtain information about the size of measurement errors for the different scales to correct for these errors (Saris and Gallhofer, 2014). However, Saris and Revilla (2016) found that for a number of important social science and marketing journals only 9% of the studies using survey data corrected for measurement errors.

When conducting cross-national research, measurement errors can also affect the comparability of results across countries. When measurement errors vary across countries, cross-national comparisons of standardized relationships can result in incorrect substantive conclusions (Saris and Revilla, 2016). According to Saris and Gallhofer (2007), the main characteristics of questions that can vary across countries and, consequently, provoke differences in measurement quality are: 1) linguistic char-

[1] Acknowledgements: We thank the ESS CST for their continuous support for this line of research. This research was funded by the ESS ERIC Work Programme 1 June 2017 - 31 May 2019.

acteristics, 2) levels of social desirability, and 3) the centrality associated with the question. Regarding linguistic differences, languages have different structures that can lead to different levels of measurement quality across countries even when questions are properly translated (Zavala-Rojas, 2016). In addition, social desirability, i.e., respondents' tendency to answer in a way they consider to be more socially acceptable than their "true" answer (DeMaio, 1984), shows systematic cross-cultural differences (Johnson and Vijver, 2003), being higher in collectivist societies in particular. Finally, question topics can have different levels of importance or be more or less present in public debate, meaning that their centrality (or saliency), i.e., the degree to which the topic of any question resonates with the respondent and the amount of information available, can also vary across countries (Couper and Leeuw, 2003).

This earlier research suggests that differences in the size of measurement errors are expected across countries and can affect cross-national comparisons. However, very few studies have explored cross-national differences in measurement errors.

This article contributes in several ways to this very limited literature on differences in the size of measurement errors across countries. First, we focus on cross-national comparisons between Spain and other European countries. As of April 2019, Spain was a fixed or rotating participant country in at least 21 active cross-national surveys based on samples of individuals or private households framing total adult national populations (GESIS, 2019). Moreover, as of August 2019, Spain was the country with the 5th most registered users of ESS data and the 6th in terms of data downloads. Additionally, 77 scientific publications have used ESS data from Spain until August 2019 (ESS, 2019). Abundant cross-national research has been done using those data.

However, there is evidence that Spain might differ in terms of survey data quality compared to other European countries. For instance, response rates (which are often used as an indicator of data quality) declined or stagnated in most participating countries from ESS rounds 1 to 7 while they increased in Spain (Beullens *et al.*, 2018). Other common indicators of data quality such as response styles, especially acquiescence and extreme response style, were found to be more present in Mediterranean countries like Spain than in other European countries such as Germany and Great Britain (Herk, Poortinga and Verhallen, 2004). This could be related to the fact that acquiescence increases when collectivism and corruption levels are higher at the country-level (Rammstedt, Danner and Bosnjak, 2017), Spain presenting moderate levels of collectivism (Beilmann, Kööts-Ausmees and Realo, 2018; Leung *et al*., 1992) and perception of corruption (Transparency International, 2019). Furthermore, considering that social desirability is higher in collectivist societies (Johnson and Vijver, 2003), this may lead to different levels of measurement errors in Spain than in other European countries. Overall, Spain can be expected to have different data quality than other European countries.

Despite the above, very few studies have analyzed the measurement quality (q^2) of survey questions (as previously defined) in Spain compared to other European countries, with two notable exceptions:

1) Saris *et al.* (2010), using MultiTrait-MultiMethod (MTMM) experiments from the ESS rounds 2 (2004) and 3 (2006), estimated the measurement quality of 12 questions on four topics: "the social distance between doctors and patients", "opinions about work", "opinions about immigration policies" and "opinion about consequences of immigration". They found that overall Spain had a higher measurement quality than the ESS average.

2) Revilla, Saris and Krosnick (2014), using MTMM experiments from the ESS round 3, estimated the quality of 12 questions on four topics: the same previously mentioned topics of "opinion about immigration policies" and "opinion about consequences of immigration", as well as "feelings about life and relationships" and "openness to the future". They found a higher measurement quality in Spain than for the ESS average.

However, both studies are very specific on the type of comparisons that they are interested in (respectively, "agree-disagree" versus item specific scales or number of answer categories in "agree-disagree" scales). In addition, neither study focuses on differences between countries nor on the implications of these differences for cross-national research between Spain and other European countries.

Secondly, we focus on different question characteristics than previous studies (e.g., correspondence between numbers and verbal labels or showing the questions on showcards or not). In this way, we provide useful information to help design different aspects of questionnaires for which previous empirical evidence was missing.

Third, we point out the implications for substantive research of not taking into account measurement errors in cross-national research. While previous research (e.g., Saris and Revilla, 2016) presented a method to correct for measurement errors, practical applications for cross-national research are still limited.

Finally, we provide practical recommendations to researchers and practitioners interested in conducting cross-national research using survey data from Spain. These recommendations are helpful both for researchers designing their own questionnaires, as well as for those using existing survey data, such as the ESS. To do so, we use data from an MTMM experiment about "attitudes towards qualifications for entry or exclusion of immigrants" conducted face-to-face in 23 countries during the ESS round 8 (2016-2017).

Method

The True Score MTMM model

To explore measurement quality in Spain compared to other European countries, we estimate measurement quality using data from an MTMM experiment. The MTMM approach, first introduced by Campbell and Fiske in 1959, consists in the repetition of a set of questions measuring correlated simple latent concepts (e.g., opinions about immigration) called traits (*Fi*) using several methods (*Mj*). In 1971, Jöreskog proposed to treat the MTMM matrices as a Confirmatory Factor Analysis (CFA) model. In 1984, Andrews suggested using the MTMM approach to evaluate the measurement quality of single questions through Structural Equation Models (SEM). He proposed using an additive method effect model. In contrast, Browne (1984) and Cudeck (1989) proposed a multiplicative method effect model. Corten *et al.* (2002) showed that a scale-dependent additive model performs better than four other multiplicative and/or scale-invariant models. Furthermore, Saris and Aalberts (2003) showed that the presence of method effects is a better explanation for correlated disturbance terms in MTMM experiments than relative answers, acquiescence or variation in response functions. Therefore, in this study, we use a scale-dependent additive method effects model. Following Andrews' approach (1984), we consider that the different methods are different answer scales (e.g., 6-point versus 11-point scale) and that the same respondents answer the same questions several times, using the different methods. More precisely, we use the True Score model proposed by Saris and Andrews (1991) that also allows for se-

parate estimates of reliability, validity and method coefficients. This is an advantage since they are often affected differently by the changes in question characteristics.

This True Score model can be summarized by the following system of equations:

$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \quad (1)$$
$$T_{ij} = v_{ij} F_i + m_{ij} M_j \quad (2)$$

where F_i is the i^{th} trait or factor, M_j is the j^{th} method, Y_{ij} is the observed answer for the i^{th} trait and the j^{th} method, T_{ij} is the true score factor or systematic component of the response, r_{ij} is the reliability coefficient (when standardized), v_{ij} is the validity coefficient (when standardized), and e_{ij} is the random error associated with Y_{ij}.

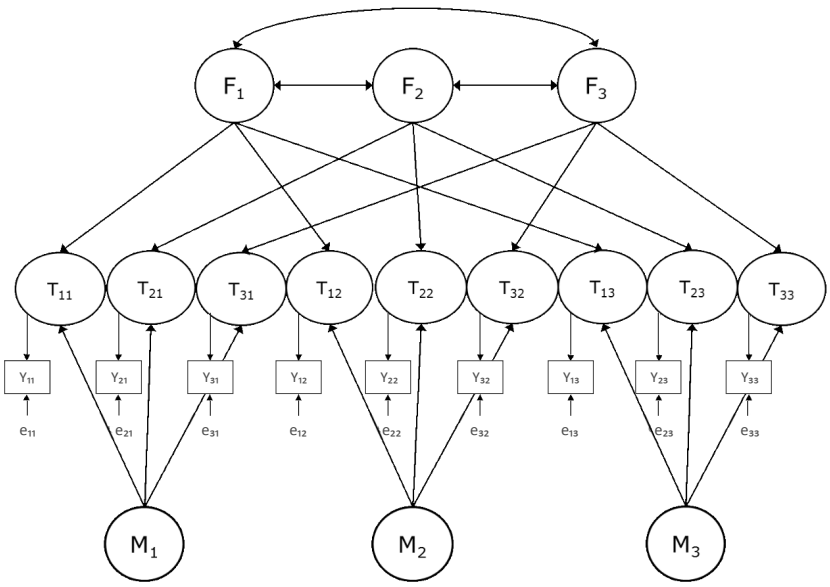
Equation (1) defines each observed variable (Y_{ij}) as the sum of the associated true score (T_{ij}) and random errors (e_{ij}). Equation (2) indicates that each true score (T_{ij}) is itself the sum of the trait component (F_i) and the effect of the method (M_j) used to measure it.

As starting point for this model, we assume that: a) random errors are uncorrelated with each other and with the independent variables in the different equations, b) the traits are correlated, c) the method factors are uncorrelated with each other and with the traits, and d) the impact of the method factor on the traits measured with a common scale is the same. Some of the assumptions made in this base model can be relaxed if needed when testing the model (see section MTMM analyses and testing) until we get a well-fitting final model.

The total measurement quality is obtained by taking the product of the reliability and validity (squared of its coefficients): $q_{ij}^2 = r_{ij}^2 * v_{ij}^2$

For identification purposes, the MTMM model usually repeats three traits, each measured with three methods, resulting in nine observed variables. Thus, each respondent needs to answer the same question three times with different scales. Figure 1 illustrates a True Score MTMM model for three traits and three methods.

FIGURE 1. *True score MTMM model for three traits and three methods*



Source: Own elaboration.

The Split-Ballot MTMM approach

In order to reduce respondents' cognitive burden and possible memory effects due to the repetition of the same questions to the same respondents (Meurs and Saris, 1990), Saris, Satorra and Coenders (2004) proposed to combine the MTMM approach with a Split-Ballot approach (SB), where respondents are randomly assigned to several groups. Each SB group gets a combination of two methods for a given set of three traits, instead of getting three methods. All reliability and validity coefficients can be estimated. The model is still identified under quite general conditions when an SB design is used (Saris, Satorra and Coenders, 2004). It is possible to split the respondents into different numbers of groups, even of unequal sample sizes (Revilla, Bosch and Weber, 2019).

Since nonconvergence problems and improper solutions occur frequently for the two-group design (Revilla and Saris, 2013), in round 8, the ESS implemented a three-group design. Group 1 answered to method 1 (M_1) at time 1 (Section C) and method 2 (M_2) at time 2 (Section I). Group 2 answered to M_2 at time 1 and method 3 (M_3) at time 2. Finally, group 3 answered to M_3 at time 1 and M_1 at time 2. With this design, all possible correlations between methods are observed.

MTMM analyses and testing

The reliability and validity coefficients are estimated through CFA (True Score model presented before) using LISREL 8.72 (Jöreskog and Sörbom, 1996). LISREL uses complex algorithms that minimize the residuals, taking all model restrictions into account. The method used for estimation in each country is Maximum Likelihood for multiple-group analyses (the different groups being the SB groups). We refer to Appendix A for an example of a base LISREL input, and to Hox and Bechger (1998) for an in-depth introduction to SEM.

In order to test if there are misspecifications, we use the JRule software (Veld, Saris and Satorra, 2008) based on the procedure developed by Saris, Satorra and Veld (2009). JRule has the advantage of taking into account the power of the test (i.e., the probability of accepting a false null hypothesis), but also of testing the misspecifications at the parameter level, i.e., testing if each specific parameter is misspecified in contrast to testing the model as a whole.

This leads, in many cases, to the introduction of corrections with respect to the base model presented in equations 1 and 2. Principally, the changes consist in: 1) allowing unequal effects of one method on the true scores corresponding to the different traits, 2) freeing error variances across SB groups to account for the fact that random errors can differ at times 1 and 2 (e.g., because respondents get tired over time), 3) adding a correlation between two method factors with similar scales characteristics, and 4) allowing correlations between error variances because of memory effects. To be able to compare results across countries and languages, we first consider making similar corrections in the different country-language groups. However, this is not always possible. The final model adjustments done in each group are summarized in Appendix B, together with different indicators of model fit.

After conducting the MTMM analyses and testing the models, we compute the measurement quality for the different traits and methods.

Correction for measurement errors

Standardized relationships between observed variables, such as correlations or coefficients of regressions, are affected by measurement errors. For example, Saris and Revilla (2016), using data from the ESS

round 3 from Great Britain, found that the correlation of "allowing more immigrants to come to Great Britain" with the opinion that immigrants make the country a worse place to live changed from –0.27 (no correction) to –0.61 (correction), whereas the correlation with the opinion that immigration is bad for the economy went from –0.13 (no correction) to 0.00 (correction).

To estimate the true relationships (i.e., relationships between the concepts of interest), correction for measurement errors is needed. Furthermore, in a framework of cross-national comparisons, a requirement in order to compare standardized relationships between observed variables across countries is to have similar levels of measurement quality. In other words, if the size of the measurement errors differs across countries, direct comparisons of standardized relationships should not be done without first correcting for measurement errors.

There are different ways to correct for measurement errors (see DeCastellarnau and Saris, 2014; Saris and Gallhofer, 2014). In addition, correction for measurement errors can be done for different types of analyses, including correlations, OLS regressions, SEM, etc. This article focuses on an illustration comparing the correlation between two simple concepts, each measured by one single question[2]. In this case, the researchers observe the correlation between the answers to the two single questions $\rho(Y_1, Y_2)$ but they are interested in the correlation between the latent concepts behind each of the two questions, i.e., the correlation corrected for measurement errors $\rho(F_1, F_2)$. Saris and Gallhofer (2014: 310) provide a formula to correct the correlation between observed variables $\rho(Y_1, Y_2)$ and obtain the correlation between latent variables $\rho(F_1, F_2)$:

$$\rho(F_1, F_2) = [\rho(Y_1, Y_2) - CMV]/q_1\, q_2 \quad (3)$$

where CMV stands for Common Method Variance and is computed as the product of the reliability coefficients (r_i) and the method effects coefficients (m_i) of both observed variables:

$$CMV = r_1\, m_1\, m_2\, r_2 \quad (4)$$

The method effect coefficients can be computed as:

$$m_i = \sqrt{(1 - v_i^2)} \quad (5)$$

Equation 3 states that the correlation between the latent variables can be obtained by subtracting the CMV to the correlation between observed variables $\rho(Y_1, Y_2)$ and then dividing by the product of the measurement quality coefficients of the two questions ($q_1\, q_2$).

CMV is expected when the two observed variables are measured with the same scale, leading to a systematic reaction of the respondents to the scale. For instance, in a scale without a neutral middle point, some respondents with a true neutral position might systematically select the closest option on the positive side whereas others systematically select the closest option on the negative side, and still others systematically skip the question. Therefore, researchers can expect an extra correlation between the observed variables, not linked to the content of the questions themselves but to the systematic reaction of respondents to a shared method.

The measurement quality coefficients of the two questions need to be estimated in a previous step, for instance using MTMM experiments or the Survey Quality Predictor (SQP) 2.1 software (Saris *et al.,* 2011), which semiautomatically generates measurement quality predictions for survey questions using a rich dataset of previous MTMM experiments and random forests algorithms.

[2] For more complex examples, we refer to DeCastellarnau and Saris (2014) and Saris and Revilla (2016).

By comparing correlations without and with correction for measurement errors across a set of different countries (including Spain), we will show how substantive conclusions change when measurement errors are or are not taken into account.

Data

European Social Survey round 8

The ESS (http://www.europeansocialsurvey.org/about/faq.html) is a cross-national survey that has been conducted in Europe every two years since 2001. The ESS conducts face-to-face interviews of around one hour and selects new cross-sectional samples for each round. The questionnaire combines a core section repeated in each round and round-specific rotating modules.

The round 8 fieldwork took place between March 2016 and December 2017 (European Social Survey round 8 Data 2016). Sample sizes range from 880 (Iceland) to 2,852 (Germany), Spain being in the middle (N=1,958, see Appendix C). The response rates for round 8 also vary across countries, ranging from 30.6% (Germany) to 74.4% (Israel), with a response rate of 67.7% in Spain and a mean response rate of 55.4% (ESS, 2017).

The SB-MTMM experiment was conducted in 23 of the participating countries. In multilingual countries, the ESS conducts the surveys in different languages (e.g., Catalan and Spanish in Spain). Since the language can affect measurement quality (Saris and Gallhofer, 2014; Zavala-Rojas, 2016), we analysed each language separately. However, the MTMM model cannot be estimated for languages with a small number of observations (e.g., Catalan in Spain, see Appendix C for a full list). Overall, we analysed 27 country-language groups.

The SB-MTMM experiment

The experiment evaluates three traits, each measured with three methods. The traits aim to measure three aspects of the complex concept "qualification for entry or exclusion of immigrants", respectively: 1) importance of having good educational qualifications, 2) a Christian background[3], and 3) work skills needed in country to qualify for entering the country. Table 1 presents the general wording of each question[4].

Regarding the methods, Table 2 summarizes the characteristics that vary across methods and provides the labels of the endpoints in each scale.

Five aspects varied:

1) The number of answer categories: M_1 is an uneven scale (11-point), while M_2 and M_3 are even scales (10 and 6 points, respectively).
2) Separate questions or battery (i.e., several items sharing the same scale are presented together, the scale being repeated only once): M_1 and M_3 present the questions in battery format whereas M_2 presents them as separate questions.
3) The number of fixed reference points (i.e., answer categories that "set no doubt about the position of the reference point on the subjective scale in the mind of the respondent"; Saris and Gallhofer, 2014: 110); M_1 and M_2 propose two fixed reference points while M_3 proposes only one.
4) The correspondence between the numbers and the verbal labels in the scale (e.g., 0 better represents the idea of "Not at all" than 1): M_1 and M_3 present a high correspondence while M_2 presents a medium correspondence.

[3] Israel changes "Christian" in this item.

[4] For the exact wording in each method, see Appendix D.

5) The presentation of the question on the showcard: usually the ESS showcards (i.e., cards presented to respondents to provide visual help in addition to the interviewer asking the questions) do not contain the questions but only the response options. In this experiment, in M_2 the questions are shown on the showcards, whereas in M_1 and M_3 they are not.

TABLE 1. *Survey questions included in the ESS round 8 SB-MTMM experiment*

Trait	Questions' general wording
Good educational qualifications	How important do you think having good educational qualifications should be in deciding whether someone born, brought up and living outside should be able to come and live here?
Christian background	How important do you think coming from a Christian background should be in deciding whether someone born, brought up and living outside should be able to come and live here?
Work skills needed in the country	How important do you think work skills that [country] needs should be in deciding whether someone born, brought up and living outside should be able to come and live here?

Source: Own elaboration.

TABLE 2. *Variation in the characteristics and labels of the endpoints in each scale*

		M_1	M_2	M_3
Characteristics	No. of points	11	10	6
	Format	Battery	Separate questions	Battery
	No. fixed reference points	2	2	1
	Correspondence	High	Medium	High
	Question in Showcard	No	Yes	No
Label endpoints	First answer category	0 Not at all important	1 Not at all important	0 Not at all important
	Last answer category	10 Extremely important	10 Extremely important	5 Very important

Source: Own elaboration.

Illustration of the implications for substantive cross-national research of not correcting for measurement errors

To illustrate the implication of not correcting for measurement errors in cross-national research, we compare the correlations between the importance given to the religious background of an individual and the importance given to his/her work skills when deciding if someone born, brought up and living outside should be able to come and live in a given country, before and after correction for measurement errors. For the sake of simplicity, we focus on one method in this illustration. We chose M_3 (6-point scale in battery format, with one fixed reference point, high correspondence in the scale, and not providing, for Spain, it presents one of the lowest measurement qualities. Moreover, we illustrate the implications for eight countries: France, Finland, Germany, Italy, Norway, Portugal, Spain and Sweden.

Results

We estimated the measurement quality for 27 country-language groups, three traits and three methods. Since presenting all 243 measurement quality estimates is not practical, first we aggregate all countries and present the results for each trait and method compared to Spain. Then, we aggregate the traits and present the results for each country and method. In that way, we can compare measurement quality, first, across traits and, secondly, across countries. Finally, we present an example of the substantive implications of not correcting for measurement errors in comparing Spain to seven other countries.

Average quality across all country-language groups, per trait and method

Table 3 presents the measurement quality in Spain as well as the average, minimum and maximum quality across the other 26 country-language groups (Spain excluded) for the different traits and methods.

For all countries, traits and methods, the highest quality obtained is 0.99 (Christian background-M_1) whereas the lowest is 0.39 (Work skills-M_1). This means that between 1% (Christian background-M_1) and 61% (Work skills-M_1) of the variance in the observed answers comes from measurement errors.

TABLE 3. *Measurement quality (q^2) in Spain and average, minimum and maximum measurement quality across the other 26 country-language groups, per trait and method*

	Education			Christian			Work skills			Average across traits		
Quality q^2	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
Average 26 groups	0.73	0.64	0.72	0.83	0.71	0.75	0.76	0.68	0.72	0.77	0.68	0.73
Max 26 groups	0.90	0.85	0.87	0.99	0.85	0.92	0.92	0.86	0.85	0.90	0.82	0.85
Min 26 groups	0.56	0.41	0.41	0.41	0.57	0.64	0.39	0.54	0.42	0.53	0.54	0.49
Spain	0.85	0.69	0.64	0.87	0.73	0.70	0.88	0.72	0.64	0.87	0.71	0.66

Note: Quality estimates take values from 0 to 1, with 1 representing a perfect relationship between the observed responses and the latent concept of interest.

Source: Own elaboration.

Furthermore, M_1 has a higher quality on average for all 26 country-language groups and for Spain for all traits. However, there are some differences between Spain and the average from the other 26 groups. First, although M_1 is the method that performs better in both cases, quality estimates in Spain are especially good for all traits. Quality estimates are also higher for M_2 in Spain than the average of all other country-language groups. However, for M_3, the quality in Spain is below the average of all other country-language groups. In addition, M_2 performs better than M_3 in Spain for all traits, whereas for the average of the other country-language groups the tendency is the opposite. Hence, although battery formats can suffer from non-differentiation (Saris and Gallhofer, 2014), in general we would recommend using an 11-point scale presented in a battery format, with two fixed reference points, high correspondence between numbers and verbal labels and no question on the showcard to measure the three indicators studied for the concept "qualification for entry or exclusion of immigrants" instead of the other two methods.

Regarding differences between traits, "Christian background" achieves the highest average measurement quality for all methods for Spain and on average for the other country-language groups. This is interesting since one might think that it would be the trait with the highest propensity to generate social desirability bias, religion being considered a sensitive topic. Finally, differences between traits are consistent for Spain and the average of the other country-language groups. Although Spain presents different quality estimates, the relationship between quality estimates and the traits is similar.

Average quality across all traits, per country-language group and method

Next, differences across countries are analysed in more detail, this time aggregating across traits. Table 4 presents the average quality across all traits, per country-language group and method and the ranking position of each country for each method.

First, measurement quality across countries and methods varies from 0.53 (Estonia-Russian-M_1) to 0.88 (Iceland-M_1). Hence, across all methods and countries, the variance explained by measurement errors ranges from 12% (Iceland-M_1) to 47% (Estonia-Russian-M_1). There are large differences in measurement quality across the different country-language groups. The general trend is that M_1 (11-point scale in battery format, with two fixed reference points, high correspondence between numbers and verbal labels and no question on the showcards) performs better than M_2 and M_3. Moreover, central and northern European countries, overall, present a higher measurement quality than their eastern and southern counterparts.

Comparing Spain to the others, Spain has the 4th highest quality for M_1 and the 10th highest for M_2. However, for M_3, Spain presents the 4th lowest quality. Therefore, important differences across methods exist for Spain and should be considered. First, using an 11-point scale presented in a battery format, with two fixed reference points, high correspondence between numbers and verbal labels and no question on the showcard, works much better in Spain than in most country-language groups. Second, a 6-point scale in battery format, with only one fixed reference point, high correspondence between numbers and verbal labels, and no question on the showcard, performs worse in Spain than for most country-language groups analysed. This suggests that cultural and linguistic differences across countries affect the size of measurement errors associated with different methods.

TABLE 4. *Average quality (q^2) across all traits, per country-language group and method*

	Average quality for the three traits			Ranking		
	M_1	M_2	M_3	M_1	M_2	M_3
Austria	0.74	0.67	0.79	20	14	7
Belgium-Dutch	0.83	0.70	0.74	9	11	13
Belgium-French	0.87	0.58	0.68	5	25	20
CzechRepublic	0.79	0.69	0.69	13	13	18
Estonia-Estonian	0.77	0.58	0.72	16	24	14
Estonia-Russian	0.53	0.63	0.69	27	19	19
Finland	0.90	0.75	0.74	1	5	12
France	0.82	0.54	0.67	11	27	21
Germany	0.77	0.80	0.76	15	3	10
Great Britain	0.72	0.73	0.71	21	8	15
Hungary	0.70	0.61	0.64	22	22	25
Iceland	0.88	0.82	0.85	2	1	1
Ireland	0.83	0.61	0.49	10	23	27
Israel-Arab	0.88	0.80	0.78	3	2	8
Israel-Hebrew	0.76	0.63	0.62	18	20	26
Italy	0.68	0.57	0.84	24	26	2
Lithuania	0.69	0.63	0.81	23	18	5
Netherlands	0.80	0.65	0.77	12	17	9
Norway	0.84	0.74	0.82	7	6	4
Poland	0.75	0.73	0.71	19	9	22
Portugal	0.67	0.61	0.75	25	21	11
Russia	0.66	0.66	0.83	26	15	3
Slovenia	0.79	0.66	0.66	14	16	24
Spain	0.87	0.71	0.66	4	10	23
Sweden	0.83	0.74	0.79	8	7	6
Switzerland-French	0.85	0.69	0.71	6	12	16
Switzerland-German	0.76	0.78	0.70	17	4	17

Source: Own elaboration.

Substantive implications for cross-national research: an illustration

Results demonstrate that there are non-negligible differences between Spain and other country-language groups. This may have important implications in cross-national research when measurement errors are not taken into account.

Table 5 presents the correlations with and without correction for measurement errors for each country, ordered from highest to lowest corrected correlations. It also presents the ranking of each country (1 meaning the highest correlation) for correlations with and without correction for measurement errors.

We can see an increase in correlations when correcting for measurement errors, except for Italy. For Spain the correlation increases from 0.41 to 0.46. However, the change is not homogeneous across countries: thus, for Finland the correlation increases 0.09 points, while for Italy it is reduced by 0.03 points. Consequently, comparing countries using correlations without correction instead of correlations with correction for measurement errors leads to different substantive conclusions. In particular, in terms of ranking, Italy presents the 4th highest correlation without correction for measurement errors, while with correction, it presents the lowest. Therefore, if researchers, for example, want to compare the correlation between "Christian background" and "Work skills needed in the country" for Spain and Italy, not correcting for measurement errors would lead to incorrect conclusions.

TABLE 5. *Correlation coefficients and rankin without and with correction*

	Correlation		Ranking	
	Without correction	With correction	Without correction	With correction
Finland	0.47	0.56	1	1
Sweden	0.41	0.47	2	2
Spain	0.41	0.46	3	3
Norway	0.39	0.44	5	4
Portugal	0.38	0.43	6	5
Germany	0.36	0.42	7	6
France	0.35	0.40	8	7
Italy	0.40	0.37	4	8

Source: Own elaboration.

Discussion and conclusions

Main results

Our main goal was to compare the measurement quality of survey questions in Spain with other European countries, since previous research focusing on Spain in a comparative perspective, even when relevant, is limited.

Overall, for the three traits considered, we found that measurement quality varies greatly across countries, from an average across methods and traits 0.62 in Estonia-Russian to 0.85 in Iceland, meaning that on average between 62% and 85% of the variance in the observed answers is due to the latent concepts of interest, while 15 to 38% is due to measurement errors.

Central and northern European countries present, in general, higher measurement quality. This could be linked to the fact that countries with low collectivism and corruption levels are less prone to social desirability (Rammstedt, Danner and Bosnjak, 2017) and/or to linguistic differences.

Furthermore, the measurement quality of the three traits considered was higher for Spain than for the average of the other 26 country-language groups analysed. These results are in line with previous research also based on ESS data on measurement quality (Saris *et al.*, 2010; Revilla, Saris and Krosnick, 2014).

However, differences across methods appear. M_1 performs especially well for Spain compared with most other countries, ranking 4th of all 27 country-language groups. However, for M_3, Spain presents the 4th lowest quality estimate. Therefore, although overall Spain presents a higher measurement quality, researchers cannot assume a higher than average quality in Spain for any method, but must consider that some methods may perform better in Spain than in other countries, while that may not be the case for other methods.

Moreover, not considering these differences in the size of measurement errors when comparing Spain with other countries affects substantive conclusions. First, in our illustration, the observed correlations were mostly underestimated. Furthermore, the ranking of countries differed when considering the correlations with and without correction. In particular, Spain and Italy presented similar correlations without correction, suggesting that the belief that a Christian background is important for immigrants to qualify to enter the country is similarly correlated with the belief that immigrants need work skills to qualify for both countries. However, the correlation corrected for measurement errors was higher for Spain than for Italy, pointing to a different substantive conclusion: although Spain and Italy have a similar level of religiosity (Evans and Baronavski, 2018), the relationship between Christian background and work skills is stronger for Spain. This difference between the correlations with and without correction for Spain and Italy is mainly related to the difference in measurement quality of both countries (0.18 points lower in Spain than in Italy), and to a lesser extent to the difference in CMV (0.04 higher in Spain than in Italy). After correction using equation 3, Spain presents a higher true correlation than Italy. This indicates that, although Spain presents a higher CMV, the notably lower quality in Spain led to an underestimation of the correlation compared to an overestimation for Italy.

Limitations and further research

These results have some limitations. First, these findings are specific for the topics analysed and the methods used and may not be generalizable to other questions or methods. Second, due to small sample sizes, some languages could not be analyzed. In particular, we were unable to use the Catalan speaking respondents, which did not permit us to compare the quality estimates across languages within Spain. However, considering that other countries present different measurement qualities depending on the language of administration, we could expect the same for Spain. Further research could specifically explore the differences between Catalan and Spanish. In addition, MTMM experiments are not well suited to explain why measurement quality is higher in some countries than others. Further research could focus on finding explanations. Furthermore, we have only illustrated how to correct correlations between two simple concepts for measurement errors. However, correction for measurement errors can be applied to more com-

plex models (e.g., regressions). We refer to DeCastellarnau and Saris (2014), Saris and Gallhofer (2014) and Saris and Revilla (2016) for examples and guidelines about how to do it for other models. Moreover, measurement quality provides information about standardised relationships. Researchers interested in comparing unstandardized relationships should look at the measurement equivalence of constructs across countries (Davidov *et al.*, 2014). Lastly, estimates of the size of measurement errors can also be affected by errors. Thus, even the corrected correlations present some errors.

To be able to draw general conclusions, future research should explore new topics and methods to see if the tendency is the same for different traits and scales. However, conducting MTMM experiments is not always possible. An alternative is to use the SQP software. Using SQP predictions, researchers could get a better picture of the effect of different methods on different questions (DeCastellarnau and Revilla, 2017). In addition, sensitivity checks of these analyses could be conducted using the SQP software in order to explore if predictions and estimates are similar, and if not, how differences affect the corrections for measurement errors.

Practical Recommendations

First, based on our results, to measure the concept "qualification for entry or exclusion of immigrants" in Spain we recommend using M_1 (11-point scale in battery format, with two fixed reference points, high correspondence between numbers and verbal labels and no question on the showcards) instead of M_2 and M_3.

Second, this case study illustrates what previous research had found (e.g., Saris and Gallhofer, 2007), that: 1) substantive researchers need to bear in mind that comparing standardized relationships across countries is only possible if the measurement quality is the same, and 2) even in this case, correcting for measurement errors is necessary to properly estimate the relationships of interest, i.e., the ones between the concepts, and not the ones between observed variables, that are only imperfect measures of the concepts of interest. Therefore, in line with both previous research and the results of this study, we recommend correcting for measurement errors whenever possible.

Bibliography

Alwin, Duane F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken, New Jersey: John Wiley and Sons, Inc.

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

Beilmann, Mai; Kööts-Ausmees, Liisi and Realo, Anu (2018). "The Relationship Between Social Capital and Individualism–Collectivism in Europe". *Social Indicators Research,* 137: 641-664. doi: 10.1007/s11205-017-1614-4

Beullens, Koen; Loosveldt, Geert; Vandenplas, Caroline and Stoop, Ineke (2018). "Response Rates in the European Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts?". *Survey Methods: Insights from the Field*. doi: 10.13094/SMIF-2018-00003

Bosch, Oriol J.; Revilla, Melanie; DeCastellarnau, Anna and Weber, Wiebke (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): 119-32. doi: 10.1177/0894439317750089

Browne, Michael W. (1984). "The Decomposition of Multitrait-multimethod Matrices". *British Journal of Mathematical and Statistical Psychology,* 37(1): 1-21. doi: 10.1111/j.2044-8317.1984.tb00785.x

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

Corten, Irmgard W.; Saris,Willem E.; Coenders,Germà; Veld, William M. van der; Aalberts, Chris E. and Kornelis, Charles (2002). "Fit of Different Models for Multitrait-Multimethod Experiments". *Structural Equation Modeling: A Multidisciplinary Journal,* 9(2): 213-32. doi: 10.1207/S15328007SEM0902_4

Couper, Mick P. and Leeuw, Edith D. de (2003). "Nonresponse in Cross-Cultural and Cross-National Surveys". In: Harness, F.; Vijver, F. van de and Mohler, P. (eds.). *Cross-Cultural Survey Methods*. New York: Wiley.

Cudeck, Robert (1989). "Analysis of Correlation Matrices Using Covariance Structure Models". *Psychological Bulletin,* 105(2): 317-27. doi: 10.1037/0033-2909.105.2.317

Davidov, Eldad; Meuleman, Bart; Cieciuch, Jan; Schmidt, Peter and Billiet, Jaak (2014). "Measurement Equivalence in Cross-National Research". *Annual Review of Sociology,* 40(1): 55-75. doi: 10.1146/annurev-soc-071913-043137

DeCastellarnau, Anna and Saris, Willem E. (2014). "A Simple Way to Correct for Measurement Errors in Survey Research". *European Social Survey Education Net (ESS Edunet). European Social Survey Education Net* (ESS EduNet). 2014. Available at: http://essedunet.nsd.uib.no/cms/topics/measurement/

DeCastellarnau, Anna, and Revilla, Melanie (2017). "Two approaches to evaluate measurement quality in online surveys: An application using the norwegian citizen panel". *Survey Research Methods*, 11(4): 415-433. doi: 10.18148/srm/2017.v11i4.7226

DeMaio, Theresea J. (1984). "Social Desirability in Survey Measurement: A Review". In: Turner, C. F. and Martin, E. (eds.). *Surveying Subjective Phenomena*. New York: Russell Sage.

ESS (2016). *Data File Edition 2.1. NSD —Norweegian Centre for Reseaerch Data, Norway— Data Archive and Distributor of ESS Data for ESSERIC.* Available at: https://www.europeansocialsurvey.org/data/download.html?r=8, access September 21, 2020.

ESS (2017). *ESS8 - 2016 Fieldwork Summary and Deviations.* Available at: https://www.europeansocialsurvey.org/data/deviations_8.html, access September 19, 2020.

ESS (2019). *ESS User Statistics*. Available at: http://www.europeansocialsurvey.org/docs/data_users/ESS_data_user_stats_aug_2019.pdf, access September 19, 2020.

Evans, Jonathan and Baronavski, Chris (2018). *How Do European Countries Differ in Religious Commitment?*. Available at: https://www.pewresearch.org/fact-tank/2018/12/05/how-do-european-countries-differ-in-religious-commitment/, access September 19, 2020.

GESIS (2019). *Overview of Comparative Surveys Worldwide*. Available at: www.gesis.org/ComparativeSurveyOverview, access September 21, 2020.

Herk, Hester van; Poortinga, Ype H. and Verhallen, Theo M. M. (2004). "Response Styles in Rating Scales: Evidence of Method Bias in Data from Six EU Countries". *Journal of Cross-Cultural Psychology,* 35(3): 346-360. doi: 10.1177/0022022104264126

Hox, Joop J. and Bechger, Timo M. (1998). "An Introduction to Structural Equation Modeling". *Family Sicence Review,* 11: 354-373.

Johnson, Timothy P. and Vijver, Fons J. R. van de (2003). "Social Desirability in Cross-Cultural Research". In: Vijver, F. van de; Mohler, P. and Wiley, J. (eds.). *Cross-Cultural Survey Methods*. Hoboken, New Jersey: Wiley-Interscience

Jöreskog, Karl G. and Sörbom, Dag (1996). *LISREL 8: User's Reference Guide*. Uppsala: Scientific Software International.

Leung, Kwok; Au, Yuk-Fai; Fernández-Dols, José M. and Iwawaki, Saburo (1992). "Preference for Methods of Conflict Processing in Two Collectivist Cultures". *International Journal of Psychology*, 27(2): 195-209. doi: 10.1080/00207599208246875

Liao, Pei-Shan; Saris, Willem E. and Zavala-Rojas, Diana (2019). "Cross-National Comparison of Equivalence and Measurement Quality of Response Scales in Denmark and Taiwan". *Journal of Official Statistics,* 35(1): 117-135. doi: 10.2478/jos-2019-0006

Meurs, A. van and Saris, Willem E. (1990). "Memory Effects in MTMM Studies". In: Saris, W. E. and Meurs, A. van (eds.). Evaluation of Measurement Instruments by Meta-Analysis of Multitraitmultimethod Studies. Amsterdam: North-Holland.

Rammstedt, Beatrice; Danner, Daniel and Bosnjak, Michael (2017). "Acquiescence Response Styles: A Multilevel Model Explaining Individual-Level and Country-Level Differences". *Personality and Individual Differences*, 107(1): 190-194. doi: 10.1016/j.paid.2016.11.038

Revilla, Melanie and Saris, Willem E. (2013). "The Split-Ballot Multitrait-Multimethod Approach: Implementation and Problems". *Structural Equa-*

tion Modeling: A Multidisciplinary Journal, 20(1): 27-46. doi: 10.1080/10705511.2013.742379

Revilla, Melanie; Saris, Willem E. and Krosnick, Jon A. (2014). "Choosing the Number of Categories in Agree-Disagree Scales". *Sociological Methods & Research,* 43(1): 73-97. doi: 10.1177/0049124113509605

Revilla, Melanie; Bosch, Oriol J. and Weber, Wiebke (2019). "Unbalanced 3-Group Split-Ballot Multitrait-Multimethod Design?". *Structural Equation Modeling,* 26(3): 437-447. doi: 10.1080/10705511.2018.1536860

Saris, Willem E. and Andrews, Frank M. (1991). "Evaluation of Measurement Instruments Using a Structural Modeling Approach". In: Biemer, P.; Groves, R.; Lyberg, L.; Mathiowetz, N. and Sudman, S. (eds.). *Measurement Errors in Surveys*. New York: John Wiley and Sons, Inc.

Saris, Willem E. and Aalberts, Chris (2003). "Different Explanations for Correlated Disturbance Terms in MTMM Studies". *Structural Equation Modeling: A Multidisciplinary Journal,* 10(2): 193-213. doi: 10.1207/S15328007SEM1002_2

Saris, Willem E. and Gallhofer, Irmtraud N. (2014 [2007]). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, New Jersey: John Wiley and Sons, Inc.

Saris, Willem E. and Revilla, Melanie (2016). "Correction for Measurement Errors in Survey Research: Necessary and Possible". *Social Indicators Research,* 127(3): 1005-1020. doi: 10.1007/s11205-015-1002-x

Saris, Willem E.; Satorra, Albert and Coenders, Germa (2004). "A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design". *Sociological Methodology,* 34(1): 311-347. doi: 10.1111/j.0081-1750.2004.00155.x

Saris, Willem E.; Satorra, Albert and Veld, William M. van der (2009). "Testing Structural Equation Models or Detection of Misspecifications?". *Structural Equation Modeling: A Multidisciplinary Journal,* 16(4): 561-582. doi: 10.1080/10705510903203433

Saris, Willem E.; Revilla, Melanie; Krosnick, Jon A. and Shaeffer, Eric 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, Willem E.; Oberski, Daniel L.; Revilla, Melanie; Zavala-Rojas, Diana; Lilleoja, Laur; Gallhofer, Irmtraud N. and Gruner, Thomas (2011). *The Development of the Program SQP 2.0 for the Prediction of the Quality of Survey Questions*. (RECSM Working Paper). Available at: http://www.upf.edu/survey/_pdf/RECSM_wp024.pdf

Transparency International (2019). *Corruption Perceptions Index 2019.* Available at: https://www.transparency.org/en/cpi/2019, access September 21, 2020.

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

Zavala-Rojas, Diana (2016). *Measurement Equivalence in Multilingual Comparative Survey Research*. Barcelona: Universitat Pompeu Fabra. [Doctoral thesisl].

RECEPTION: November 19, 2019
REVIEW: May 6, 2020
ACCEPTANCE: September 11, 2020

Appendix A: example of initial LISREL input

Analysis Group 1

```
Data ng=3 ni=9 no=221 ma=cm
km file=ATGER-group1.corr
mean file=ATGER-group1.mean
sd file=ATGER-group1.sd
model ny=9 ne=9 nk=6 ly=fu,fi te=sy,fi ps=sy,fi be=fu,fi ga=fu,fi ph=sy,fi
value 1 ly 1 1 ly 2 2 ly 3 3 ly 4 4 ly 5 5 ly 6 6 te 7 7 te 8 8 te 9 9
fr te 1 1 te 2 2 te 3 3 te 4 4 te 5 5 te 6 6
value 0 ly 7 7 ly 8 8 ly 9 9
fr ga 1 1 ga 2 2 ga 3 3 ga 4 1 ga 5 2 ga 6 3 ga 7 1 ga 8 2 ga 9 3
value 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 ph 1 1 ph 2 2 ph 3 3
fr ph 2 1 ph 3 1 ph 3 2 ph 4 4 ph 5 5 ph 6 6
start .5 all
out mi iter= 300 adm=off sc
```

Analysis Group 2

```
Data ni=9 no=219 ma=cm
km file=ATGER-group2.corr
mean file=ATGER-group2.mean
sd file=ATGER-group2.sd
model ny=9 ne=9 nk=6 ly=fu,fi te=sy,fi ps=in be=in ga=in ph=in
fr te 4 4 te 5 5 te 6 6 te 7 7 te 8 8 te 9 9
va 1 ly 4 4 ly 5 5 ly 6 6 ly 7 7 ly 8 8 ly 9 9 te 1 1 te 2 2 te 3 3
equal te 1 4 4 te 4 4
equal te 1 5 5 te 5 5
equal te 1 6 6 te 6 6
value 0 ly 1 1 ly 2 2 ly 3 3
out mi iter= 300 adm=off sc
```

Analysis Group 3

```
Data ni=9 no=228 ma=cm
km file=ATGER-group3.corr
mean file=ATGER-group3.mean
sd file=ATGER-group3.sd
model ny=9 ne=9 nk=6 ly=fu,fi te=sy,fi ps=in be=in ga=in ph=in
fr te 1 1 te 2 2 te 3 3 te 7 7 te 8 8 te 9 9
va 1 ly 1 1 ly 2 2 ly 3 3 ly 7 7 ly 8 8 ly 9 9 te 4 4 te 5 5 te 6 6
equal te 1 1 1 te 1 1
equal te 1 2 2 te 2 2
equal te 1 3 3 te 3 3
equal te 2 7 7 te 7 7
equal te 2 8 8 te 8 8
equal te 2 9 9 te 9 9
value 0 ly 4 4 ly 5 5 ly 6 6
out mi iter= 300 adm=off sc
```

Appendix B: Split Ballot-True Score-MTMM model analysis adjustments, fit and JRule evaluation

Country-Language group	Model adjustments (LISREL notations)	df	$\chi 2$	No. JRule misspecifications
Austria	Free GA 2 4 TE 8 8 TE 5 5	36	78.90	0
Belgium-Dutch	Free TE 1 1	38	92.20	0
Belgium-French	Free GA 8 6	38	60.59	2
Czech Republic	Free TE 7 7 TE 2 2 TE 5 5 GA 5 5	35	86.50	0
Estonia-Estonian	Free TE 6 6 TE 5 5 TE 8 5	36	95.20	0
Estonia-Russian	Free TE 9 9	38	55.87	4
Finland	Free TE 8 8 TE 7 7 TE 4 4 GA 5 5 TE 8 2	34	91.10	2
France	Free TE 4 4 TE 7 7 GA 4 1 TE 3 1	35	77.32	3
Germany	Free TE 7 7 TE 2 2 TE 4 4 TE 6 6 GA 2 4	34	79.69	3
Great Britain	Free TE 4 4 TE 1 1 TE 3 3 TE 6 6 GA 8 6 GA 2 4 GA 6 5	32	83.00	4
Hungary	Free TE 4 4 TE 7 7 TE 5 5 TE 9 9 GA 7 6 (G2)	34	83.27	1
Ireland	Free TE 7 7 TE 8 8 TE 4 4 GA 7 6 GA 8 6 GA 9 6 (G2); fix TE 2 2 (va 0)	34	88.41	2
Israel-Arab	Free TE 8 8 TE 8 8 GA 2 4 PH 5 4 GA 9 3 (G3)	34	41.70	1
Israel-Hebrew	Free TE 1 1 GA 8 6 GA 5 5	36	81.45	2
Iceland	Free TE 4 4 TE 6 6 TE 1 1	36	71.45	0
Italia	Free TE 4 4 TE 7 7 TE 5 5 PH 5 4	37	98.26	1
Lithuania	PH 5 4	38	74.01	0
Netherlands	TE 6 6 TE 4 4 GA 5 5 GA 6 5 PH 5 4	36	58.20	1
Norway	TE 4 4 TE 8 8 TE 7 7 GA 2 4	37	88.08	2
Poland	TE 4 4 TE 1 1 TE 7 7 TE 7 1	37	85.50	0
Portugal	None	39	69.03	0
Russia	TE 4 4 TE 5 5 TE 2 2 TE 1 1 GA 6 5 (G2) PH 5 4	33	83.35	1
Slovenia	TE 1 1	38	60.48	3
Spain	TE 4 4 GA 8 6 PH 6 5	36	92.04	0
Sweden	TE 1 1 TE 7 7 TE 2 2 TE 4 4 GA 5 5	36	68.07	2
Switzerland-French	TE 4 4 TE 6 6 TE 3 3	36	46.95	2
Switzerland-German	TE 4 4 TE 7 7 GA 8 6	36	87.70	2

Source: Own elaboration.

Appendix C: ESS round 8 sample sizes per country-language group

Country-Language	Group 1	Group 2	Group 3	Total cases
Austria	221	219	228	668
Belgium-Dutch	345	365	342	1,052
Belgium-French	236	241	237	714
Czech Republic	763	777	726	2,266
Estonia-Estonian	522	510	520	1,552
Estonia-Russian	169	139	159	467
Finland	608	603	590	1,801
France	685	680	696	2,061
Germany	959	958	935	2,852
Great Britain	662	653	644	1,959
Hungary	527	557	530	1,614
Ireland	913	933	911	2,757
Israel-Arab	175	168	182	525
Israel-Hebrew	660	663	685	2,008
Iceland	295	293	292	880
Italia	935	843	848	2,626
Lithuania	654	662	627	1,943
Netherlands	557	554	570	1,681
Norway	497	563	485	1,545
Poland	591	585	516	1,692
Portugal	408	451	411	1,270
Russia	822	812	796	2,430
Slovenia	447	424	436	1,307
Spain	648	599	590	1,837
Sweden	541	506	504	1,551
Switzerland-French	123	133	129	385
Switzerland-German	361	367	350	1,078

Note: Finland-Swedish, Israel-Russian, Lithuania-Russian, Spain-Catalan and Switzerland-Italian were not analysed because the sample size was < 100 cases per split-ballot group.

Source: Own elaboration.

Appendix D: Questionnaire questions' formulations by method

Introduction (similar in all cases)

People come to live in (country) from other countries for different reasons. Some have ancestral ties. Others come to work here, or to join their families. Others come because they're under threat. Here are some questions about this issue.

Method 1

How important do you think each of these things should be in deciding whether someone born, brought up and living outside (country) should be able to come and live here. Please use this card. Firstly, how important should it be for them to…**Read out…**

		Not at all important									Extremely important	(Refusal)	(Don't know)
C33	… have good educational qualifications?	00	01	02	03	04	05	06	07	08	09	10	77	88
C34	… come from a Christian background?	00	01	02	03	04	05	06	07	08	09	10	77	88
C35	...have work skills that [country] needs?	00	01	02	03	04	05	06	07	08	09	10	77	88

Method 2

How important do you think having good educational qualifications should be in deciding whether someone born, brought up and living outside (country) should be able to come and live here?

Not at all important									Extremely important	(Refusal)	(Don't know)
01	02	03	04	05	06	07	08	09	10	77	88

How important do you think coming from a Christian background should be in deciding whether someone should be able to come and live here?

Not at all important									Extremely important	(Refusal)	(Don't know)
01	02	03	04	05	06	07	08	09	10	77	88

How important do you think having work skills that (country) needs should be in deciding whether someone should be able to come and live here?

Not at all important									Extremely important	(Refusal)	(Don't know)
01	02	03	04	05	06	07	08	09	10	77	88

Method 3

CARD 30 How important do you think each of these things should be in deciding whether someone born, brought up and living outside (country) should be able to come and live here. Firstly, how important should it be for them to…**read out…**

		Not at all important					Very important	(Refusal)	(Don't know)
C39	… have good educational qualifications?	00	01	02	03	04	05	7	8
C40	… come from a Christian background?	00	01	02	03	04	05	7	8
C41	…have work skills that (country) needs?	00	01	02	03	04	05	7	8

doi:10.5477/cis/reis.175.3

La calidad de las preguntas de encuesta en España: una comparación transnacional

The Quality of Survey Questions in Spain: A Cross-National Comparison

Oriol J. Bosch y Melanie Revilla

Palabras clave

Calidad de los datos
- Errores de medición
- Experimento MultiRasgo-MultiMétodo
- Investigación transnacional
- Metodología de encuestas

Resumen

La mayoría de la investigación social estudia conceptos abstractos (p. ej., actitudes) mediante preguntas de encuestas. No obstante, las encuestas adolecen de errores de medición que afectan a las conclusiones sustantivas. Cuando dichos errores difieren entre países, comparar relaciones estadísticas estandarizadas entre países puede resultar en conclusiones incorrectas. Sin embargo, la calidad de medición de las preguntas de encuestas en España no ha sido investigada de forma comparada. Utilizando un experimento MultiRasgo-MultiMétodo, realizado en la Encuesta Social Europea (ESS), comparamos la calidad de las preguntas en España con la de otros países. En general, la calidad de medición en España es superior a la mayoría de países participantes. Además, si no se tienen en cuenta los errores de medición al comparar España con otros países, las conclusiones sustantivas pueden ser erróneas.

Key words

Data Quality
- Measurement Errors
- MultiTrait-MultiMethod Experiment
- Cross-National Research
- Survey Methodology

Abstract

Most social research collects data about abstract concepts (e.g., attitudes) using survey questions. However, survey data suffer from measurement errors that affect substantive conclusions. When measurement errors differ across countries, cross-national comparisons of standardized relationships can result in incorrect substantive conclusions. However, no research has analysed the measurement quality of survey questions in Spain in a comparative perspective. Using a Split-Ballot Multitrait-Multimethod experiment conducted in the European Social Survey round 8, we compare the quality of questions in Spain with their quality in other participating countries. The average measurement quality in Spain is higher than the overall average for all ESS countries. In addition, when comparing Spain with other countries, substantive conclusions can be incorrect if differences in the size of measurement errors are not taken into account.

Cómo citar

Bosch, Oriol J. y Revilla, Melanie (2021). «La calidad de las preguntas de encuesta en España: una comparación transnacional». *Revista Española de Investigaciones Sociológicas*, 175: 3-26. (http://dx.doi.org/10.5477/cis/reis.175.3)

La versión en inglés de este artículo puede consultarse en http://reis.cis.es

Oriol J. Bosch: The London School of Economics and Political Science y Research and Expertise Centre for Survey Methodology (RECSM) - Universitat Pompeu Fabra | o.bosch-jover@lse.ac.uk

Melanie Revilla: Research and Expertise Centre for Survey Methodology (RECSM) - Universitat Pompeu Fabra | melanie.revilla@upf.edu

Introducción[1]

La mayoría de la investigación social requiere de la recopilación de datos sobre conceptos abstractos como actitudes, sentimientos u opiniones. Estos conceptos, que corresponden a representaciones mentales no directamente observables, suelen operacionalizarse mediante indicadores empíricos, siendo las preguntas de encuestas el tipo más común (Saris y Gallhofer, 2014).

Una operacionalización adecuada de estos conceptos implica diseñar preguntas que maximicen la fuerza de la relación estadística entre el concepto latente que los investigadores quieren medir (por ejemplo, felicidad, *F*) y los indicadores observados (respuestas a las preguntas, *Y*). La fuerza de esta relación entre *F* e *Y*, cuando está estandarizada, se llama calidad de medición (q^2) y puede calcularse como el producto de la fiabilidad (r^2) y la validez (v^2) (Saris y Andrews, 1991). La fiabilidad representa la fuerza de la relación entre la respuesta observada (*Y*) y el valor verdadero (*T*), es decir, el valor de una pregunta de encuesta con una escala determinada si no se hubieran producido errores aleatorios al responder. La validez representa la fuerza de la relación estadística entre el concepto de interés latente (*F*) y el valor verdadero (*T*) de una pregunta determinada. La calidad de medición toma valores de 0 a 1.

Idealmente, la calidad de medición debería ser igual a 1 (la pregunta mide perfectamente el concepto de interés). Sin embargo, en la práctica, los datos de las encuestas adolecen de errores de medición aleatorios y sistemáticos, que son el complemento de la calidad de medición y, por lo tanto, se pueden calcular como $1 - q^2$.

Alwin (2007) sugiere que el 50% de la varianza (es decir, la dispersión o variabilidad de la distribución) de las variables observadas en las encuestas se debe a errores de medición. Por tanto, existen grandes diferencias entre la variable que los investigadores quieren medir (*F*) y la que realmente mide la pregunta (*Y*).

El tamaño de estos errores de medición depende de cómo se diseñan las preguntas de encuesta (por ejemplo, formulación exacta o escalas de respuesta), el idioma y el país donde se administra la encuesta (Liao, Saris y Zavala-Rojas, 2019), el modo (cara a cara, teléfono, etc.) de recopilación de datos y, para las encuestas *online*, también el tipo de dispositivo utilizado para responder (Bosch *et al.,* 2019). Esto, a su vez, puede tener serias implicaciones para las conclusiones de la investigación. Saris y Gallhofer (2007) ilustraron este punto utilizando datos de un experimento realizado en la ronda 1 de la Encuesta Social Europea (European Social Survey, ESS) en Gran Bretaña: mientras que la correlación entre la confianza interpersonal y la confianza en el Parlamento medida usando una escala de cuatro puntos era negativa y significativa (–0,15), al usar una escala de 11 puntos la misma correlación era positiva y significativa (0,29). Sin embargo, ambas correlaciones contienen errores de medición. Para saber cuál es la verdadera correlación entre confianza interpersonal y confianza en el Parlamento, es necesario obtener información sobre el tamaño de los errores de medición de las diferentes escalas para corregir estos (Saris y Gallhofer, 2014). Sin embargo, Saris y Revilla (2016) encontraron que, para varias revistas importantes de ciencias sociales y *marketing*, solo el 9% de los estudios que utilizaron datos de encuestas corrigió dichos errores.

[1] Agradecimientos: Queremos agradecer al equipo científico central (Core Scientific Team, CST) de la Encuesta Social Europea por su apoyo continuo a esta línea de investigación. Esta investigación ha sido financiada por el ESS ERIC Work Programme 1 de junio 2017 - 31 de mayo 2019.

Al realizar investigaciones transnacionales, los errores de medición pueden afectar la comparabilidad de los resultados entre países. Cuando los errores de medición varían de un país a otro, las comparaciones de relaciones estandarizadas entre países pueden dar lugar a conclusiones sustantivas erróneas (Saris y Revilla, 2016). Según Saris y Gallhofer (2007), las principales características de una pregunta que pueden variar entre países y, en consecuencia, provocar diferencias en la calidad de medición son: 1) las características lingüísticas, 2) los niveles de deseabilidad social y 3) el nivel de centralidad de dicha pregunta. Con respecto a las diferencias lingüísticas, los idiomas tienen distintas estructuras, lo que puede conducir a diferentes niveles de calidad de medición entre países, incluso si se traducen correctamente (Zavala-Rojas, 2016). Igualmente, la deseabilidad social, es decir, la tendencia de los encuestados a responder de una manera que consideran socialmente más aceptable que su respuesta «verdadera» (DeMaio, 1984), muestra diferencias interculturales sistemáticas (Johnson y Vijver, 2003), siendo más alto, en particular, en sociedades colectivistas. Finalmente, los temas de las preguntas pueden tener diferentes niveles de importancia o estar más o menos presentes en el debate público, lo que significa que su centralidad (o prominencia), es decir, el grado en que el tema de cualquier pregunta resuena con el encuestado y la cantidad de información disponible, también puede variar entre países (Couper y Leeuw, 2003).

La investigación realizada hasta ahora, por ende, sugiere que existen diferencias en el tamaño de los errores de medición entre países y que, dichas diferencias, pueden afectar a las comparaciones entre países. Aun con todo, solo unos pocos estudios han explorado las diferencias transnacionales en cuanto al tamaño de los errores de medición.

Este artículo contribuye de varias formas a enriquecer la escasa literatura existente sobre las diferencias en el tamaño de los errores de medición entre países. En primer lugar, nos centramos en comparar España con otros países europeos. Por un lado, en abril de 2019, España era un participante fijo o rotativo de al menos 21 encuestas transnacionales, activas hoy en día, centradas en muestras —de individuos u hogares privados— de la población general (GESIS, 2019). Asimismo, en agosto de 2019, España era el quinto país con más usuarios registrados utilizando datos de la ESS y el sexto en términos de descargas de datos. Además, 77 publicaciones científicas han utilizado datos de la ESS de España hasta agosto de 2019 (ESS, 2019). Así pues, abundante investigación transnacional se realiza utilizando dichos datos.

Por otro lado, existen evidencias de que España puede diferir en términos de la calidad de los datos de encuestas en comparación con otros países europeos. Por ejemplo, las tasas de respuesta (que a menudo se utilizan como un indicador de la calidad de los datos) disminuyeron o se estancaron en la mayoría de los países participantes de las rondas 1 a 7 de la ESS, mientras que aumentaron en España (Beullens *et al.*, 2018). Para otros indicadores comúnmente utilizados para inferir la calidad de los datos, tales como la aquiescencia y el estilo de respuesta extremo, se ha descubierto que son más presentes en países mediterráneos como España que en otros países europeos como Alemania o Gran Bretaña (Herk, Poortinga y Verhallen, 2004). Esto podría estar relacionado con el hecho de que la aquiescencia aumenta cuando los niveles de colectivismo y corrupción son más elevados en un país (Rammstedt, Danner y Bosnjak, 2017), presentando España niveles moderados de colectivismo (Beilmann, Kööts-Ausmees y Realo, 2018; Leung *et al.*, 1992) y de percepción de la

corrupción (Transparency International, 2019). Además, teniendo en cuenta que la deseabilidad social es mayor en las sociedades colectivistas (Johnson y Vijver, 2003), esto podría dar lugar a niveles de errores de medición en España distintos comparado con otros países europeos. En general, pues, se puede esperar que España muestre una calidad de datos diferente a la de otros países europeos.

Sin embargo, muy pocos estudios han analizado la calidad de medición (q^2) de las preguntas de encuesta (como se definió anteriormente) en España en comparación con otros países europeos, con dos notables excepciones:

1) Saris *et al*. (2010), utilizando experimentos MultiRasgo-MultiMétodo (*MultiTrait-MultiMethod*, MTMM) de las rondas 2 (2004) y 3 (2006) de la ESS, estimaron la calidad de medición de 12 preguntas sobre cuatro temas: «la distancia social entre médicos y pacientes», «opinión sobre el trabajo», «opinión sobre políticas de inmigración» y «opinión sobre las consecuencias de la inmigración». Los autores descubrieron que, en general, España tiene una calidad de medición superior a la media de la ESS.
2) Revilla, Saris y Krosnick (2014), utilizando experimentos MTMM de la ronda 3 de la ESS, estimaron la calidad de 12 preguntas sobre cuatro temas: los mismos temas mencionados anteriormente de «opinión sobre políticas de inmigración» y «opinión sobre consecuencias de la inmigración», así como «sentimientos sobre la vida y las relaciones» y «apertura al futuro». Encontraron una calidad de medición superior en España a la media de los países participantes en dicha ronda de la ESS.

Sin embargo, ambos artículos son muy específicos respecto al tipo de comparaciones que les interesan (respectivamente, escalas «de acuerdo-en desacuerdo» [*agree-disagree scales*] versus escalas específicas [*item-specific scales*] y variaciones en el número de categorías de respuesta en escalas «de acuerdo-en desacuerdo»). Además, ninguno de ellos se centra en las diferencias entre países ni en las implicaciones de estas diferencias para la investigación transnacional, específicamente cuando se trata de comparar España con otros países europeos.

En segundo lugar, nos centramos, a diferencia de anteriores estudios, en distintas características de las preguntas (por ejemplo, el nivel de correspondencia entre los números y las etiquetas verbales o el hecho de mostrar las preguntas en las tarjetas que se facilitan a los participantes). De esta manera, podemos brindar información útil para ayudar a diseñar diferentes aspectos de los cuestionarios, sobre los que aún falta evidencia empírica.

En tercer lugar, ilustramos las implicaciones para la investigación sustantiva (en particular transnacional) de no tener en cuenta los errores de medición. Si bien investigaciones anteriores (por ejemplo, Saris y Revilla, 2016) presentaron un método para corregir los errores de medición, las aplicaciones prácticas para la investigación transnacional aún son escasas.

Por último, proporcionamos recomendaciones prácticas a investigadores y profesionales interesados en realizar investigaciones transnacionales utilizando datos de encuestas de España. Estas recomendaciones son útiles tanto para los investigadores que diseñan sus propios cuestionarios como para aquellos que utilizan datos de encuestas existentes, como la ESS. Para ello, utilizamos datos de un experimento MTMM sobre «actitudes hacia la calificación de entrada o exclusión de inmigrantes» que se realizó mediante entrevistas presenciales en 23 países durante la ronda 8 de la ESS (2016-2017).

Método

El modelo *True Score* MTMM

Para explorar la calidad de medición en España en comparación con otros países europeos, estimamos la calidad de medición utilizando datos de un experimento MTMM. El enfoque MTMM, introducido por primera vez por Campbell y Fiske en 1959, consiste en repetir un conjunto de preguntas que miden conceptos latentes simples correlacionados entre ellos (por ejemplo, opiniones sobre inmigración), llamados rasgos (*Fi*), utilizando varios métodos (*Mj*). En 1971, Jöreskog propuso tratar las matrices MTMM como un modelo de Análisis Factorial Confirmatorio (*Confirmatory Factor Analysis*, CFA). En 1984, Andrews sugirió utilizar el enfoque MTMM para evaluar la calidad de medición de preguntas individuales a través de Modelos de Ecuaciones Estructurales (*Structural Equation Modeling*, SEM), utilizando un modelo en el que los efectos de los métodos se suman (*additive method effect model*). En contraste, Browne (1984) y Cudeck (1989) propusieron un modelo en el que esos efectos se multiplican (*multiplicative method effect model*). Corten *et al*. (2002) mostraron que un modelo aditivo dependiente de la escala (*scale-dependent additive model*) funciona mejor que otros cuatro modelos multiplicativos y/o invariantes de escala (*scale-invariant*). Por otro lado, Saris y Aalberts (2003) demostraron que la presencia de efectos de método es una mejor explicación para los términos perturbativos correlacionados en los experimentos MTMM en comparación con otras posibles explicaciones como las respuestas relativas, la aquiescencia o variaciones en las funciones de respuesta. Por lo tanto, en este estudio utilizamos un modelo en el que los efectos de los métodos se suman y son dependientes de la escala (*scale-dependent additive method effects model*). Siguiendo el enfoque de Andrews (1984), consideramos que cada método corresponde a una escala de respuesta (por ejemplo, escala de 6 puntos o de 11 puntos) y que los mismos encuestados responden a las mismas preguntas varias veces, utilizando los diferentes métodos. Más precisamente, utilizamos el modelo *True Score* («Valor verdadero») propuesto por Saris y Andrews (1991) que además permite estimar por separado los coeficientes de fiabilidad, validez y método. Esto es una ventaja, ya que a menudo se ven afectados de manera diferente por los cambios en las características de la pregunta.

El modelo *True Score* se puede resumir con el siguiente sistema de ecuaciones:

$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \quad (1)$$

$$T_{ij} = v_{ij} F_i + m_{ij} M_j \quad (2)$$

donde F_i es el rasgo o factor i, M_j es el método j, Y_{ij} es la respuesta observada por el rasgo i y el método j, T_{ij} es el componente sistemático de la respuesta por el rasgo i y método j llamado *true score factor*, r_{ij} es el coeficiente de fiabilidad (cuando se estandariza), v_{ij} es el coeficiente de validez (cuando se estandariza), y e_{ij} es el error aleatorio asociado con Y_{ij}.

La ecuación (1) define cada variable observada (Y_{ij}) como la suma de los asociados *true score* (T_{ij}) y los errores aleatorios (e_{ij}). La ecuación (2) indica que cada *true score* (T_{ij}) es en sí mismo la suma del componente del rasgo (F_i) y el efecto de método usado para medirlo (M_j).

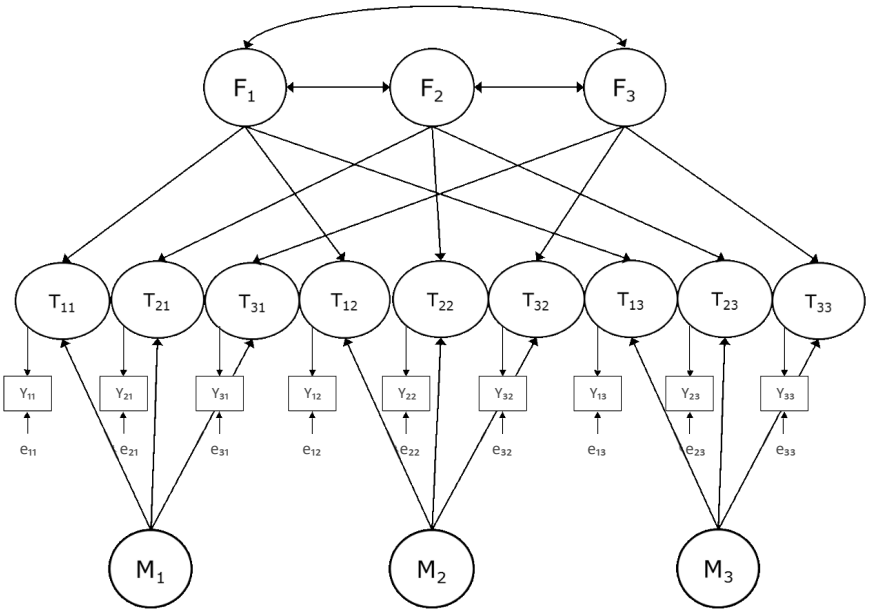
Como punto de partida para este modelo, asumimos que: a) los errores aleatorios no están correlacionados entre sí ni con las variables independientes en las diferentes ecuaciones, b) los rasgos están correlacionados, c) los factores del método no están correlacionados entre ellos ni con los rasgos, y d) el impacto del factor del método sobre los rasgos medidos con una escala común es el mismo. Al testear el

modelo, algunas de las asunciones hechas en este modelo base se pueden relajar si es necesario (ver sección «Análisis y testeo de los MTMM»), hasta que se pueda obtener un modelo final con buen ajuste.

Una vez hecho esto, la calidad total de medición se puede obtener tomando el producto de la fiabilidad y la validez (el cuadrado de sus coeficientes): $q_{ij}^2 = r_{ij}^2 * v_{ij}^2$

En el interés de conseguir un modelo identificado, el modelo MTMM generalmente repite tres rasgos, cada uno medido con tres métodos, resultando en nueve variables observadas. Por tanto, cada encuestado debe responder la misma pregunta tres veces con diferentes escalas. La figura 1 ilustra un modelo *True Score* MTMM para tres rasgos y tres métodos.

FIGURA 1. *Modelo* True Score *MTMM para tres rasgos y tres métodos*



Fuente: Elaboración propia.

El enfoque *Split-Ballot* MTMM

Con el fin de reducir la carga cognitiva de los encuestados y los posibles efectos de memoria debido a la repetición de las mismas preguntas a los mismos encuestados (Meurs y Saris, 1990), Saris, Satorra y Coenders (2004) propusieron combinar el enfoque MTMM con un diseño donde los encuestados se asignan al azar a varios grupos (llamado diseño *Split-Ballot;* SB), los cuales reciben un cuestionario ligeramente distinto al otro. Cada grupo obtiene una combinación de dos métodos para un conjunto dado de tres rasgos, en lugar de obtener tres métodos. Con esto se pueden estimar todos los coeficientes de fiabilidad y validez. El modelo es normalmente identificado en condiciones generales cuando se utiliza un diseño de *Split-Ballot* (Saris, Satorra y Coenders, 2004). Es posible dividir a los encuestados en diferentes números de grupos, incluso con tamaños de muestra desiguales (Revilla, Bosch y Weber, 2019).

Dado que los problemas de no convergencia y valores inválidos ocurren con fre-

cuencia para el diseño de dos grupos (Revilla y Saris, 2013), en la ronda 8, la ESS implementó un diseño de tres grupos. El grupo 1 respondió al método 1 (M_1) en el momento 1 (Sección C) y al método 2 (M_2) en el momento 2 (Sección I). El grupo 2 respondió al M_2 en el momento 1 y al método 3 (M_3) en el momento 2. Finalmente, el grupo 3 respondió al M_3 en el momento 1 y al M_1 en el momento 2. Con este diseño se observan todas las posibles correlaciones entre métodos.

Análisis y testeo de los MTMM

Los coeficientes de fiabilidad y validez se estiman usando CFA (modelo *True Score* presentado anteriormente) y LISREL 8,72 (Jöreskog y Sörbom, 1996). LISREL utiliza algoritmos complejos que minimizan los residuos, teniendo en cuenta todas las restricciones del modelo. El método utilizado para la estimación en cada país es el de Máxima Verosimilitud (*Maximum Likelihood*) para análisis de grupos múltiples (los diferentes grupos son los grupos de *Split-Ballot*). Nos referimos al apéndice A para un ejemplo del código base de LISREL y a Hox y Bechger (1998) para una introducción en profundidad a los Modelos de Ecuaciones Estructurales.

Para probar si hay problemas debidos a especificaciones incorrectas, utilizamos el *software* JRule (Veld, Saris y Satorra, 2008) basado en el procedimiento desarrollado por Saris, Satorra y Veld (2009). JRule tiene la ventaja de tener en cuenta el poder estadístico (es decir, la probabilidad de aceptar una hipótesis nula falsa). También testea las especificaciones incorrectas a nivel de parámetro, es decir, testea si cada parámetro está mal especificado en vez de testear todo el modelo a la vez.

Esto lleva en muchos casos a la introducción de correcciones con respecto a las asunciones del modelo base presentado en las ecuaciones 1 y 2. Principalmente, los cambios consisten en: 1) permitir efectos desiguales de un método sobre los valores verdaderos correspondientes a los diferentes rasgos, 2) liberar las varianzas de los términos de error entre los grupos de *Split-Ballot* para tener en cuenta el hecho de que los errores aleatorios pueden diferir en los momentos 1 y 2 (por ejemplo, porque los encuestados se cansan con el tiempo), 3) agregar una correlación entre dos factores de método con características de escalas similares, y 4) permitir correlaciones entre varianzas de los términos de error debido a efectos de memoria. Para poder comparar los resultados entre países e idiomas, primero consideramos introducir correcciones similares en los diferentes grupos de países e idiomas. Sin embargo, no es siempre posible. Las correcciones finales del modelo realizadas en cada análisis se resumen en el apéndice B, junto con diferentes indicadores del ajuste del modelo.

Después de realizar los análisis y testear los modelos MTMM, calculamos la calidad de medición para los diferentes rasgos y métodos.

Corrección por errores de medición

Las relaciones estadísticas estandarizadas entre variables observadas, como las correlaciones o los coeficientes de regresión, se ven afectadas por los errores de medición. Por ejemplo, Saris y Revilla (2016), utilizando datos de la ronda 3 de la ESS de Gran Bretaña, encontraron que la correlación de «permitir que más inmigrantes vengan a Gran Bretaña» con la opinión de que los inmigrantes hacen del país un lugar peor para vivir cambió de –0,27 (sin corrección) a –0,61 (corrección), mientras que la correlación con la opinión de que la inmigración es mala para la economía pasó de 0,13 (sin corrección) a 0,00 (corrección).

Para estimar las verdaderas relaciones (es decir, las relaciones entre los concep-

tos de interés), es necesario corregir por los errores de medición. En el marco de las comparaciones entre países, asimismo, un requisito para comparar las relaciones estadísticas estandarizadas usando variables observadas es tener niveles similares de calidad de medición en dichos países. Dicho de otra manera, si el tamaño de los errores de medición difiere entre países, no se debería realizar comparaciones directas de relaciones estadísticas estandarizadas sin corregir primero por los errores de medición.

Hay diferentes formas de corregir por los errores de medición (véanse DeCastellarnau y Saris, 2014; Saris y Gallhofer, 2014). La corrección de dichos errores de medición se puede realizar para diferentes tipos de análisis, incluidas correlaciones, regresiones lineales simples, SEM, etc. Este artículo se centra en una ilustración que compara la correlación entre dos conceptos simples, cada uno medido por una sola pregunta[2]. En el caso que ilustramos, imaginemos que unos investigadores observan la correlación entre las respuestas a dos preguntas individuales $\rho(Y_1, Y_2)$, pero están interesados en la correlación entre los conceptos latentes detrás de cada una de las dos preguntas, es decir, la correlación corregida por errores de medición $\rho(F_1, F_2)$. Saris y Gallhofer (2014: 310) proporcionan una fórmula para corregir la correlación entre las variables observadas $\rho(Y_1, Y_2)$ y obtener la correlación entre las variables latentes $\rho(F_1, F_2)$:

$$\rho(F_1, F_2) = [\rho(Y_1, Y_2) - CMV]/q_1\, q_2 \quad (3)$$

donde *CMV* significa Varianza del Método Común (*Common Method Variance*) y se calcula como el producto de los coeficientes de fiabilidad (r_i) y los del efecto de método (m_i) de ambas variables observadas:

$$CMV = r_1\, m_1\, m_2\, r_2 \quad (4)$$

Los coeficientes del efecto de método se pueden calcular como:

$$m_i = \sqrt{(1 - v_i^2)} \quad (5)$$

La ecuación 3 establece que la correlación entre las variables latentes se puede obtener restando el CMV a la correlación entre las variables observadas $\rho(Y_1, Y_2)$ y luego dividir por el producto de los coeficientes de calidad de medición de las dos preguntas ($q_1\, q_2$).

Se espera CMV cuando las dos variables observadas se miden con la misma escala, lo que lleva a una reacción sistemática de los encuestados a la escala. Por ejemplo, en una escala sin un punto medio neutral, algunos encuestados con una verdadera posición neutral pueden seleccionar sistemáticamente la opción más cercana en el lado positivo, mientras que otros seleccionan sistemáticamente la opción más cercana en el lado negativo, y otros se saltan sistemáticamente la pregunta. Por lo tanto, los investigadores pueden esperar una correlación adicional entre las variables observadas, no vinculada al contenido de las preguntas en sí, sino a la reacción sistemática de los encuestados a un método compartido.

Los coeficientes de calidad de medición de las dos preguntas deben estimarse en un paso anterior, por ejemplo, utilizando experimentos MTMM o el *software Survey Quality Predictor* (SQP) 2.1 (Saris *et al.*, 2011), que genera semiautomáticamente predicciones de calidad de medición de preguntas de encuesta utilizando un rico conjunto de datos de experimentos MTMM previos y algoritmos de bosques aleatorios (*random forests*).

Al comparar las correlaciones sin y con corrección por errores de medición en un conjunto de países diferentes (incluido España), mostraremos cómo cambian las con-

[2] Para ejemplos mas complejos, nos referimos a DeCastellarnau y Saris (2014) y Saris y Revilla (2016).

clusiones sustantivas cuando los errores de medición se tienen en cuenta o no.

Datos

Ronda 8 de la Encuesta Social Europea

La ESS (http://www.europeansocialsurvey.org/about/faq.html) es una encuesta internacional realizada en Europa cada dos años desde 2001. La ESS realiza entrevistas cara a cara de aproximadamente una hora y selecciona nuevas muestras transversales para cada ronda. El cuestionario combina una sección central que se repite en cada ronda y módulos rotativos específicos de cada ronda.

El trabajo de campo de la octava ronda se llevó a cabo entre marzo de 2016 y diciembre de 2017 (datos de la octava ronda de la Encuesta Social Europea, 2016). Los tamaños de las muestras oscilan entre 880 (Islandia) y 2.852 (Alemania), estando España en el medio (N = 1.958, véase el apéndice C). Las tasas de respuesta para la ronda 8 también varían entre países, oscilando entre el 30,6% (Alemania) y el 74,4% (Israel), con una tasa de respuesta del 67,7% en España y una tasa de respuesta media del 55,4% (ESS, 2017).

El experimento MTMM se llevó a cabo en 23 de los países participantes. En países multilingües, la ESS realiza las encuestas en diferentes idiomas (por ejemplo, catalán y español en España). Dado que el idioma puede afectar la calidad de la medición (Saris y Gallhofer, 2014; Zavala-Rojas, 2016), analizamos cada idioma por separado. Sin embargo, el modelo MTMM no se puede estimar para idiomas con un número reducido de observaciones (por ejemplo, catalán en España; consulte el apéndice C para obtener una lista completa). Así pues, analizamos 27 grupos correspondientes a los grupos lingüísticos con tamaño muestral suficiente existentes en cada uno de los países disponibles (grupos país-idioma).

TABLA 1. *Preguntas de encuesta incluidas en el experimento MTMM de la ronda 8 de la ESS*

Rasgo	Formulación general de las preguntas
Nivel educativo	¿Qué importancia debería darse a tener un buen nivel educativo en la decisión de permitir o no a una persona que ha nacido y vivido siempre fuera de [país], venir a vivir aquí?
Tradición cristiana	¿Qué importancia debería darse a ser de un país de tradición cristiana en la decisión de permitir o no a una persona venir a vivir aquí?
Cualificación laboral	¿Qué importancia debería darse a tener una cualificación de las que [país] necesita en la decisión de permitir o no a una persona venir a vivir aquí?

Fuente: Elaboración propia.

El experimento MTMM

El experimento evalúa tres rasgos medidos cada uno con tres métodos. Los rasgos pretenden medir tres aspectos del concepto complejo «calificación para la entrada o exclusión de inmigrantes», respectivamente la importancia de tener: 1) un buen nivel educativo, 2) una tradición cristiana[3],

[3] Para Israel, «cristiana» se sustituye.

y 3) cualificaciones laborales necesarias en el país, para estar cualificado para entrar en dicho país. La tabla 1 presenta el redactado general de cada pregunta[4].

Con respecto a los métodos, la tabla 2 resume las características que varían entre métodos y proporciona las etiquetas de los puntos finales para cada escala.

TABLA 2. *Variación en las características y etiquetas de los puntos finales en cada escala*

		M_1	M_2	M_3
Características	Núm. de puntos	11	10	6
	Formato	Batería	Preguntas separadas	Batería
	Núm. puntos de referencia fijos	2	2	1
	Correspondencia	Alta	Media	Alta
	Preguntas en la tarjeta	No	Sí	No
Etiquetas	Primera categoría	0 Nada importante	1 Nada importante	0 Nada importante
	Última categoría	10 Extremadamente importante	10 Extremadamente importante	5 Muy importante

Fuente: Elaboración propia.

Cinco aspectos varían:

1) El número de categorías de respuesta: M_1 es una escala impar (11 puntos), mientras que M_2 y M_3 son escalas pares (10 y 6 puntos, respectivamente).
2) Preguntas separadas o batería (es decir, varias preguntas que comparten la misma escala se presentan juntas, la escala se repite solo una vez): M_1 y M_3 presentan las preguntas en formato de batería, mientras que M_2 las presenta como preguntas separadas.
3) El número de puntos de referencia fijos (es decir, categorías de respuesta que «no establecen ninguna duda sobre la posición del punto de referencia en la escala subjetiva en la mente del encuestado»; Saris y Gallhofer, 2014: 110): M_1 y M_2 presentan dos puntos de referencia fijos, mientras que M_3 presenta solo uno.
4) La correspondencia entre los números y las etiquetas verbales en la escala (por ejemplo, 0 representa mejor la idea de «Para nada» que 1): M_1 y M_3 presentan una correspondencia alta mientras que M_2 presenta una correspondencia media.
5) La presentación de la pregunta en las tarjetas que se enseñan a los participantes: por lo general, las tarjetas que la ESS proporciona (es decir, las tarjetas que se presentan a los encuestados para brindar ayuda visual a la vez que el entrevistador hace las preguntas) no contienen la pregunta sino solo las opciones de respuesta. En este experimento, en M_2 las preguntas se muestran en las tarjetas, mientras que en M_1 y M_3 no.

[4] Para el redactado específico de cada método, véase el apéndice D.

Ilustración de las implicaciones para la investigación sustantiva transnacional de no corregir por los errores de medición

Para ilustrar la implicación de no corregir por los errores de medición en la investigación transnacional, comparamos las correlaciones, antes y después de corregirlas por los errores de medición, entre la importancia que se le da a que un individuo venga de tradición cristiana y la importancia que se les da a sus cualificaciones laborales al momento de decidir si alguien nacido, criado y que vive en el exterior debe poder ir a vivir a un país determinado. En aras de la simplicidad, en esta ilustración nos enfocamos únicamente en uno de los métodos. Elegimos M_3 (escala de 6 puntos en formato de batería, con un punto de referencia fijo, alta correspondencia en la escala y que no proporciona la pregunta en la tarjeta que se enseña al participante), porque, para España, presenta una de las calidades más bajas. Ilustramos las implicaciones para ocho países: Alemania, España, Francia, Finlandia, Italia, Noruega, Portugal y Suecia.

Resultados

Nuestros análisis estiman la calidad de medición para 27 grupos país-idioma, tres rasgos y tres métodos. Dado que presentar todas las 243 estimaciones de calidad de medición no es práctico, primero agregamos todos los países y presentamos los resultados para cada rasgo y método en comparación con España. Luego, agregamos los rasgos y presentamos los resultados para cada país y método. De esa manera, podemos comparar la calidad de las mediciones, primero, entre rasgos y, segundo, entre países. Finalmente, presentamos un ejemplo de las implicaciones sustantivas de no corregir los errores de medición al comparar España con otros siete países.

Calidad promedio en todos los grupos de países e idiomas, por rasgo y método

La tabla 3 presenta la calidad de medición en España, así como la calidad media, mínima y máxima en los otros 26 grupos país-idioma (excluido España) para los diferentes rasgos y métodos.

Para todos los grupos país-idioma, rasgos y métodos, la calidad más alta obtenida es 0,99 (tradición cristiana-M_1) mientras que la más baja es 0,39 (cualificaciones laborales-M_1). Esto significa que entre el 1% (tradición cristiana-M_1) y el 61% (cualificaciones laborales-M_1) de la varianza en las respuestas observadas proviene de errores de medición.

TABLA 3. *Calidad de medición (q^2) en España y media, mínima y máxima calidad de medición de los otros 26 grupos país-idioma, por rasgo y método*

	Nivel educativo			Tradición cristiana			Cualificación laboral			Media de los rasgos		
Calidad q^2	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
Media 26 grupos	0,73	0,64	0,72	0,83	0,71	0,75	0,76	0,68	0,72	0,77	0,68	0,73
Máximo 26 grupos	0,90	0,85	0,87	0,99	0,85	0,92	0,92	0,86	0,85	0,90	0,82	0,85
Mínimo 26 grupos	0,56	0,41	0,41	0,41	0,57	0,64	0,39	0,54	0,42	0,53	0,54	0,49
España	0,85	0,69	0,64	0,87	0,73	0,70	0,88	0,72	0,64	0,87	0,71	0,66

Nota: Las estimaciones de calidad toman valores entre 0 y 1, representando 1 una relación perfecta entre la respuesta observada y el concepto de interés latente.

Fuente: Elaboración propia.

Asimismo, M_1 presenta una calidad más alta en promedio para los 26 grupos país-idioma y para España, para todos los rasgos. Sin embargo, existen algunas diferencias entre España y la media de los otros 26 grupos. Primero, aunque M_1 es el que tiene un mejor rendimiento en ambos casos, las estimaciones de calidad en España son especialmente buenas para todos los rasgos. Las estimaciones de calidad también son más altas para M_2 en España que la media de todos los demás grupos país-idioma. Sin embargo, para M_3, la calidad en España está por debajo de la media de todos los demás grupos país-idioma. Por otro lado, en España, M_2 presenta una mayor calidad que M_3 en todos los rasgos, mientras que para la media de los demás grupos país-idioma la tendencia es opuesta. Por lo tanto, aunque los formatos de batería pueden sufrir del fenómeno de no diferenciación (Saris y Gallhofer, 2014), en general recomendamos utilizar una escala de 11 puntos presentada en formato de batería, con dos puntos de referencia fijos, alta correspondencia entre números y etiquetas verbales, y ninguna pregunta en la tarjeta en lugar de los otros dos métodos, para medir los tres indicadores estudiados para el concepto «calificación de ingreso o exclusión de inmigrantes».

En cuanto a las diferencias entre rasgos, la «tradición cristiana» alcanza la calidad de medición promedio más alta para todos los métodos, para España y en promedio para los demás grupos de países e idiomas. Esto es interesante ya que se podría pensar que sería el rasgo con mayor propensión a generar sesgos de deseabilidad social, considerándose la religión un tema delicado. Finalmente, las diferencias entre rasgos son consistentes para España y la media de los otros grupos de países e idiomas. Aunque España presenta diferentes estimaciones de calidad, la relación entre las estimaciones de calidad y los rasgos es similar.

Calidad promedio en todos los rasgos, por grupo país-idioma y método

A continuación, las diferencias entre países se analizan con más detalle, esta vez agregando a nivel de rasgos. La tabla 4 presenta la calidad promedio en todos los rasgos, por grupo país-idioma y método, así como la posición de cada grupo país-idioma en el *ranking*, para cada método.

Primeramente, la calidad de la medición entre países y métodos varía de 0,53 (Estonia-Ruso-M_1) a 0,88 (Islandia-M_1). Por lo tanto, en todos los métodos y países, la varianza explicada por los errores de medición va del 12% (Islandia-M_1) al 47% (Estonia-Ruso-M_1). Entonces, existen grandes diferencias en la calidad de medición entre los diferentes grupos país-idioma. La tendencia general es que M_1 (escala de 11 puntos en formato de batería, con dos puntos de referencia fijos, alta correspondencia entre números y etiquetas verbales y sin preguntas en las tarjetas) se desempeña mejor que M_2 y M_3. Además, los países del centro y norte de Europa presentan, en general, una calidad de medición más alta que sus homólogos del este y del sur.

Comparando España con los demás, España tiene la cuarta calidad más alta para M_1 y la décima más alta para M_2. Sin embargo, para M_3, España presenta la cuarta calidad más baja. Por lo tanto, existen diferencias importantes entre los métodos para España, los cuales deben tenerse en cuenta. Primero, usar una escala de 11 puntos presentada en un formato de batería, con dos puntos de referencia fijos, alta correspondencia entre números y etiquetas verbales, y ninguna pregunta en la tarjeta funciona mucho mejor en España que en la mayoría de los grupos de países e idiomas. En segundo lugar, una escala de 6 puntos en formato de batería, con un solo punto de referencia fijo, alta correspondencia entre números y etiquetas verbales, y sin preguntas en la tarjeta, se comporta peor en

España que en la mayoría de los grupos de países e idiomas analizados. Esto sugiere que las diferencias culturales y lingüísticas entre países afectan el tamaño de los errores de medición asociados con diferentes métodos.

TABLA 4. *Calidad media (q^2) de los rasgos agregados, por grupo país-idioma y método*

	Calidad media de los tres rasgos			*Ranking*		
	M_1	M_2	M_3	M_1	M_2	M_3
Austria	0,74	0,67	0,79	20	14	7
Bélgica-Neerlandés	0,83	0,70	0,74	9	11	13
Bélgica-Francés	0,87	0,58	0,68	5	25	20
República Checa	0,79	0,69	0,69	13	13	18
Estonia-Estonio	0,77	0,58	0,72	16	24	14
Estonia-Ruso	0,53	0,63	0,69	27	19	19
Finlandia	0,90	0,75	0,74	1	5	12
Francia	0,82	0,54	0,67	11	27	21
Alemania	0,77	0,80	0,76	15	3	10
Gran Bretaña	0,72	0,73	0,71	21	8	15
Hungría	0,70	0,61	0,64	22	22	25
Islandia	0,88	0,82	0,85	2	1	1
Irlanda	0,83	0,61	0,49	10	23	27
Israel-Árabe	0,88	0,80	0,78	3	2	8
Israel-Hebreo	0,76	0,63	0,62	18	20	26
Italia	0,68	0,57	0,84	24	26	2
Lituania	0,69	0,63	0,81	23	18	5
Países Bajos	0,80	0,65	0,77	12	17	9
Noruega	0,84	0,74	0,82	7	6	4
Polonia	0,75	0,73	0,71	19	9	22
Portugal	0,67	0,61	0,75	25	21	11
Rusia	0,66	0,66	0,83	26	15	3
Eslovenia	0,79	0,66	0,66	14	16	24
España	0,87	0,71	0,66	4	10	23
Suecia	0,83	0,74	0,79	8	7	6
Suiza-Francés	0,85	0,69	0,71	6	12	16
Suiza-Alemán	0,76	0,78	0,70	17	4	17

Fuente: Elaboración propia.

Implicaciones sustantivas para la investigación transnacional: una ilustración

Los resultados demuestran que existen diferencias no despreciables entre España y otros grupos país-idioma. Esto puede tener importantes implicaciones en la investigación transnacional cuando no se tienen en cuenta los errores de medición.

La tabla 5 presenta las correlaciones sin y con corrección por errores de medición para cada país, ordenadas de mayor a menor correlación corregida. Además, presenta el *ranking* de cada país (1 significa la correlación más alta) para correlaciones sin y con corrección por errores de medición.

Podemos ver un aumento en las correlaciones al corregir por errores de medición, excepto en Italia. Para España la correlación pasa de 0,41 a 0,46. Sin embargo, el cambio no es homogéneo entre países: mientras que para Finlandia la correlación aumenta 0,09 puntos, para Italia se reduce en 0,03 puntos. En consecuencia, comparar países que utilizan correlaciones sin corrección en lugar de correlaciones con corrección por errores de medición lleva a conclusiones sustantivas diferentes. En particular, en términos del *ranking*, sin corrección por errores de medición, Italia presenta la cuarta correlación más alta mientras que con corrección, presenta la más baja. Por lo tanto, si los investigadores quisieran comparar la correlación entre «tradición cristiana» y «cualificación laboral» para España e Italia, no corregir por los errores de medición llevaría a conclusiones erróneas.

TABLA 5. *Coeficientes de correlación y* ranking *sin y con corrección*

	Correlación		*Ranking*	
	Sin corrección	Con corrección	Sin corrección	Con corrección
Finlandia	0,47	0,56	1	1
Suecia	0,41	0,47	2	2
España	0,41	0,46	3	3
Noruega	0,39	0,44	5	4
Portugal	0,38	0,43	6	5
Alemania	0,36	0,42	7	6
Francia	0,35	0,40	8	7
Italia	0,40	0,37	4	8

Fuente: Elaboración propia.

Discusión y conclusiones

Resultados principales

Nuestro principal objetivo ha sido comparar la calidad de medición de las preguntas de las encuestas en España con otros países europeos, ya que la investigación existente centrada en España desde una perspectiva comparativa, aunque relevante, es aún escasa.

En general, para los tres rasgos considerados, encontramos que la calidad de medición varía mucho entre países, desde un promedio (de todos los métodos y rasgos) de 0,62 en Estonia-Ruso a 0,85 en Islandia, lo que significa que en promedio entre el 62% y el 85% de la varianza en las respuestas observadas se debe a los conceptos de interés latentes mientras que del 15% al 38% se debe a errores

de medición. Los países del centro y norte de Europa presentan, en general, una mayor calidad de medición. Esto podría estar relacionado con el hecho de que los países con bajos niveles de colectivismo y corrupción son menos propensos a la deseabilidad social (Rammstedt, Danner y Bosnjak, 2017) y/o a las diferencias lingüísticas.

Además, la calidad de medición de los tres rasgos considerados fue superior para España que para la media de los otros 26 grupos país-idioma analizados. Estos resultados van en línea con investigaciones previas sobre la calidad de medición (Saris *et al.*, 2010; Revilla, Saris y Krosnick, 2014) basadas también en datos de ESS.

Sin embargo, existen diferencias entre los métodos. M_1 presenta un rendimiento especialmente bueno en España en comparación con la mayoría de los demás países, y ocupa el cuarto lugar de los 27 grupos país-idioma. Sin embargo, para M_3, España presenta la cuarta estimación de calidad más baja. Por tanto, aunque en general España presenta una calidad de medición superior, los investigadores no deberían asumir una calidad superior a la media en España para cualquier método. Al contrario, deberían considerar que algunos métodos pueden funcionar mejor y otros peor en España que en otros países.

No considerar las potenciales diferencias en el tamaño de los errores de medición al comparar España con otros países afecta a las conclusiones de fondo. En primer lugar, en nuestra ilustración, las correlaciones observadas fueron mayormente subestimadas. Además, los *rankings* de países con mayor y menor correlación con y sin corrección difirieron substancialmente. En particular, sin corrección, España e Italia presentaron correlaciones similares. Esto sugiere que, para ambos países, la creencia de que venir de tradición cristiana es importante para que los inmigrantes estén cualificados para ingresar al país esta correlacionada en similar medida con la creencia de que los inmigrantes que presentan cualificaciones laborales necesarias en dicho país están más cualificados para su ingreso. Sin embargo, la correlación corregida por errores de medición fue mayor para España que para Italia, lo que apunta a una conclusión sustantiva diferente: aunque España e Italia tienen un nivel similar de religiosidad (Evans y Baronavski, 2018), la relación entre la proveniencia de una tradición cristiana y las cualificaciones laborales es más fuerte para España. Esta diferencia entre las correlaciones con y sin corrección entre España e Italia se relaciona principalmente con la diferencia en la calidad de medición de ambos países (0,18 puntos menos en España que en Italia), y en menor medida a la diferencia en CMV (0,04 más alto en España que en Italia). Después de corregir utilizando la ecuación 3, España presenta una correlación verdadera más alta que Italia. Esto indica que, aunque España presenta un CMV algo más alto, la calidad notablemente más baja en España llevó a una subestimación de la correlación en comparación con la sobreestimación de Italia.

Límites y futura investigación

Estos resultados presentan algunas limitaciones. Primero, estos hallazgos son específicos para los temas analizados y los métodos utilizados y no deben ser generalizados a otras preguntas o métodos. En segundo lugar, debido al reducido tamaño de la muestra, no se pudieron analizar algunos idiomas. En particular, no hemos podido utilizar a los encuestados de habla catalana, lo que no permite comparar las estimaciones de calidad entre los idiomas de España. Sin embargo, considerando que otros países presentan diferentes calidades de medición en función del idioma de administración, podríamos esperar lo mismo para España. Una investigación adicional podría explorar específicamente las diferencias entre el catalán y el español. Ade-

más, los experimentos MTMM no son los más adecuados para explicar por qué en algunos países la calidad es más alta que en otros. Futuras investigaciones podrían centrarse en encontrar explicaciones. También, solo hemos ilustrado cómo corregir las correlaciones entre dos conceptos simples para errores de medición. Sin embargo, la corrección por errores de medición se puede aplicar a modelos más complejos (por ejemplo, regresiones). Nos referimos a DeCastellarnau y Saris (2014), Saris y Gallhofer (2014) y Saris y Revilla (2016) para obtener ejemplos y pautas sobre cómo hacerlo para otros modelos. Asimismo, la calidad de la medición proporciona información sobre relaciones estandarizadas. Los investigadores interesados en comparar relaciones no estandarizadas deben estudiar la equivalencia de medición de constructos entre países (Davidov *et al.*, 2014). Finalmente, la estimación del tamaño de los errores de medición también puede verse afectada por errores. Así, incluso las correlaciones corregidas presentan algunos errores.

Para poder sacar conclusiones generales, futuras investigaciones deben explorar nuevos temas y métodos para ver si la tendencia es la misma para diferentes rasgos y escalas. Sin embargo, no siempre es posible realizar experimentos MTMM. Una alternativa es utilizar el software SQP. Utilizando predicciones de SQP, los investigadores podrían obtener una imagen más clara del efecto de diferentes métodos para diferentes preguntas (DeCastellarnau y Revilla, 2017). Asimismo, la sensibilidad de los análisis se podría testear utilizando el software SQP para explorar si las predicciones y estimaciones son similares y, de no ser así, cómo las diferencias afectan las correcciones por errores de medición.

Recomendaciones prácticas

Primero, basándonos en nuestros resultados, para medir el concepto «calificación para la entrada o exclusión de inmigrantes» en España recomendamos utilizar M_1 (escala de 11 puntos en formato de batería, con dos puntos de referencia fijos, alta correspondencia entre números y etiquetas verbales y ninguna pregunta en las tarjetas que se enseñan a los participantes) en lugar de M_2 y M_3.

Segundo, este estudio de caso ilustra lo que previamente se había afirmado en otras investigaciones (p. ej., Saris y Gallhofer, 2007), es decir, que: 1) los investigadores sustantivos deben tener en cuenta que la comparación de relaciones estadísticas estandarizadas entre países solo es posible si la calidad de la medición es la misma, y 2) incluso en este caso, es necesario corregir por los errores de medición para estimar adecuadamente las relaciones de interés; es decir, las que existen entre los conceptos, y no entre las variables observadas, que son solo medidas imperfectas de los conceptos de interés. Por lo tanto, de acuerdo con investigaciones anteriores y los resultados de este nuevo estudio, recomendamos corregir por los errores de medición siempre que sea posible.

Bibliografía

Alwin, Duane F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken, New Jersey: John Wiley and Sons, Inc.

Andrews, Frank M. (1984). «Construct Validity and Error Components of Survey Measures: A Structural Modelling Approach». *Public Opinion Quarterly,* 48(2): 409-442. doi: 10.1086/268840

Beilmann, Mai; Kööts-Ausmees, Liisi y Realo, Anu (2018). «The Relationship Between Social Capital and Individualism-Collectivism in Europe». *Social Indicators Research,* 137: 641-664. doi: 10.1007/s11205-017-1614-4

Beullens, Koen; Loosveldt, Geert; Vandenplas, Caroline y Stoop, Ineke (2018). «Response Rates in the European Social Survey: Increasing, Decreasing or a Matter of Fieldwork Efforts?». *Survey Methods: Insights from the Field*. doi: 10.13094/SMIF-2018-00003

Bosch, Oriol J.; Revilla, Melanie; DeCastellarnau, Anna y Weber, Wiebke (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. doi: 10.1177/0894439317750089

Browne, Michael W. (1984). «The Decomposition of Multitrait-Multimethod Matrices». *British Journal of Mathematical and Statistical Psychology,* 37(1): 1-21. doi: 10.1111/j.2044-8317.1984.tb00785.x

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

Corten, Irmgard W.; Saris, Willem E.; Coenders, Germà; Veld, William M. van der; Aalberts, Chris E. y Kornelis, Charles (2002). «Fit of Different Models for Multitrait-Multimethod Experiments». *Structural Equation Modeling: A Multidisciplinary Journal,* 9(2): 213-232. doi: 10.1207/S15328007SEM0902_4

Couper, Mick P. y Leeuw, Edith D. de (2003). «Nonresponse in Cross-Cultural and Cross-National Surveys». En: Harkness, J.; Vijver, F. van de y Mohler, P. (eds.). *Cross-Cultural Survey Methods*. New York: Wiley.

Cudeck, Robert (1989). «Analysis of Correlation Matrices Using Covariance Structure Models». *Psychological Bulletin,* 105(2): 317-327. doi: 10.1037/0033-2909.105.2.317

Davidov, Eldad; Meuleman, Bart; Cieciuch, Jan; Schmidt, Peter y Billiet, Jaak (2014). «Measurement Equivalence in Cross-National Research». *Annual Review of Sociology,* 40(1): 55-75. doi: 10.1146/annurev-soc-071913-043137

DeCastellarnau, Anna y Saris, Willem E. (2014). «A Simple Way to Correct for Measurement Errors in Survey Research». *European Social Survey Education Net (ESS Edunet). European Social Survey Education Net (ESS EduNet).* Disponible en: http://essedunet.nsd.uib.no/cms/topics/measurement/

DeCastellarnau, Anna y Revilla, Melanie (2017). «Two approaches to evaluate measurement quality in online surveys: An application using the norwegian citizen panel». *Survey Research Methods*, 11(4): 415-433. doi: 10.18148/srm/2017.v11i4.7226

DeMaio, Theresea J. (1984). «Social Desirability in Survey Measurement: A Review». En: Turner, C. F. y Martin, E. (eds.). *Surveying Subjective Phenomena*. New York: Russell Sage.

ESS (2016). *Data File Edition 2.1. NSD —Norweegian Centre for Reseaerch Data, Norway— Data Archive y Distributor of ESS Data for ESSERIC.* Disponible en: https://www.europeansocialsurvey.org/data/download.html?r=8, acceso el 21 de septiembre de 2020.

ESS (2017). *ESS8 - 2016 Fieldwork Summary and Deviations.* Disponible en: https://www.europeansocialsurvey.org/data/deviations_8.html, acceso el 19 de septiembre de 2020.

ESS (2019). *ESS User Statistics* . Disponible en: http://www.europeansocialsurvey.org/docs/data_users/ESS_data_user_stats_aug_2019.pdf, acceso el 21 de septiembre de 2020.

Evans, Jonathan y Baronavski, Chris (2018). *How Do European Countries Differ in Religious Commitment?* Disponible en: https://www.pewresearch.org/fact-tank/2018/12/05/how-do-european-countries-differ-in-religious-commitment/, acceso el 21 de septiembre de 2020.

GESIS (2019). *Overview of Comparative Surveys Worldwide*. Disponible en: www.gesis.org/ComparativeSurveyOverview, acceso el 21 de septiembre de 2020.

Herk, Hester van; Poortinga, Ype H. y Verhallen, Theo M. M. (2004). «Response Styles in Rating Scales: Evidence of Method Bias in Data from Six EU Countries». *Journal of Cross-Cultural Psychology,* 35(3): 346-360. doi: 10.1177/0022022104264126

Hox, Joop J. y Bechger, Timo M. (1998). «An Introduction to Structural Equation Modeling». *Family Science Review,* 11: 354-373.

Johnson, Timothy P. y Vijver, Fons J. R. van de (2003). «Social Desirability in Cross-Cultural Research». En: Vijver, F. van de; Mohler, P. y Wiley, J. (eds.). *Cross-Cultural Survey Methods*. Hoboken, New Jersey: Wiley-Interscience.

Jöreskog, Karl G. y Sörbom, Dag (1996). *LISREL 8: User's Reference Guide*. Uppsala: Scientific Software International.

Leung, Kwok; Au, Yuk-Fai; Fernández Dols, José Miguel e Iwawaki, Saburo (1992). «Preference for Methods of Conflict Processing in Two Collectivist Cultures». *International Journal of Psychology*, 27(2): 195-209. doi: 10.1080/00207599208246875

Liao, Pei-Shan; Saris, Willem E. y Zavala-Rojas, Diana (2019). «Cross-National Comparison of Equivalence and Measurement Quality of Response Scales in Denmark y Taiwan». *Journal of Official Statistics,* 35(1): 117-135. doi: 10.2478/jos-2019-0006

Meurs, A. van y Saris, Willem E. (1990). «Memory Effects in MTMM Studies». En: Saris, W. E. y Meurs, A. van (eds.). *Evaluation of Measurement Instruments by Meta-Analysis of Multitraitmultimethod Studies*. Amsterdam: North-Holland.

Rammstedt, Beatrice; Danner, Daniel y Bosnjak, Michael (2017). «Acquiescence Response Styles: A Multilevel Model Explaining Individual-Level y Country-Level Differences». *Personality and Individual Differences*, 107(1): 190-194. doi: 10.1016/j.paid.2016.11.038

Revilla, Melanie y Saris, Willem E. (2013). «The Split-Ballot Multitrait-Multimethod Approach: Implementation and Problems». *Structural Equation Modeling: A Multidisciplinary Journal,* 20(1): 27-46. doi: 10.1080/10705511.2013.742379

Revilla, Melanie; Saris, Willem E. y Krosnick, Jon A. (2014). «Choosing the Number of Categories in Agree-Disagree Scales». *Sociological Methods & Research,* 43(1): 73-97. doi: 10.1177/0049124113509605

Revilla, Melanie; Bosch, Oriol J. y Weber, Wiebke (2019). «Unbalanced 3-Group Split-Ballot Multitrait–Multimethod Design?». *Structural Equation Modeling,* 26(3): 437-447. doi: 10.1080/10705511.2018.1536860

Saris, Willem E. y Andrews, Frank M. (1991). «Evaluation of Measurement Instruments Using a Structural Modeling Approach». En: Biemer, P.; Groves, R.; Lyberg, L.; Mathiowetz, N. y Sudman, S. (eds.). *Measurement Errors in Surveys*. New York: John Wiley and Sons, Inc.

Saris, Willem E. y Aalberts, Chris (2003). «Different Explanations for Correlated Disturbance Terms in MTMM Studies». *Structural Equation Modeling: A Multidisciplinary Journal,* 10(2): 193-213. doi: 10.1207/S15328007SEM1002_2

Saris, Willem E. y Gallhofer, Irmtraud N. (2014 [2007]). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, New Jersey: John Wiley and Sons, Inc.

Saris, Willem E. y Revilla, Melanie (2016). «Correction for Measurement Errors in Survey Research: Necessary and Possible». *Social Indicators Research,* 127(3): 1005-1020. doi: 10.1007/s11205-015-1002-x

Saris, Willem E.; Satorra, Albert y Coenders, Germa (2004). «A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design». *Sociological Methodology*, 34(1): 311-347. doi: 10.1111/j.0081-1750-.2004.00155.x

Saris, Willem E.; Satorra, Albert y Veld, William M. van der (2009). «Testing Structural Equation Models or Detection of Misspecifications?». *Structural Equation Modeling: A Multidisciplinary Journal,* 16(4): 561-582. doi: 10.1080/10705510903203433

Saris, Willem E.; Revilla, Melanie; Krosnick, Jon A. y 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. doi: 10.18148/srm/2010.v4i1.2682

Saris, Willem E.; Oberski, Daniel L.; Revilla, Melanie; Zavala-Rojas, Diana; Lilleoja, Laur; Gallhofer, Irmtraud, N. y Gruner,Thomas (2011). *The Development of the Program SQP 2.0 for the Prediction of the Quality of Survey Questions*. (RECSM Working Paper). Disponible en: http://www.upf.edu/survey/_pdf/RECSM_wp024.pdf

Transparency International (2019). *Corruption Perceptions Index 2019.* Disponible en: https://www.transparency.org/en/cpi/2019, acceso el 21 de septiembre de 2020.

Veld, William M. van der; Saris, Willem E. y Satorra, Albert (2008). *Judgement Rule Aid for Structural Equation Models.* (Versión 3.0.4 Beta).

Zavala-Rojas, Diana (2016). *Measurement Equivalence in Multilingual Comparative Survey Research.* Barcelona: Universitat Pompeu Fabra. [Tesis doctoral].

RECEPCIÓN: 19/11/2019
REVISIÓN: 06/05/2020
APROBACIÓN: 11/09/2020

Apéndice A: ejemplo del código base de LISREL

```
Analysis Group 1
Data ng=3 ni=9 no=221 ma=cm
km file=ATGER-group1.corr
mean file=ATGER-group1.mean
sd file=ATGER-group1.sd
model ny=9 ne=9 nk=6 ly=fu,fi te=sy,fi ps=sy,fi be=fu,fi ga=fu,fi ph=sy,fi
value 1 ly 1 1 ly 2 2 ly 3 3 ly 4 4 ly 5 5 ly 6 6 te 7 7 te 8 8 te 9 9
fr te 1 1 te 2 2 te 3 3 te 4 4 te 5 5 te 6 6
value 0 ly 7 7 ly 8 8 ly 9 9
fr ga 1 1 ga 2 2 ga 3 3 ga 4 1 ga 5 2 ga 6 3 ga 7 1 ga 8 2 ga 9 3
value 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 ph 1 1 ph 2 2 ph 3 3
fr ph 2 1 ph 3 1 ph 3 2 ph 4 4 ph 5 5 ph 6 6
start .5 all
out mi iter= 300 adm=off sc
Analysis Group 2
Data ni=9 no=219 ma=cm
km file=ATGER-group2.corr
mean file=ATGER-group2.mean
sd file=ATGER-group2.sd
model ny=9 ne=9 nk=6 ly=fu,fi te=sy,fi ps=in be=in ga=in ph=in
fr te 4 4 te 5 5 te 6 6 te 7 7 te 8 8 te 9 9
va 1 ly 4 4 ly 5 5 ly 6 6 ly 7 7 ly 8 8 ly 9 9 te 1 1 te 2 2 te 3 3
equal te 1 4 4 te 4 4
equal te 1 5 5 te 5 5
equal te 1 6 6 te 6 6
value 0 ly 1 1 ly 2 2 ly 3 3
out mi iter= 300 adm=off sc
Analysis Group 3
Data ni=9 no=228 ma=cm
km file=ATGER-group3.corr
mean file=ATGER-group3.mean
sd file=ATGER-group3.sd
model ny=9 ne=9 nk=6 ly=fu,fi te=sy,fi ps=in be=in ga=in ph=in
fr te 1 1 te 2 2 te 3 3 te 7 7 te 8 8 te 9 9
va 1 ly 1 1 ly 2 2 ly 3 3 ly 7 7 ly 8 8 ly 9 9 te 4 4 te 5 5 te 6 6
equal te 1 1 1 te 1 1
equal te 1 2 2 te 2 2
equal te 1 3 3 te 3 3
equal te 2 7 7 te 7 7
equal te 2 8 8 te 8 8
equal te 2 9 9 te 9 9
value 0 ly 4 4 ly 5 5 ly 6 6
out mi iter= 300 adm=off sc
```

Apéndice B: modificaciones del análisis del modelo *Split Ballot-True Score*-MTMM, ajuste del modelo y evaluación de JRule

Grupo país-idioma	Modificaciones del modelo (notación LISREL)	df	$\chi 2$	Núm. de errores de especificación por JRule
Austria	Free GA 2 4 TE 8 8 TE 5 5	36	78,90	0
Bélgica-Neerlandés	Free TE 1 1	38	92,20	0
Bélgica-Francés	Free GA 8 6	38	60,59	2
República Checa	Free TE 7 7 TE 2 2 TE 5 5 GA 5 5	35	86,50	0
Estonia-Estonio	Free TE 6 6 TE 5 5 TE 8 5	36	95,20	0
Estonia-Ruso	Free TE 9 9	38	55,87	4
Finlandia	Free TE 8 8 TE 7 7 TE 4 4 GA 5 5 TE 8 2	34	91,10	2
Francia	Free TE 4 4 TE 7 7 GA 4 1 TE 3 1	35	77,32	3
Alemania	Free TE 7 7 TE 2 2 TE 4 4 TE 6 6 GA 2 4	34	79,69	3
Gran Bretaña	Free TE 4 4 TE 1 1 TE 3 3 TE 6 6 GA 8 6 GA 2 4 GA 6 5	32	83,00	4
Hungría	Free TE 4 4 TE 7 7 TE 5 5 TE 9 9 GA 7 6 (G2)	34	83,27	1
Irlanda	Free TE 7 7 TE 8 8 TE 4 4 GA 7 6 GA 8 6 GA 9 6 (G2); fix TE 2 2 (va 0)	34	88,41	2
Israel-Árabe	Free TE 8 8 TE 8 8 GA 2 4 PH 5 4 GA 9 3 (G3)	34	41,70	1
Israel-Hebreo	Free TE 1 1 GA 8 6 GA 5 5	36	81,45	2
Islandia	Free TE 4 4 TE 6 6 TE 1 1	36	71,45	0
Italia	Free TE 4 4 TE 7 7 TE 5 5 PH 5 4	37	98,26	1
Lituania	PH 5 4	38	74,01	0
Países Bajos	TE 6 6 TE 4 4 GA 5 5 GA 6 5 PH 5 4	36	58,20	1
Noruega	TE 4 4 TE 8 8 TE 7 7 GA 2 4	37	88,08	2
Polonia	TE 4 4 TE 1 1 TE 7 7 TE 7 1	37	85,50	0
Portugal	Ninguna	39	69,03	0
Rusia	TE 4 4 TE 5 5 TE 2 2 TE 1 1 GA 6 5 (G2) PH 5 4	33	83,35	1
Eslovenia	TE 1 1	38	60,48	3
España	TE 4 4 GA 8 6 PH 6 5	36	92,04	0
Suecia	TE 1 1 TE 7 7 TE 2 2 TE 4 4 GA 5 5	36	68,07	2
Suiza-Francés	TE 4 4 TE 6 6 TE 3 3	36	46,95	2
Suiza-Alemán	TE 4 4 TE 7 7 GA 8 6	36	87,70	2

Fuente: Elaboración propia.

Apéndice C: tamaño muestral de la ronda 8 de la ESS por grupo país-idioma

Grupo país-idioma	Grupo 1	Grupo 2	Grupo 3	Casos totales
Austria	221	219	228	668
Bélgica-Neerlandés	345	365	342	1.052
Bélgica-Francés	236	241	237	714
República Checa	763	777	726	2.266
Estonia-Estonio	522	510	520	1.552
Estonia-Ruso	169	139	159	467
Finlandia	608	603	590	1.801
Francia	685	680	696	2.061
Alemania	959	958	935	2.852
Gran Bretaña	662	653	644	1.959
Hungría	527	557	530	1.614
Irlanda	913	933	911	2.757
Israel-Árabe	175	168	182	525
Israel-Hebreo	660	663	685	2.008
Islandia	295	293	292	880
Italia	935	843	848	2.626
Lituania	654	662	627	1.943
Países Bajos	557	554	570	1.681
Noruega	497	563	485	1.545
Polonia	591	585	516	1.692
Portugal	408	451	411	1.270
Rusia	822	812	796	2.430
Eslovenia	447	424	436	1.307
España	648	599	590	1.837
Suecia	541	506	504	1.551
Suiza-Francés	123	133	129	385
Suiza-Alemán	361	367	350	1.078

Nota: Finlandia-Sueco, Israel-Ruso, Lituania-Ruso, España-Catalán y Suiza-Italiano no se pudieron analizar debido a que el tamaño de la muestra era < 100 casos por grupo.

Fuente: Elaboración propia.

Apéndice D: Formulaciones de las preguntas en el cuestionario, por método

Introducción (similar en todos los casos)

Hay personas de otros países que vienen a vivir a [país] por distintas razones. Algunas tienen antepasados que eran de aquí. Otras vienen a trabajar o para reunirse con sus familias. Otras vienen porque están amenazadas. Aquí hay algunas preguntas sobre este tema.

Método 1

¿Qué importancia debería tener cada uno de los siguientes aspectos en la decisión de permitir o no a una persona que ha nacido y vivido siempre fuera de España, venir a vivir aquí? En primer lugar, ¿qué importancia debería tener que esa persona… **Leer cada frase…**

		Nada importante										Extremadamente importante	(NC)	(NS)
C33	… tenga un buen nivel educativo?	00	01	02	03	04	05	06	07	08	09	10	77	88
C34	… sea de un país de tradición cristiana?	00	01	02	03	04	05	06	07	08	09	10	77	88
C35	...tenga una cualificación laboral de las que [país] necesita?	00	01	02	03	04	05	06	07	08	09	10	77	88

Método 2

¿Qué importancia debería darse a tener un buen nivel educativo en la decisión de permitir o no a una persona que ha nacido y vivido siempre fuera de [país], venir a vivir aquí?

Nada importante									Extremadamente importante	(NC)	(NS)
01	02	03	04	05	06	07	08	09	10	77	88

¿Qué importancia debería darse a ser de un país de tradición cristiana en la decisión de permitir o no a una persona venir a vivir aquí?

Nada importante									Extremadamente importante	(NC)	(NS)
01	02	03	04	05	06	07	08	09	10	77	88

¿Qué importancia debería darse a tener una cualificación de las que [país] necesita en la decisión de permitir o no a una persona venir a vivir aquí?

Nada importante									Extremadamente importante	(NC)	(NS)
01	02	03	04	05	06	07	08	09	10	77	88

Método 3

Tarjeta 30 ¿Qué importancia debería tener cada uno de los siguientes aspectos en la decisión de permitir o no, a una persona que ha nacido y vivido siempre fuera de [país], venir a vivir aquí? En primer lugar, ¿qué importancia debería tener que esa persona... **leer cada frase…**

		Nada importante					Muy importante	(NC)	(NS)
C39	… tenga un buen nivel educativo?	00	01	02	03	04	05	7	8
C40	… sea de un país de tradición cristiana?	00	01	02	03	04	05	7	8
C41	...tenga una cualificación laboral de las que [país] necesita?	00	01	02	03	04	05	7	8