The results of the ESS Round 7 two-group SB-MTMM experiments

Anna DeCastellarnau

Universitat Pompeu Fabra

Barcelona

In Round 7 of the European Social Survey (ESS) three Split-Ballot Multitrait-Multimethod (SB-MTMM) experiments were conducted to evaluate the measurement quality of survey questions. In this report, we first, define measurement quality, the experimental SB-MTMM approach used to evaluate it, and we will explain how it can be estimated. Second, we describe the design of the different experiments. Finally, we report the results of these experiments and discuss the differences in measurement quality of the responses for the different experimental conditions and by countries.

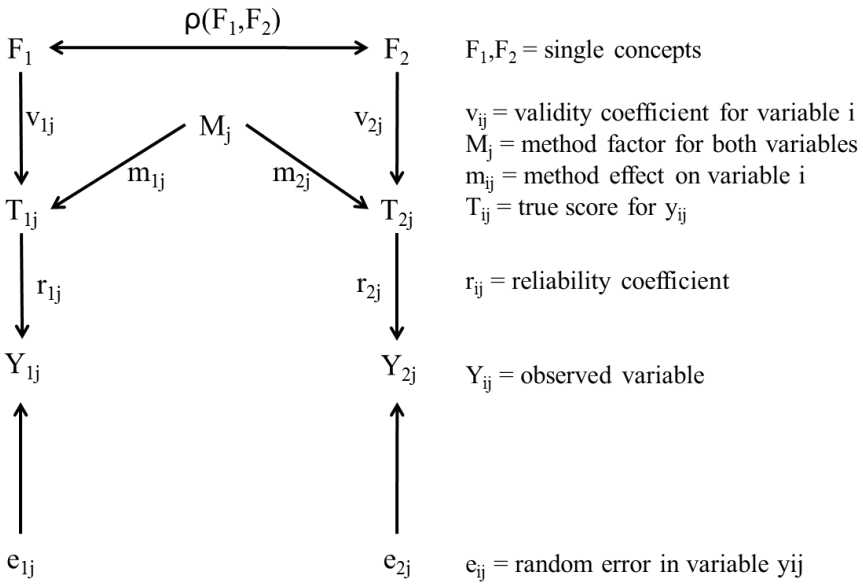
1. Evaluation of measurement quality of single concepts

In the 7th Round, three SB-MTMM experiments have been administrated. Each experiment consists of nine survey questions, which measure three different single concepts by three different methods. The first experiment aims to measure an ambit of attitudes towards immigration, the 'qualification for entry or exclusion of immigrants', by the single concepts 'ability to speak language', 'being white', and the 'commitment to the way of life'. The second experiment aims to measure an ambit of political efficacy, the 'system responsiveness or external political efficacy', by single concepts 'people have a say about the government', 'people have an influence in politics', and 'politicians care what people think'. The third experiment aims to measure another ambit of political efficacy, the 'subjective competence or internal political efficacy', by the single concepts 'ability to take an active role in a group about political issues', the 'confidence in ability to participate in politics', and the 'facility to take part in politics'. The aim of these experiments is to evaluate the measurement quality of each single concept using different formulations (i.e. different methods), i.e. the nine survey questions.

1.1. The measurement quality criteria: definition

The evaluation of single concepts is done through the quantification of survey questions' measurement quality. Figure 1 presents the basic response model used as starting point to evaluate the measurement quality of survey questions. This is the true score measurement model as proposed by Saris and Andrews (1991) for two single concepts measured by the same method.

Figure 1: The measurement model for two single concepts measured with the same method



In Figure 1, F_i is the i[th] single concept of interest; M_j is the j[th] method factor; Y_{ij} is the observed variable for the i[th] single concept and the j[th] method; and T_{ij} is the systematic component or true score of the response to $Y_{ij.}$ Figure 1 allows disentangling the proportions of measurement quality in a survey question, as follows:

The difference between the observed response (Y_{ij}) and the true score (T_{ij}) corresponds to random measurement error (e_{ij}). The effect r_{ij} represents the reliability coefficient and its squared is the reliability (r_{ij}^2), i.e. the strength of the relationship between the true score and the observed variable, and its complement is the random error.

The true score (T_{ij}) is separated from the single concept (F_i) because it is affected by the method (m_{ij}) used to measure it. The effect v_{ij} represents the validity coefficient and its squared is the validity (v_{ij}^2), i.e. the strength of the relationship between the variable of interest and the true score, and its complement is the method effect.

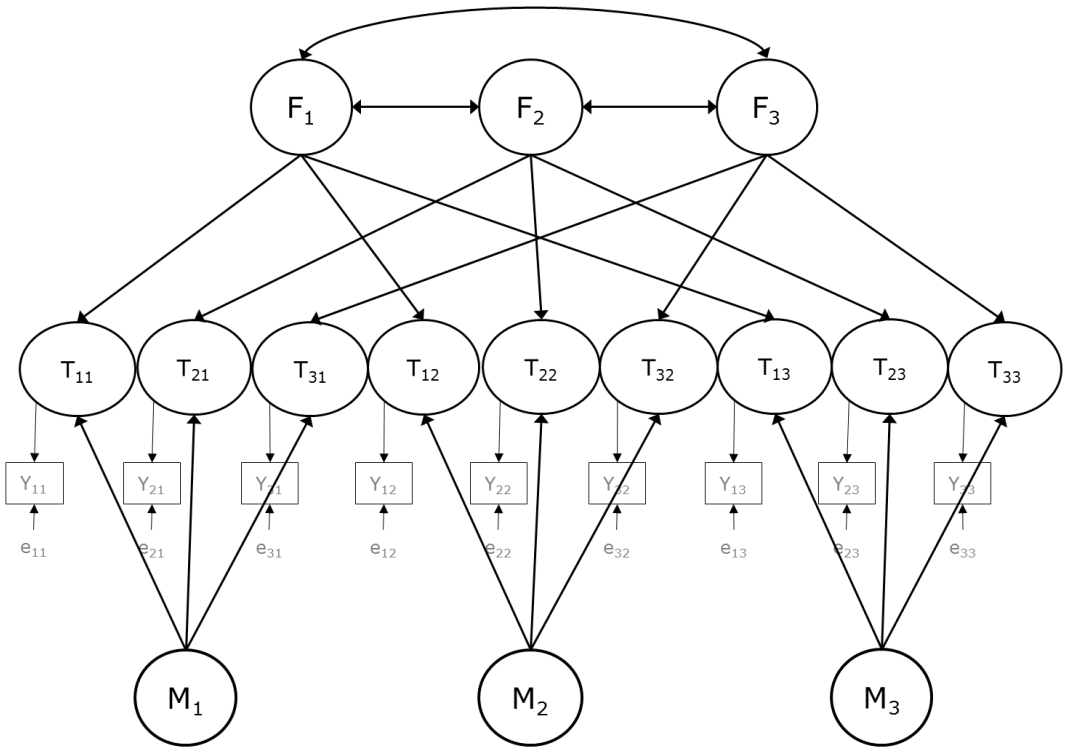
The measurement quality of a question (q_{ij}^2), defined as the strength of the relationship between the variable of interest and the observed variable, can be computed as the product of reliability and validity: $q_{ij}^2 = r_{ij}^2 \cdot v_{ij}^2$. We call q_{ij} the quality coefficient.

The measurement reliability, validity and quality take values between 0 and 1. The closer to one, the better the measurement instrument is.

1.2. Split-Ballot Multitrait-Multimethod (SB-MTMM) approach

The model presented in Figure 1 is not identified. Therefore, to be able to estimate reliability and validity coefficients, it is necessary to repeat several questions, measuring different single concepts (from now on called *traits*) using several methods, for instance, 2-point scale, 6-point scale, 11-point scale, etc. This is the MTMM approach, first developed by Campbell and Fiske (1959) and implemented by Andrews (1984) for structural equation models. Figure 2 illustrates the true score MTMM model for 3 traits each measured with 3 methods.

Figure 2: True score MTMM model for 3 traits and 3 methods



Following Figure 2, each trait (F_i) is measured three times, each of the times measured with a different method (M_j), resulting in nine different true scores (T_j). Respondents are required to answer three times to the same questions first with method 1 (M_1), second with method 2 (M_2), and third with method 3 (M_3). Overall, 9 survey questions are evaluated in each experiment and their responses are identified as Y_{ij}.

To avoid memory effects, reduce costs, and have shorter questionnaires, among others, Saris, Satorra and Coenders (2004) proposed to randomly assign the respondents to

different split ballot groups, each group getting a different combination of only 2 methods. This SB-MTMM approach was implemented in the ESS, since it allows asking only two times the same respondent the same questions and still all reliability and validity coefficients can be estimated. It is possible to split the sample in different numbers of groups, here we use the 2-group design as presented in Table 1.

Table 1: The 2-group SB-MTMM design

	Main Q.	**Suppl. Q. A**	**Suppl. Q. B**
Group 1	Method 1	Method 2	
Group 2	Method 1		Method 3

In this 2-group design, all respondents answer to the main questionnaire using method 1. Only in the supplementary questionnaires the two groups get different methods. These different methods are summarized in Table 5.

2. Data and analysis

2.1. Country-language groups

In Round 7, the SB-MTMM experiments were conducted in 21 of the ESS participating countries. Since the language can have an impact on the validity of the data (Saris & Gallhofer, 2007; Zavala-Rojas, 2016), in multilingual countries, the data was not only split by country but also by language. Table 2 summarizes the combinations of countries and languages available in Round 7. In brackets are the short names used for the country-language combinations for the remainder of the report[1].

Table 2: ESS Round 7 countries and languages available

Country	Language 1	Language 2	Language 3
Austria	German [ATGER]		
Belgium	Dutch [BEDUT]	French [BEFRE]	
Switzerland	German [CHGER]	French [CHFRE]	Italian*
Czech Republic	Czech [CZCZE]		
Germany	German [DEGER]		
Denmark	Danish [DKDAN]		
Estonia	Estonian [EEEST]	Russian [EERUS]	
Spain	Spanish [ESSPA]	Catalan*	
Finland	Finnish [FIFIN]	Swedish*	
France	French [FRFRE]		
United Kingdom	English [GBENG]		
Hungary	Hungarian [HUHUN]		
Ireland	England [IEENG]		
Israel	Arabic [ILARA]	Hebrew [ILHEB]	Russian*
Lithuania	Lithuanian [LTLIT]	Russian*	
Netherlands	Dutch [NLDUT]		
Norway	Norwegian [NONOR]		
Poland	Polish [PLPOL]		
Portugal	Portuguese [PTPOR]		
Sweden	Swedish [SESWE]		
Slovenia	Slovene [SISLV]		

The cases with an asterisk (*) were not analysed because the sample size was too small (<100 cases per split-ballot group). Thus, taking into account the significant country-language combinations, we could analyse the 25 country-language groups presented in Table 2 in brackets. The sample size of each group per country-language group is summarized in Table 3.

[1] The first two letters belong to the country ISO code and the last three letters belong to the corresponding language ISO code.

Table 3: ESS Round 7 sample sizes per country-language group

Country-Language	SB-Group 1	SB-Group 2	Total cases
ATGER	904	891	1795
BEDUT	508	470	978
BEFRE	408	375	783
CHFRE	184	160	344
CHGER	542	574	1116
CZCZE	1073	1075	2148
DEGER	1518	1512	3030
DKDAN	749	753	1502
EEEST	616	646	1262
EERUS	378	411	789
ESSPA	879	901	1780
FIFIN	1009	964	1973
FRFRE	956	961	1917
GBENG	1177	1084	2261
HUHUN	736	716	1452
IEENG	1169	1221	2390
ILARA	231	230	461
ILHEB	1023	1002	2025
LTLIT	1020	1074	2094
NLDUT	854	865	1719
NONOR	697	698	1395
PLPOL	801	808	1609
PTPOR	618	640	1258
SESWE	889	891	1780
SISLV	610	611	1221

2.2. Experimental questions

In Round 7, the following three SB-MTMM experiments were implemented: 1) attitudes towards immigration: qualification for entry or exclusion of immigrants ("Immigration"), 2) external political efficacy or system Responsiveness ("system Responsiveness"), and internal political efficacy or subjective Competence ("Subjective Competence")

In Table 4, the wording of the survey questions' requests for an answer for each experiment is presented, as they are included in the Round 7 questionnaires. Each experiment is measured by three traits. Each trait is formulated three times in the questionnaire as a survey question: once in the main questionnaire and twice in the supplementary questionnaire. By the 2 group SB design, respondents get twice the same survey question during the interview. Each time using a different answer scale. The randomization has been of the questions by group and method is presented in Table 1.

Table 4: Survey questions included in the ESS Round 7 main and supplementary questionnaires

Experiment	Trait[2]	ID[3]	Variable[4]	Questions' request for an answer
Immigration	Ability to speak language	D2 IF1 IF10	qfimlng testf1 testf10	How important do you think being able to speak the country's language should be in deciding whether someone born, brought up and living outside should be able to come and live here.
	To be white	D4 IF2 IF11	qfimwht testf2 testf11	How important you think being white should be in deciding whether someone born, brought up and living outside should be able to come and live here?
	Commitment way of life	D6 IF3 IF12	qfimcmt testf3 testf12	How important you think being committed to the way of life should be in deciding whether someone born, brought up and living outside should be able to come and live here?
System Responsiveness	People have a say about the government	B1a IF4 IF13	psppsgv testf4 testf13	How much would you say that the political system in [country] allows people like you to have a say in what the government does?
	People have an influence in politics	B1c IF5 IF14	psppipl testf5 testf14	How much would you say that the political system in [country] allows people like you to have an influence on politics?
	Politicians care what people think	B1e IF6 IF15	ptcpplt testf6 testf15	How much would you say that politicians care about what people like you think?
Subjective Competence	Ability to take an active role in a group about political issues	B1b IF7 IF16	actrolg testf7 testf16	How able do you think you are to take an active role in a group involved with political issues?
	Confidence in ability to participate in politics	B1d IF8 IF17	cptppol testf8 testf17	How confident are you in your own ability to participate in politics?
	Facility to take part in politics	B1f IF9 IF18	etapapl testf9 testf18	How easy do you find it personally to take part in politics?

[2] The Trait column indicates the names given, in this report, to the set of questions measuring the same single concepts.
[3] The ID column provides the identifier name given in the ESS questionnaires to the questions; the Bs and Ds refer to the questions presented in the main questionnaire, while the IFs refer to those presented in the supplementary.
[4] The Variable column indicates the names given in the ESS dataset for each of the questions used.

In Table 4 we present the questions' request for an answer, which do not change across the different variables, except for the experiment *immigration.* In this experiment, the three questions in the main questionnaire are presented in the form of a battery, i.e. the request for an answer appears only once and then the different items are presented as short statements. The six other remaining questions are presented in the supplementary questionnaire as direct requests (see Appendix A for the exact wording of each survey question).

In each experiment, each of the three requests for an answer presented in Table 4 are asked using three different answer scales. The variations in the design of the answer scales used in each experiment are presented in Table 5. Each of the requests for an answer was first presented using method 1 in the main questionnaire (for the complete sample), and later using method 2, for a random half of the sample, and method 3, for the other half, in the supplementary questionnaire.

The *immigration* questions are presented to the respondents, for the first time, in the main questionnaire in a battery of related questions using a unipolar 11-point item-specific (IS)[5] scale, which is presented in a horizontal layout and with only the end-points labelled (method 1). In the supplementary questionnaire, these questions are repeated but presented as direct requests, for one random half of the sample using the same scale in method 1 but in a vertical layout (method 2), and for a second random half of the sample maintaining the vertical layout but fully labelling all points (method 3). Because all points are labelled in Method 3, and the endpoint labels had been worded in a bipolar way, an explicit neutral category 'Neither unimportant nor important' was included. However, theoretically, unipolar concepts such as importance, should not have a neutral category (Dolnicar, 2013), and doing so can affect the meaning one attaches to these options (Alwin, 2007). Given these variations, our expectations are that method 3 shall result in lower measurement quality compared to methods 1 and 2. We also expect to find small differences between methods 1 and 2 since scales only vary in the layout display and whether the questions were presented within a battery or as single questions. The French in Switzerland group (CHFRE) did not find an appropriate way to formulate method 3 in a fully-labelled scale and instead used an end-points labelled

[5] An IS response scale is used to ask a direct question in a simple and informative form. This type of scale is called item-specific because the categories used to express the opinion are exactly those answers we would like to obtain for this question (Saris et al., 2010).

scale from ‘Pas du tout important’ to ‘Extrêmement important’. Although the labels used in method 2 were ‘Très peu important’ to ‘Très important’, we expect those methods to be more similar in CHFRE. This group will be excluded from the overall groups comparison.

Table 5: Answer scale formulations of the items included in the ESS Round 7 main and supplementary questionnaires

Experiment	Method 1	Method 2	Method 3
Immigration	*D2, D4, D6*	*IF1, IF2, IF3*	*IF10, IF11, IF12*
	Extremely unimportant 0 … Extremely important 10	0 Extremely unimportant … 10 Extremely important	0 Extremely unimportant 1 Very unimportant 2 Quite unimportant 3 Rather unimportant 4 A bit unimportant 5 Neither unimportant nor important 6 A bit important 7 Rather important 8 Quite important 9 Very important 10 Extremely important
System Responsiveness	*B1a, B1c, B1e*	*IF4, IF5, IF6*	*IF13, IF14, IF15*
	Not at all 0 … Completely 10	1 Not at all 2 Very little 3 Some 4 Quite a lot 5 A lot	1 Not at all 2 Very little 3 Some 4 A lot 5 A great deal
Subjective Competence	*B1b, B1d, B1f*	*IF7, IF8, IF9*	*IF16, IF17, IF18*
	Not at all able/confident /easy 0 … Completely able/confident/ easy 10	Completely unable/ unconfident/ difficult 0 … Completely able/ confident/ easy 10	1 Not at all able/confident/easy 2 A little able/confident/easy 3 Quite able/confident/easy 4 Very able/confident/easy 5 Completely able/confident/easy

The *system responsiveness* questions presented to the respondents, for the first time, in the main questionnaire use a unipolar and horizontal 11-point IS scale, where only the two end-points are labelled as fixed reference points (method 1). In the supplementary questionnaire, these same items are first presented using a unipolar and vertical 5-point IS scale, fully labelled and with only one fixed reference point (method 2), and second with a very similar unipolar and vertical 5-point IS scale, also fully labelled and with only one fixed reference point (method 3). The fixed reference point in both cases is the first label 'Not at all'. The only differences between these two last scales is that the last two labels in method 2 are 'Quite a lot' and 'A lot' and in method 3 are 'A lot' and 'A great deal', respectively. Given this, we expect small differences between method 2 and method 3, and larger differences between these and method 1.

In the *subjective competence* experiment the scales' labels are specifically developed for each trait. Theoretically, the first two traits, 'ability' and 'confidence', are unipolar concepts, while the third, 'facility', is a bipolar concept. Taking this into account, the questions in the main questionnaire use a unipolar 11-point IS scale, which is only labelled at the two end-points as fixed reference points (method 1). In the first repetition of the supplementary questionnaire, the scale also uses an 11-point IS scale, only labelled at the two end-points as fixed reference points. However, this time the scale has a within trait variation. For the first two traits, the scales provide a "fake-bipolar" formulation, since 'unable' and 'unconfident' mean the lack of ability or confidence, and the third uses a truly bipolar formulation, i.e. 'difficult' towards 'easy'. We use the term "fake-bipolar" to refer to theoretically unipolar scales using a bipolar formulation. The distinction in method 2 between "fake-bipolar" and bipolar scales is relevant because it has been argued that "one common error is to measure unipolar attributes on a bipolar answer scales" (Rossiter, 2011, p. 105). Moreover, because this third trait 'Facility to take part in politics' measures a theoretically bipolar concept, its scale has an implicit neutral category, and thus a third fixed reference point. Summarizing, the first repetition of the supplementary questionnaire consists, on the one hand, on a "fake-bipolar" 11-point IS scale, only labelled at the two end-points as fixed reference points (method 2 for traits 1 and 2), and on the other hand, on a bipolar uses an 11-point IS scale, only labelled at the two end-points, and with 3 fixed reference points (method 2 for trait 3). Finally, those questions are again provided in the supplementary questionnaire using a third formulation of the scale: a unipolar 5-point IS scale, fully

labelled and with two fixed reference points (method 3). Given these variations, we expect method 2 to have lower qualities than the other methods, especially for traits 1 and 2. We also expect differences between methods 1 and 3 since they vary in the number of categories, the layout display and the number of labels provided.

The exact formulation of the questions and scales per experiment and method is provided in Appendix A.

2.3. SB-MTMM analysis procedure

We analyse each of the three experiments by country-language-groups. The SB-MTMM structural equation models are estimated in LISREL 8.72 using the Maximum Likelihood (Jöreskog & Sörbom, 1996). In order to test if there are misspecifications, we use the JRule software (van der Veld, Saris, & Satorra, 2008) based on the procedure developed by Saris, Satorra and van der Veld (2009). JRule has the advantage of taking into account both type I and type II errors (i.e. analysis of the power), but also to test the misspecifications at the parameter level, i.e. test if each specific parameter is misspecified and not the model as a whole. This leads in many cases to the introduction of corrections with respect to the general model presented earlier in Figure 2. Principally, the changes on the SB-MTMM analyses consist in 1) allowing unequal effects of one method on the different traits, or 2) adding a correlation between two methods when they are very similar. To solve cases of negative variances or non-convergence, we sometimes need to fix one or two of the method variances to zero when it is not significantly different from zero. In order to be able to compare results across countries and languages, we first consider making the same corrections in all country-language groups for one specific experiment. However, this is not always possible and sometimes we have to allow differences across country-language groups. The final model adjustments done in each group and experiment are summarized in the Appendix B, together with the model fit indices, i.e. degrees of freedom (df) and chi-square (χ^2), and the still misspecified parameters detected in JRule.

3. Results

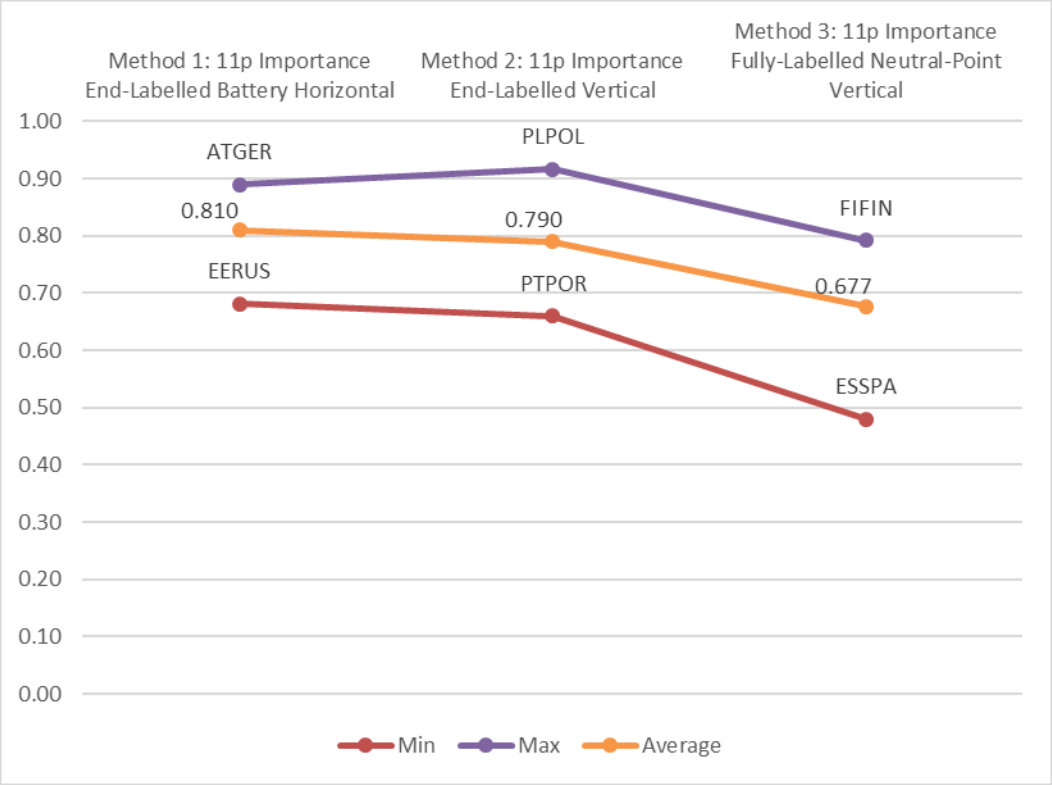
In this section, the results of the Round 7 SB-MTMM analyses will be presented. We will provide the impact on measurement quality (q^2) of the different methods used by experiment, trait and country. To interpret the following results, we can use the thresholds specified for Cronbach's alpha: $\alpha \geq 0.9$ is excellent; $0.8 \leq \alpha < 0.9$ is good; $0.7 \leq \alpha < 0.8$ is acceptable; $0.6 \leq \alpha < 0.7$ is questionable, $0.5 \leq \alpha < 0.6$ is poor, and $\alpha < 0.5$ is unacceptable

Given that these experiments vary a specific set of design characteristics and that they are implemented in three specific concepts, the results cannot be generalized nor extrapolated. They can only serve to have an idea of which types of formulations work best for the survey questions under evaluation.

3.1. Immigration experiment

The *immigration* experiment compares three unipolar 11-point IS scales labelled from "Extremely unimportant" to "Extremely important". First, the scale is presented within a battery of questions using a horizontal layout and with only the end-points labelled (method 1). In the supplementary questionnaire, the scale is provided for single questions using a vertical layout: for a random half of the sample, the scale provides only the end-points labelled (method 2), while for the other half the scale is fully labelled and includes an explicit neutral point (method 3), except for the French in Switzerland group (CHFRE), which provided also a scale with only the end-points labelled for method 3. The overall *immigration* results per method are presented in Figure 3. The overall average quality (q^2) is around 0.76.

Figure 3: Average quality (q^2), over all country-language groups and traits, per method with its country-language specific extremes in the immigration experiment[6]



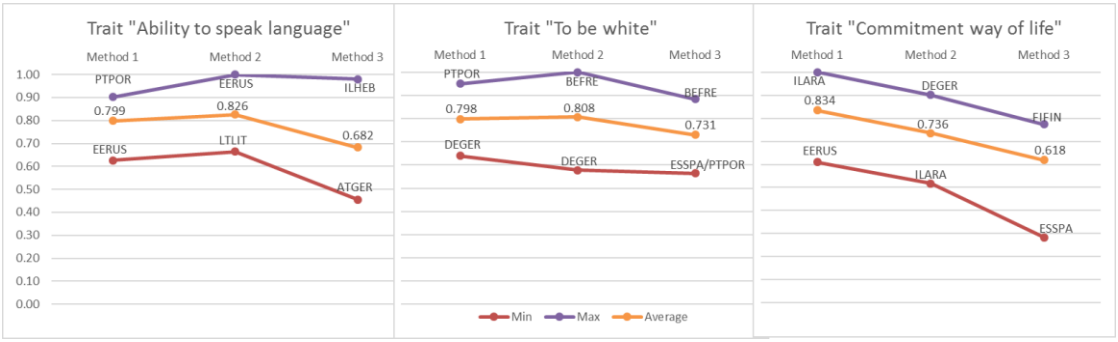
The average qualities in method 1 and method 2 presented in Figure 3 are not significantly different[7], while method 3 is significantly lower. Overall, the differences across methods of the average qualities are up to 0.13.

In the same figure, we can observe the minimum and maximum quality values obtained per method and country-language group. For instance, for method 1, where the average quality is 0.81, Austria (ATGER) obtained the highest quality overall groups (0.89) and Russian in Estonia (EERUS) the minimum (0.68). The range between the minimum and the maximum quality obtained across the different country-language groups is, in all methods, between 0.21 and 0.31. The average quality for the different traits and methods and over all country-language groups is presented in Figure 4.

[6] The country-language group CHFRE not included because it deviated from the original formulation of method 3.

[7] Test of significance: two-tailed z-test for two means, with a significance level (α) of 0.05.

Figure 4: Average quality (q^2), over all country-language groups, per trait and method and extremes in the immigration experiment[8]



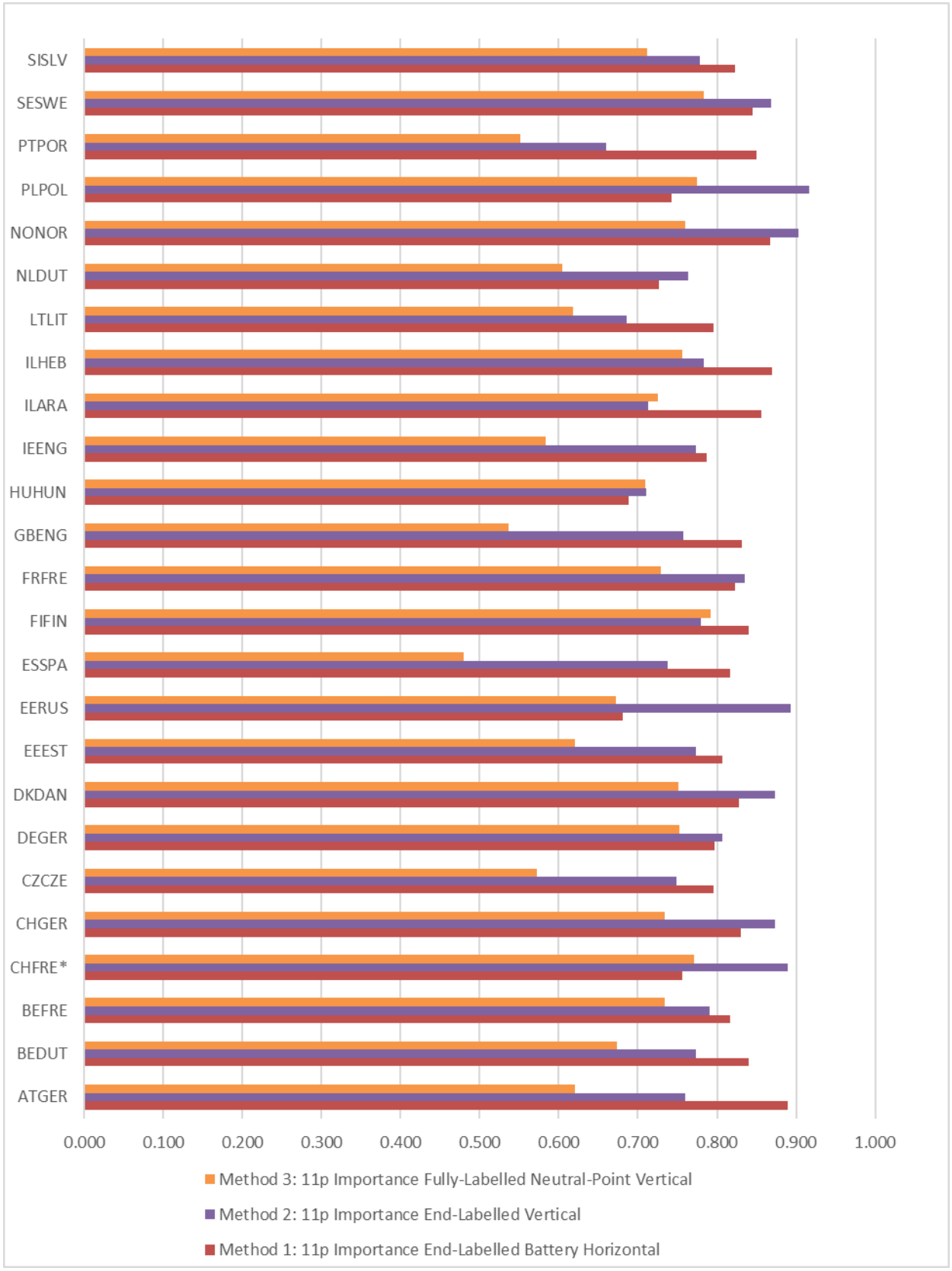
In Figure 4, we observe for the three traits 'Ability to speak language', 'To be white' and 'Commitment way of life' that the average results are stable only within the first two traits. Figure 4 also illustrates the big differences across country-language groups, which indicates that it is important considering the country-language specific results. Moreover, we can observe that the biggest difference between the minimum and the maximum quality comes from the trait 'Ability to speak language' with method 3 and the lower from the same trait with method 1, which could suggest that method 3 is more sensitive to different interpretations. However, we would need to repeat this several times to see the robustness of this conclusion.

Moreover, the average quality of the questions across methods for the different country-language groups is presented in Figure 5, which shows that there are large differences in quality across the different country-language groups. The general trend is that method 1 and method 2 perform better than method 3, except for Finland (FIFIN), Hungary (HUHUN), Arabic in Israel (ILARA) and Poland (PLPOL), where method 3 obtained a higher quality than method 1 and/or method 2. That means that for most groups the end-point labelled scales have higher quality compared to the fully labelled scale.

In the Swiss French version (CHFRE), the 11-point end-labelled scale ranging from 'Très peu important' to 'Très important' (method 2) has a higher quality, more than 0.1 difference, than the same scale ranging from 'Pas du tout important' to 'Extrêmement important' (method 3).

[8] The country-language group CHFRE not included because it deviated from the original formulation of method 3.

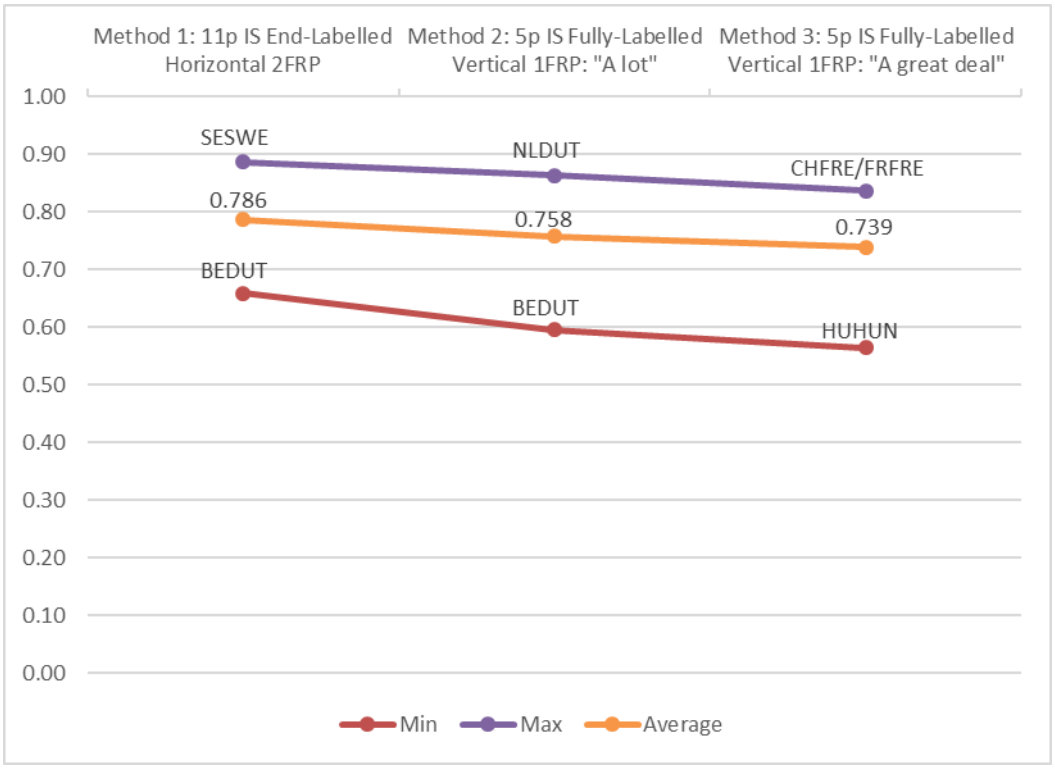
Figure 5: Average quality (q^2), over all traits, per country-language group and method in the immigration experiment



3.2. System responsiveness experiment

The *system responsiveness* experiment allows studying the effect of the varying number of answer categories (11 points versus 5 points), the use of verbal labels (end-labelled versus fully-labelled), the number of fixed reference points (2FRP versus 1FRP) and the layout display (horizontal versus vertical) on the quality of unipolar IS scales. This can be done by comparing a horizontal 11-point IS scale, end-labelled and with two fixed reference points (method 1), and two vertical 5-point IS scales, fully-labelled and with one fixed reference points (methods 2 and 3). As indicated in Table 5, the only difference between methods 2 and 3 are the labels used, i.e. “Quite a lot” and “A lot” in method 2 and “A lot” and “A great deal” in method 3. The overall *system responsiveness* results per method are presented in Figure 6. The overall the quality (q^2) is also around 0.76.

Figure 6: Average quality (q^2), over all country-language groups and traits, per method with its country-language specific extremes in the system responsiveness experiment

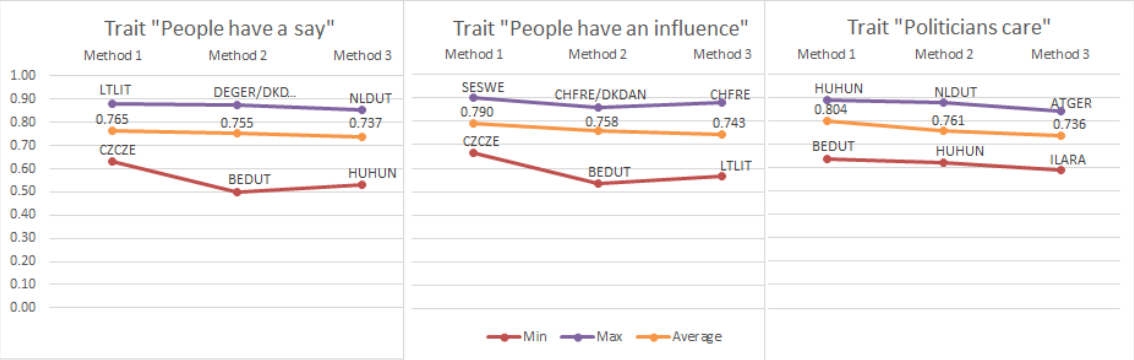


The average qualities, reported in Figure 6, between method 1 and method 2 and method 2 and method 3 are not significantly different but differences are significant between 1 and 3. Overall, the differences across methods of the average qualities is up

to 0.05.

In the same figure, we can observe the minimum and maximum quality values obtained per method and country-language group. For instance, for method 1the average quality is 0.79, ranging from 0.66 in France (FRFRE) to 0.89 in Sweden (SESWE). The range between the lowest and the highest quality obtained across the different country-language groups is, in all methods, between 0.23 and 0.27.

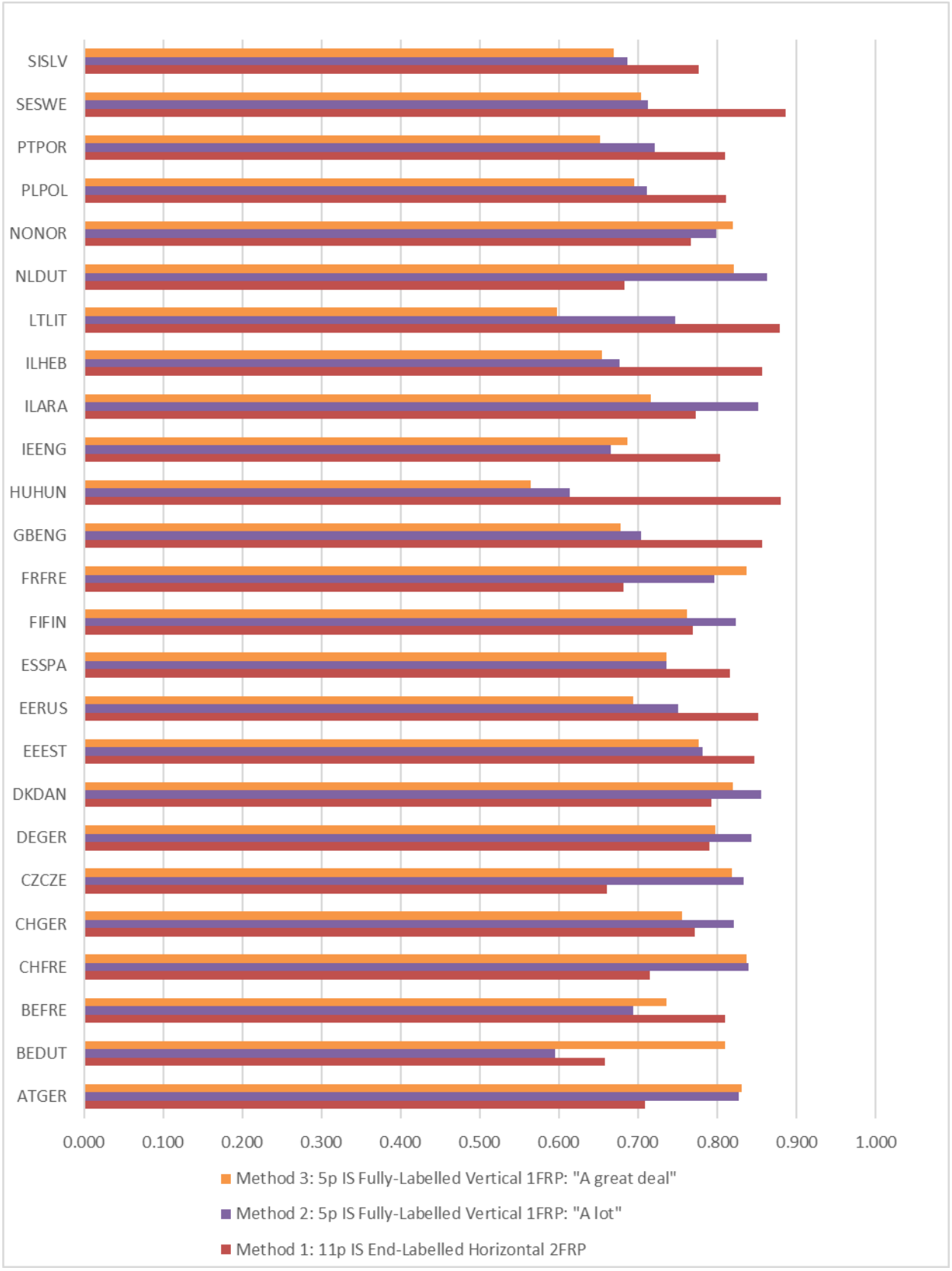
Figure 7: Average quality (q^2), over all country-language groups, per trait and method and extremes in the system responsiveness experiment



The average quality for the different traits and methods and over all country-language groups is presented in Figure 7. We can observe that the average results are stable within and across traits, the differences are not significant. Moreover, we can observe that the bigger differences between the minimum and the maximum quality comes from the trait ‘People have a say about the government’ with method 2 and the lower from the trait ‘Politicians care what people think’ with method 1. Figure 7 also illustrates the high differences across country-language groups, which indicates that it is important taking into account the country-language specific results, as presented in Figure 8.

Most country-language groups, in Figure 8, do not present relevant differences between method 2 and 3 (less than 0.1 difference), except for Dutch in Belgium (BEDUT), Arabic in Israel (ILARA) and Lithuania (LTLIT). Moreover, while in groups such as BEFRE, EEEST, EERUS, ESSPA, GBENG, HUHUN, IEENG, ILHEB, LTLIT, PLPOL, PTPOR, SESWE and SISLV the higher quality is obtained by method 1, in the other groups methods 2 and/or 3 performed better than method 1. This means that for about half of the groups the scale with higher quality is the 11-point scale, while for the other half the scale with higher qualities are the 5-point scales.

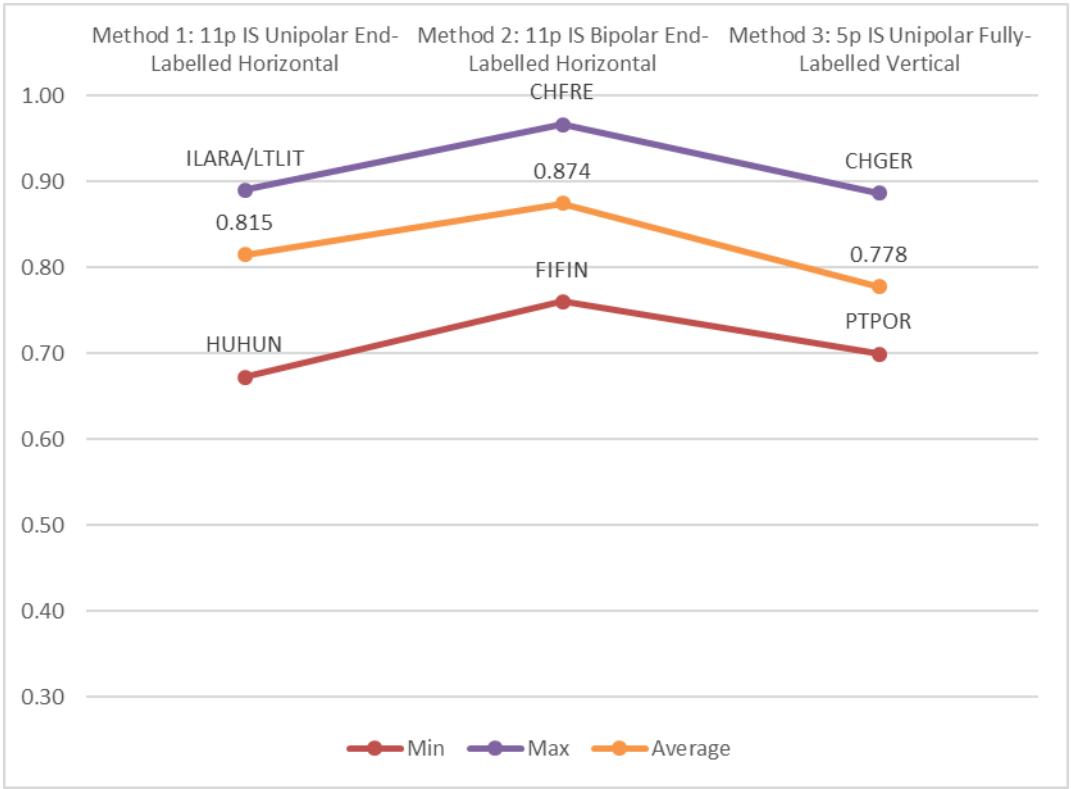
Figure 8: Average quality (q^2), over all traits, per country-language group and method in the system responsiveness experiment



3.3. Subjective competence experiment

The *subjective competence* experiment allows studying the effect of the varying number of answer categories (11 points versus 5 points), the use of verbal labels (end-labelled versus fully-labelled), the polarity of the scale used (unipolar, "fake-bipolar"[9] or bipolar versus unipolar), and the layout display (horizontal versus vertical) on the quality of unipolar IS scales. This can be done by comparing a unipolar and horizontal 11-point IS scale only labelled at the end-points (method 1), a "fake-bipolar" versus bipolar and horizontal 11-point IS scale also only labelled at the end-points (method 2) and a unipolar and vertical 5-point IS fully-labelled scale (method 3). The overall *subjective competence* results per method are presented in Figure 9. The overall the quality (q^2) is higher than in the other two experiments, 0.82.

Figure 9: Average quality (q^2), over all country-language groups and traits, per method with its country-language specific extremes in the subjective competence experiment



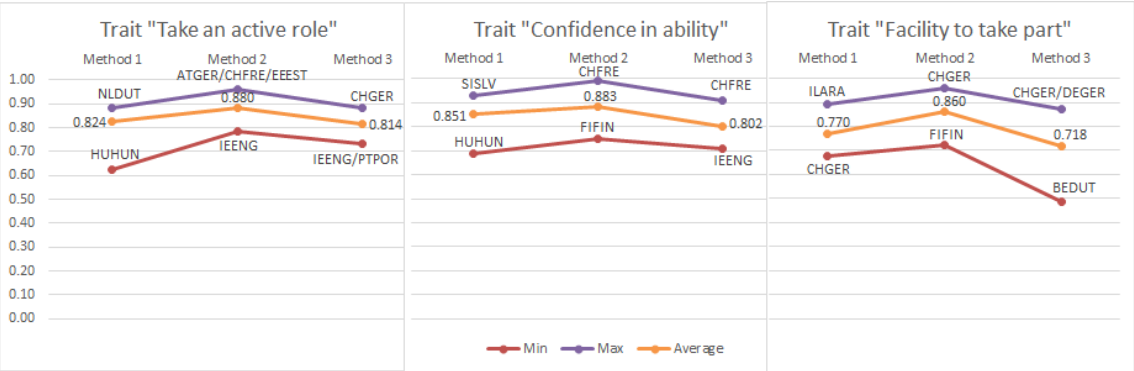
The average qualities, reported in Figure 9, are all significantly different from each

[9] As noted above, traits 1 and 2 cannot be formulated in a real bipolar connotation. We refer to these as "fake bipolar".

other. It shows that overall the bipolar 11-point end-labelled formulation of the scale (method 2) performs better than the unipolar formulation (method 1), and than the vertically displayed, unipolar 5-point fully-labelled scale (method 3). Overall, the differences across methods of the average qualities are up to 0.1.

In this same figure, we can observe the minimum and maximum quality values obtained per method and country-language group. For instance, for method 1, where the average quality is 0.82, both Arabic in Israel (ILARA) and Lithuania (LTLIT)[10] obtained the highest quality overall groups (0.89) and Hungary (HUHUN) the lowest (0.67). The difference between the lowest and the highest quality obtained across the different country-language groups is, in all methods, 0.2. Even if the variations in one method should be the same for all traits using this method, we have observed that questions using method 2, the scale gets a different connotation depending on the trait being measured. For traits 1 and 2, which measure theoretically unipolar concepts, a "fake bipolar" scale is used, while for trait 3, which measures a theoretically bipolar concept, the scale used is bipolar. To see if it has an impact on the quality, it is especially interesting to look at the results by trait as presented in Figure 10.

Figure 10: Average quality (q^2), over all country-language groups, per trait and method and extremes in the subjective competence experiment



In Figure 10, we can observe for the questions measuring the three traits 'Take an active role in a group', 'Confidence in ability to participate' and 'Facility to take part' that the average results are stable within and across traits: method 2 performs better than method 1 and method 1 performs better than method 3. Moreover, we can observe that the bigger differences between the minimum and the maximum quality comes from the trait

[10] For Lithuania, in this experiment, it should be noticed that the model fit indices reported the worst fit for this model compared to all other analyses performed with the ESS Round 7 SB-MTMM data.

‘Facility to take part’ with method 3 and the lower from the trait ‘Take an active role in a group’ with method 3.

Focusing on method 2, i.e. the “fake bipolar” versus bipolar scale, we see that the average quality is higher, around 0.88, for the two questions measuring theoretical unipolar concepts (traits 1 and 2) and a bit lower for the theoretically bipolar question (trait 3), 0.86. Thus, this suggests that the use of “fake bipolar” scales does not reduce the quality of these questions. However, problems in the translation process of these scales were reported for some of the languages. Not all languages were able to formulate such bipolar formulations of the scale for the unipolar concepts ‘ability’ (trait 1) and ‘confidence’ (trait 2). Because of that, we looked at the relationship between the quality of this experiment and the translation of the bipolar scale (method 2) in the different countries.

More specifically, some countries indicated difficulties of translating the bipolar scale (method 2) from the supplementary questionnaire for the first two items IF7 (trait ‘Active role’) and IF8 (trait ‘Participate’). In the English version of the questionnaire IF7 used a scale from ‘Completely unable’ to ‘Completely able’ and IF8 a scale from ‘Completely unconfident’ to ‘Completely confident’. We have identified, in Appendix C, whether the scale was formulated as “fake-bipolar” or as unipolar in each of the country-language groups and divided them in four possible scenarios:

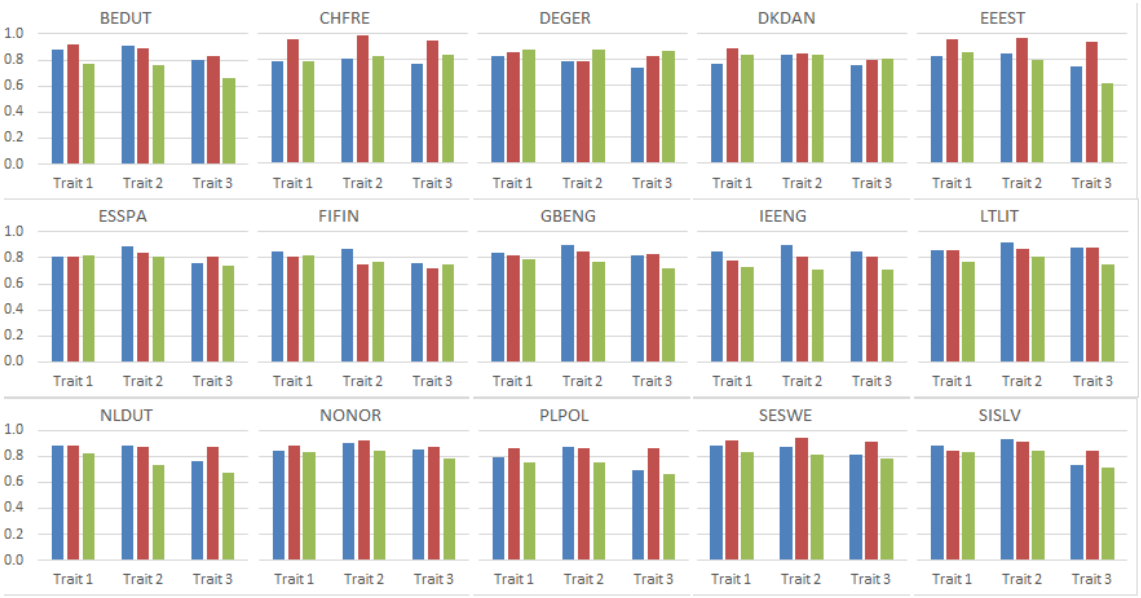
1) Traits 1 and 2 using a “fake-bipolar” formulation, and trait 3 using a bipolar formulation (15 groups)
2) For all scales, a unipolar formulation of the scale was used (5 groups)
3) Only scales in trait 1 or trait 2 uses a unipolar formulation of the scale (5 groups)

If they formulated the scale as unipolar for traits 1 and 2 (scenarios 2 and 3) it means that the country-language group did not find an equivalent bipolar translation. Thus, it makes sense to not only look at the results by method for each country-language group but also by trait.

First, 15 groups belonging to the first scenario, i.e. a “fake-bipolar” scale for traits 1 and 2 in method 2, are presented in Figure 11. Theoretical unipolar concepts should not be measured with bipolar scales (Rossiter, 2011), it could be expected that for these traits the quality was lower or equal (differences < 0.05) in method 2 than in method 1 and/or 3, i.e. lower because the scale is not understood by respondents or equal because the

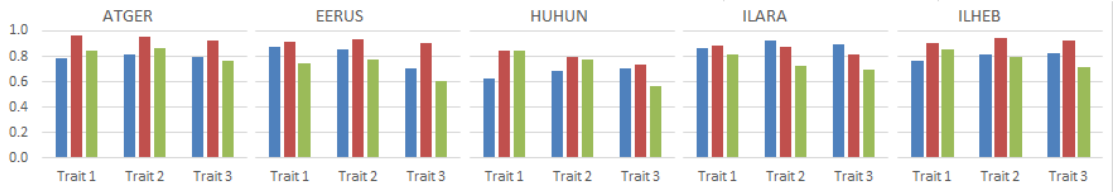
scale is threated as unipolar. However, as shown in Figure 11, for trait 1, this is true in Germany (DEGER), Spain (ESSPA), Finland (FIFIN) , Great Britain (GBENG), Ireland (IEENG), Netherlands (NLDUT), Norway (NONOR), and Slovenia (SISLV), 8 out of the 15 groups. For trait 2, this is only true in Germany (DEGER), Denmark (DKDAN), Spain (ESSPA), Finland (FIFIN) and Great Britain (GBENG), 5 out of the 15 groups.

Figure 11: Quality (q^2), for each country-language group in Scenario 1, per trait and method in the subjective competence experiment



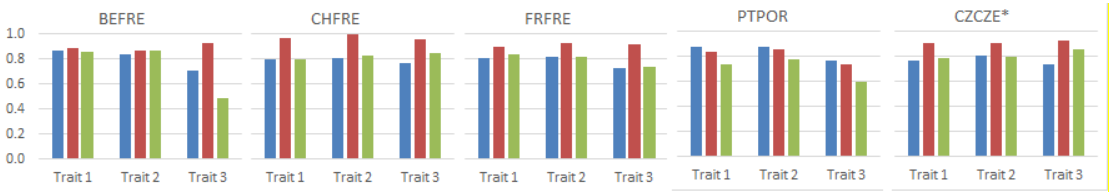
Second, five groups belonging to the second scenario, i.e. a unipolar scale for traits 1 and 2 in method 2, are presented in Figure 12. In this case, we would expect no differences (differences < 0.05) between method 1 and method 2. If method 2 is formulated as unipolar the scales for traits 1 and 2 will be no different from those in method 1, i.e. a unipolar 11-point IS scale, which is only labelled at the two end-points as fixed reference points However, this is only true for both traits in Arabic in Israel (ILARA), and for trait 1 in Russian in Estonia (EERUS).

Figure 12: Quality (q^2), for each country-language group in Scenario 2, per trait and method in the subjective competence experiment



Third, five more groups belonging to the third scenario, i.e. a "fake-bipolar" scale for trait 1 and a unipolar for trait 2[11] in method 2, are presented in Figure 12. In this case, we would expect that method 2, for trait 1 works worse or equal to method 1 and that method 2 for trait 2 has a similar quality to method 1. That is true for: French in Belgium (BEFRE) and Portugal (PTPOR). In the Czech (CZCZE) case, we similarly expect that, method 2 in trait 1 is worse or equal to method 1 and that method 2 for trait 2 is similar to method 1, but we find the opposite.

Figure 13: Quality (q^2), for each country-language group in Scenario 3, per trait and method in the subjective competence experiment



Overall, we cannot observe any clear effect of the different polarity formulations in the quality obtained.

[11] Only the country-group CZCZE, provided a unipolar scale for trait 1 and a "fake-bipolar" scale for trait 2.

4. Main conclusions from SB-MTMM experiments

Overall the quality of the three experiments was between acceptable and good, 0.76 and 0.82. Following, we summarize the main conclusions by experiment.

First, the *immigration* experiment, compared three unipolar 11-point IS scales labelled from "Extremely unimportant" to "Extremely important". The Method 1 scale was presented in a battery of questions, in a horizontal layout, and only the end-points were labelled. Method 2 was a vertical scale with only end-points labelled and Method 3 was a fully labelled, vertical scale with an explicit neutral point. Overall, the pattern shows, as expected, comparable qualities among Method 1 and Method 2 and significant smaller quality for Method 3. However, we cannot disentangle if the quality is negatively affected using fully labelled points or the use of a neutral category for the theoretically unipolar concept 'importance' or both.

Second, the *system responsiveness* experiment, was designed with the purpose to test three alternative formulations of the items, in relation to the different response categories, the use of verbal labels, the number of fixed reference points and the layout display. For this, the experiment compared: 1) a horizontal 11-point IS scale, end point-labelled and with two fixed reference points, 2) a vertical 5-point IS scale, fully-labelled and with one fixed reference point (labelling the end-points as "Quite a lot" and "A lot"), and 3) another vertical 5-point IS scale, fully-labelled and with one fixed reference points (but this time labelling the end-points as "A lot" and "A great deal"). As expected, we did not find differences between the scales used in methods 2 and 3. Although we might have expected higher differences between the 11-point (method 1) and 5-point scales (methods 2 and 3), the qualities were quite similar too.

Third, from the *subjective competence* experiment, a unipolar and horizontal 11-point IS scale only labelled at the end-points (Method 1), a "fake-bipolar" and horizontal 11-point IS scale also only labelled at the end-points (Method 2) and a unipolar and vertical 5-point IS fully-labelled scale (Method 3) have been compared. It has been shown that, over all traits and in most country-language groups, one can get a higher quality by using the 11-point bipolar scale. As the formulation of the bipolar scale (method 2) was not possible in all languages, country-by-country analysis was necessary, showing large differences between countries in the quality of the different scales. First, comparing countries who designed method 2 using a "fake-bipolar" formulation, we see that, on

average, method 1 and method 2 are not significantly different and that method 3 has a lower quality. Second, comparing countries who designed method 2 using a unipolar formulation, we see that, on average method 2 has a significant higher quality than method 1 or method 3. Third, comparing countries who designed method 2 in trait 1 with a "fake-bipolar" formulation scale and trait 2 with a unipolar formulation scale, we see that, on average, method 2 is significantly higher than methods 1 and 3.

Finally, over all experiments, the quality reached a maximum of 0.97, in French in Switzerland (CHFRE) for method 2 in the *subjective competence experiment* and a minimum of 0.48, in Spain (ESSPA) for method 3 in the *immigration* experiment. Moreover, we have observed that there are large deviations not only across countries but also within countries for different languages. Groups might therefore not be comparable if they are not corrected for measurement error. Correction for measurement error can be done using the quality estimation of this analyses. However, to ensure comparability of the groups, an equivalence test across groups should be conducted.

To conclude, it is important to highlight that these findings are specific for the topics analyzed and the methods used. To be able to draw general conclusions, more topics would need to be studied, to get a better picture of the effect of methods for different topics.

References

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

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrices. *Psychological Bulletin*, *56*(2), 81–105.

Dolnicar, S. (2013). Asking Good Survey Questions. *Journal of Travel Research*, *52*(5), 551–574. https://doi.org/10.1177/0047287513479842

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's Reference Guide*. Uppsala, Sweden: Scientific Software International.

Rossiter, J. R. (2011). *Measurement for the Social Sciences: The C-OAR-SE Method and Why it Must Replace Psycometrics*. New York, NY, US: Springer-Verlag.

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

Saris, W. E., & Gallhofer, I. N. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ, US: John Wiley and Sons, Inc.

Saris, W. E., Satorra, A., & Coenders, G. (2004). A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design. *Sociological Methodology*, *34*(1), 311–347. https://doi.org/10.1111/j.0081-1750.2004.00155.x

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705510903203433

van der Veld, W. M., Saris, W. E., & Satorra, A. (2008). Judgement rule aid for structural equation models version 3.0.4 beta.

Zavala-Rojas, D. (2016). *Measurement equivalence in multilingual comparative survey*

research. Universitat Pompeu Fabra.

Appendix A: Questionnaire questions' formulations by experiment and method

- *EXPERIMENT 1: ATTITUDES TOWARDS IMMIGRATION*

Method 1

D2 …be able to speak [country's official language(s)]?

Extremely unimportant									Extremely important		(Don't know)
00	01	02	03	04	05	06	07	07	09	10	88

D4 ...be white?

Extremely unimportant									Extremely important		(Don't know)
00	01	02	03	04	05	06	07	07	09	10	88

D6 ...be committed to the way of life in [country]?

Extremely unimportant									Extremely important		(Don't know)
00	01	02	03	04	05	06	07	07	09	10	88

Method 2

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. The first few questions are about this issue.

IS1 How important do you think being able to speak [country's official language(s)] should be in deciding whether someone born, brought up and living outside [country] should be able to come and live here. Please tick one box.

0 Extremely unimportant

1

2

3

4

5

6

7

8

9

10 Extremely important

IS2 And how important you think being white should be in deciding whether someone born, brought up and living outside [country] should be able to come and live here? Please tick one box.

0 Extremely unimportant

1

2

3

4

5

6

7

8

9

10 Extremely important

IS3 Now, how important you think being committed to the way of life in [country] should be in deciding whether someone born, brought up and living outside [country]

should be able to come and live here? Please tick one box.

0 Extremely unimportant

1

2

3

4

5

6

7

8

9

10 Extremely important

Method 3

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. The first few questions are about this issue.

IS10 How important do you think being able to speak [country's official language(s)] should be in deciding whether someone born, brought up and living outside [country] should be able to come and live here. Please tick one box.

0 Extremely unimportant

1 Very unimportant

2 Quite unimportant

3 Rather unimportant

4 A bit unimportant

5 Neither unimportant nor important

6 A bit important

7 Rather important

8 Quite important

9 Very important

10 Extremely important

IS11 And how important you think being white should be in deciding whether someone born, brought up and living outside [country] should be able to come and live here? Please tick one box.

0 Extremely unimportant

1 Very unimportant

2 Quite unimportant

3 Rather unimportant

4 A bit unimportant

5 Neither unimportant nor important

6 A bit important

7 Rather important

8 Quite important

9 Very important

10 Extremely important

IS12 Now, how important you think being committed to the way of life in [country] should be in deciding whether someone born, brought up and living outside [country] should be able to come and live here? Please tick one box.

0 Extremely unimportant

1 Very unimportant

2 Quite unimportant

3 Rather unimportant

4 A bit unimportant

5 Neither unimportant nor important

6 A bit important

7 Rather important

8 Quite important

9 Very important

10 Extremely important

- *EXPERIMENT 2: POLITICAL EFFICACY – SYSTEM RESPONSIVENESS*

Method 1

B1a CARD 5 How much would you say the political system in [country] allows people like you to have a say in what the government does? Please use this card.

Not at all									Completely		(Don't know)
00	01	02	03	04	05	06	07	07	09	10	88

B1c CARD 7 And how much would you say that the political system in [country] allows people like you to have an influence on politics? Please use this card.

Not at all									Completely		(Don't know)
00	01	02	03	04	05	06	07	07	09	10	88

B1e CARD 9 How much would you say that politicians care what people like you think? Please use this card.

Not at all									Completely		(Don't know)
00	01	02	03	04	05	06	07	07	09	10	88

Method 2

The next few questions are on a different topic.

IS4 How much would you say the political system in [country] allows people like you to have a say in what the government does? Please tick one box.

1 Not at all

2 Very little

3 Some

4 Quite a lot

5 A lot

IS5 And how much would you say that the political system in [country] allows people like you to have an influence on politics? Please tick one box.

1 Not at all

2 Very little

3 Some

4 Quite a lot

5 A lot

IS6 How much would you say that politicians care what people like you think? Please tick one box.

1 Not at all

2 Very little

3 Some

4 Quite a lot

5 A lot

Method 3

The next few questions are on a different topic.

IS13 How much would you say the political system in [country] allows people like you to have a say in what the government does? Please tick one box.

1 Not at all

2 Very little

3 Some

4 Quite a lot

5 A great deal

IS14 And how much would you say that the political system in [country] allows people like you to have an influence on politics? Please tick one box.

1 Not at all

2 Very little

3 Some

4 Quite a lot

5 A great deal

IS15 How much would you say that politicians care what people like you think? Please tick one box.

1 Not at all

2 Very little

3 Some

4 Quite a lot

5 A great deal

- *EXPERIMENT 3: POLITICAL EFFICACY – SUBJECTIVE COMPETENCE*

Method 1

B1b CARD 6 How able do you think you are to take an active role in a group involved with political issues? Please use this card.

Not at all able										Completely able	(Don't know)
00	01	02	03	04	05	06	07	07	09	10	88

B1d CARD 8 And using this card, how confident are you in your own ability to participate in politics?

Not at all confident										Completely confident	(Don't know)
00	01	02	03	04	05	06	07	07	09	10	88

B1f CARD 10 Using this card, how easy do you personally find it to take part in politics?

Not at all easy										Extremely easy	(Don't know)
00	01	02	03	04	05	06	07	07	09	10	88

Method 2

IS7 How able do you think you are to take an active role in a group involved with political issues? Please tick one box.

Completely unable										Completely able	(Don't know)
00	01	02	03	04	05	06	07	07	09	10	88

IS8 And how confident are you in your own ability to participate in politics? Please tick

one box.

Completely unconfident									Completely confident		(Don't know)
00	01	02	03	04	05	06	07	07	09	10	88

IS9 How easy do you personally find it to take part in politics? Please tick one box.

Extremely difficult									Extremely easy		(Don't know)
00	01	02	03	04	05	06	07	07	09	10	88

Method 3

IS16 How able do you think you are to take an active role in a group involved with political issues? Please tick one box.

1 Not at all able

2 A little able

3 Quite able

4 Very able

5 Completely able

IS17 And how confident are you in your own ability to participate in politics? Please tick one box.

1 Not at all confident

2 A little confident

3 Quite confident

4 Very confident

5 Completely confident

IS18 How easy do you personally find it to take part in politics? Please tick one box.

1 Not at all easy

2 A little easy

3 Quite easy

4 Very easy

5 Completely easy

Appendix B: SB-MTMM model analysis adjustments, fit and JRule evaluation

Experiment	Country-Language group	Model adjustments	df	$\chi 2$	JRule
Immigration	ATGER	2M 0PH44 GA55 GA86	17	44.44 (P=0.98)	3
	BEDUT	2M 0PH44 GA55 GA86	17	22.04 (P=1)	Not relevant
	BEFRE	2M 0PH44 GA55 GA76	17	22.83 (P=1)	2
	CHFRE	1M PH66 0TE88 TE93	20	28.15 (P=1)	1
	CHGER	3M GA65 GA86	16	14.13 (P=1)	3
	CZCZE	3M PH54 GA55 GA76 GA96	14	19.46 (P=1)	Not relevant
	DEGER	3M GA86 GA96 TE52 TE63	14	34.09 (P=1)	2
	DKDAN	3M	14	28.92 (P=1)	Not relevant
	EEEST	3M GA55 GA86	16	23.86 (P=1)	Not relevant
	EERUS	2M 0PH55 0TE44	20	33.42 (P=1)	1
	ESSPA	3M PH54 GA55 GA86 TE93	14	21.27 (P=1)	Not relevant
	FIFIN	3M GA24 GA65 GA86 TE41	14	24.18 (P=1)	2
	FRFRE	2M 0PH44 GA55 GA86	17	8.74 (P=1)	Any
	GBENG	3M GA55 GA86	16	16.61 (P=1)	Not relevant
	HUHUN	3M GA45 GA86	16	19.78 (P=1)	Not relevant
	IEENG	2M 0PH44 GA65 GA86	17	36.48 (P=1)	1
	ILARA	2M 0PH44 0TE33 GA55 GA86	18	37.06 (P=1)	2
	ILHEB	2M 0PH55 0TE33 PH64 GA86 TE71	17	42.52 (P=0.99)	Not relevant
	LTLIT	3M	18	53.14 (P=0.87)	2
	NLDUT	3M PH54 PH64 GA55 GA86	14	38.08 (P=0.99)	Not relevant
	NONOR	3M GA65 TE71	16	21.3 (P=1)	1
	PLPOL	2M 0PH55	19	13.72 (P=1)	Any
	PTPOR	3M GA24	17	11.85 (P=1)	Not relevant
	SESWE	3M	18	44.9 (P=0.98)	2
	SISLV	2M 0PH44 GA55 GA86	17	44.44 (P=0.98)	3

System Responsiveness	ATGER	2M 0PH44 GA86 TE41	17	13.06 (P=1)	0	Any
	BEDUT	3M	18	39.57 (P=1)	0	1
	BEFRE	3M PH54 GA65	16	19.02 (P=1)	0	Not relevant
	CHFRE	2M 0PH44 TE62	18	18.01 (P=1)	0	Not relevant
	CHGER	1M PH44 TE93	19	24.24 (P=1)	0	2
	CZCZE	3M GA96	17	34.08 (P=1)	0	4
	DEGER	3M	18	17.24 (P=1)	0	Any
	DKDAN	3M GA34	17	27.18 (P=1)	0	1
	EEEST	3M	18	32.22 (P=1)	0	1
	EERUS	2M 0PH44 GA55 GA65 GA96	16	33.43 (P=1)	0	Not relevant
	ESSPA	3M GA76	17	17.13 (P=1)	0	Any
	FIFIN	2M 0PH44 GA55 GA96 TE93	16	20.23 (P=1)	0	Not relevant
	FRFRE	2M 0PH55 GA24 GA86 TE93	16	41.88 (P=0.99)	0	2
	GBENG	2M 0PH66 GA65 TE87	17	47.29 (P=0.95)	0	2
	HUHUN	2M 0PH44 GA65 GA86 GA96	16	12.7 (P=1)	0	Any
	IEENG	2M 0PH44 GA65 GA96	17	30.18 (P=1)	0	Not relevant
	ILARA	3M GA34	17	31.3 (P=1)	0	3
	ILHEB	3M GA34 TE93	16	26.01 (P=1)	0	2
	LTLIT	3M GA55	17	37.07 (P=1)	0	2
	NLDUT	3M PH64 GA65 GA96	15	27.68 (P=1)	0	Not relevant
	NONOR	2M 0PH55 GA34 TE54	17	19.06 (P=1)	0	Any
	PLPOL	2M 0PH66 GA65	18	18.75 (P=1)	0	Not relevant
	PTPOR	3M GA76 GA96	16	45.94 (P=0.96)	0	4
	SESWE	2M 0PH44 GA65 GA96 TE41	16	18.91 (P=1)	0	Not relevant
	SISLV	2M 0PH44 GA65 GA96	17	25.35 (P=1)	0	Not relevant

Subjective Competence	ATGER	2M 0PH55 GA96	18	27.75 (P=1)	0	2
	BEDUT	3M GA96	17	34.13 (P=1)	0	Not relevant
	BEFRE	3M GA65 TE93	16	35.06 (P=1)	0	3
	CHFRE	2M 0PH55 GA96	18	23.84 (P=1)	0	Not relevant
	CHGER	3M PH54	17	19.18 (P=1)	0	Not relevant
	CZCZE	3M GA24 GA34 GA45 GA86	14	28.38 (P=1)	0	Not relevant
	DEGER	3M PH54 GA55 GA96	15	29.7 (P=1)	0	Not relevant
	DKDAN	3M PH54 GA45 GA86	15	35.6 (P=1)	0	Not relevant
	EEEST	3M GA34 GA76 GA86 TE93	14	25.04 (P=1)	0	1
	EERUS	3M GA96 TE93	16	19.64 (P=1)	0	Not relevant
	ESSPA	3M PH54 GA65 GA86	15	32.18 (P=1)	0	2
	FIFIN	3M PH54 GA55 GA86	15	33.85 (P=1)	0	2
	FRFRE	3M GA34 GA65	16	26.36 (P=1)	0	Not relevant
	GBENG	2M 0PH44 GA65 GA86	17	38.25 (P=1)	0	Not relevant
	HUHUN	3M GA34 GA55 GA76	15	24.8 (P=1)	0	Not relevant
	IEENG	3M GA65 GA76 GA96	15	14.37 (P=1)	0	Not relevant
	ILARA	2M 0PH44 GA65 GA76 TE93	16	20.87 (P=1)	0	Not relevant
	ILHEB	3M GA34 GA76	16	46.57 (P=0.95)	0	2
	LTLIT	3M GA65	17	114.28 (P=0.00016)	0.027	1
	NLDUT	3M GA45 GA55 GA96 TE93	13	20.27 (P=1)	0	Any
	NONOR	1M PH44 GA14	19	33.34 (P=1)	0	Not relevant
	PLPOL	3M GA65 GA96 TE93	15	23.46 (P=1)	0	Not relevant
	PTPOR	3M GA55	17	39.27 (P=1)	0	1
	SESWE	3M GA96	17	13.43 (P=1)	0	Any
	SISLV	3M GA65 GA96	16	15.21 (P=1)	0	Any

Appendix C: Polarity formulation of the scale in method 2 for each country-language group in the subjective competence experiment.

	Method 2	
	Trait 1 "ability"	**Trait 2 "confidence"**
ATGER	Unipolar	Unipolar
BEDUT	"Fake-bipolar"	"Fake-bipolar"
BEFRE	"Fake-bipolar"	Unipolar
CHFRE	"Fake-bipolar"	Unipolar
CHGER	"Fake-bipolar"	"Fake-bipolar"
CZCZE	Unipolar	"Fake-bipolar"
DEGER	"Fake-bipolar"	"Fake-bipolar"
DKDAN	"Fake-bipolar"	"Fake-bipolar"
EEEST	"Fake-bipolar"	"Fake-bipolar"
EERUS	Unipolar	Unipolar
ESSPA	"Fake-bipolar"	"Fake-bipolar"
FIFIN	"Fake-bipolar"	"Fake-bipolar"
FRFRE	"Fake-bipolar"	Unipolar
GBENG	"Fake-bipolar"	"Fake-bipolar"
HUHUN	Unipolar	Unipolar
IEENG	"Fake-bipolar"	"Fake-bipolar"
ILARA	Unipolar	Unipolar
ILHEB	Unipolar	Unipolar
LTLIT	"Fake-bipolar"	"Fake-bipolar"
NLDUT	"Fake-bipolar"	"Fake-bipolar"
NONOR	"Fake-bipolar"	"Fake-bipolar"
PLPOL	"Fake-bipolar"	"Fake-bipolar"
PTPOR	"Fake-bipolar"	Unipolar
SESWE	"Fake-bipolar"	"Fake-bipolar"
SISLV	"Fake-bipolar"	"Fake-bipolar"