ESS ERIC DELIVERABLE

ESS ERIC WORK PROGRAMME 02 DECEMBER 2013 – 31 MAY 2015

DELIVERABLE NUMBER: 7.8

DELIVERALBE TITLE: A detailed report outlining the results of the Round 6 MTMM experiments for all countries.

WORK PACKAGE Number: 7

SUBMITTED BY: Universitat Pompeu Fabra (UPF)

AUTHOR(S): Anna de Castellarnau, Melanie Revilla, Willem E. Saris, and Henrik Dobewall

DISSEMINATION STATUS: open

SUBMITTED: 11/03/2015

ACCEPTED: 20 May 2015; amended version resubmitted 30 September 2015

Results of the ESS Round 6 Split-Ballot MTMM experiments for all countries

Anna de Castellarnau
Melanie Revilla
Willem Saris
Henrik Dobewall

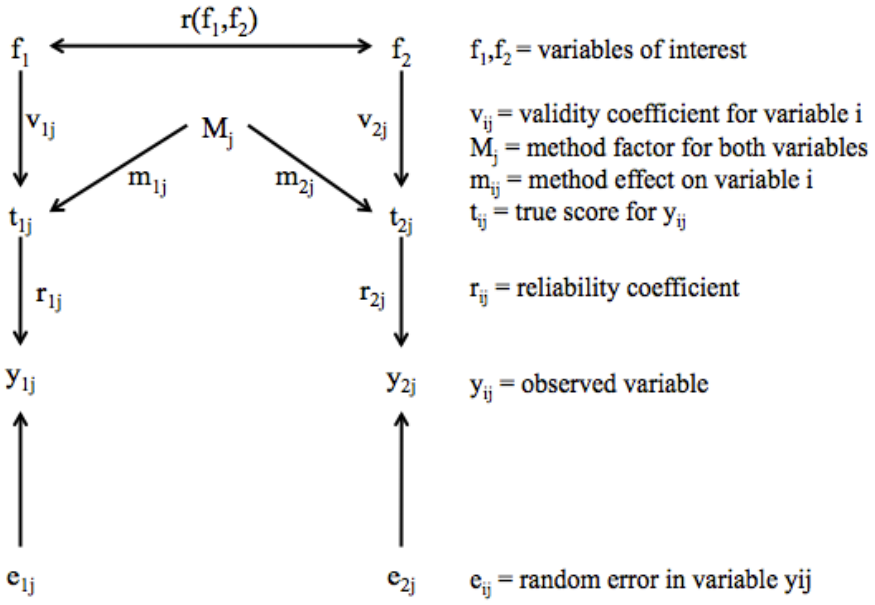
Universitat Pompeu Fabra
Barcelona

In the sixth Round of the European Social Survey (ESS) four Split-Ballot Multitrait-Multimethod (SB-MTMM) experiments have been done to evaluate the quality of survey questions. In this report we will, first, define the quality criteria used and explain how we could estimate it. Then, we will describe the different experiments analysed. Finally, we will report the results of these experiments and discuss the differences in measurement quality of the responses for the different countries.

The quality criteria

Figure 1 presents the basic response model used as starting point to evaluate the quality of a survey question. This is the true score model as proposed by Saris and Andrews (1991).

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



The difference between the observed response (y_{ij}) and the so-called "true score" (t_{ij}) corresponds to random measurement error (e_{ij}). The coefficient r_{ij} represents the reliability coefficient and r_{ij}^2 is the reliability, i.e. the strength of the relationship between the true score and the observed variable.

The true score is separated from the variable of interest (f_i) because it is affected by the method (M_j) used to measure this variable of interest. The coefficient v_{ij} represents the validity coefficient and v_{ij}^2 is the validity, i.e. the strength of the relationship between

the variable of interest and the true score.

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: $q_{ij}^2 = r_{ij}^2 \cdot v_{ij.}$ We call q_{ij} the quality coefficient.

The correlation between the latent variables of interest (also called "factors") is denoted by $\rho(f_1,f_2)$.

Using the decomposition rule, we can express the correlation $r_{(y_{1j,}\ y_{2j})}$ between two observed variables y_{1j} and y_{2j} as a function of reliability, validity, method effect coefficients, and the correlation between the latent factors:

$$r_{(y_{1j,}\ y_{2j})} = r_{1j} \cdot v_{1j} \cdot \rho(f_1,f_2) \cdot r_{2j} \cdot v_{2j} + r_{1j} \cdot m_{1j} \cdot m_{2j} \cdot r_{2j} \quad (1)$$

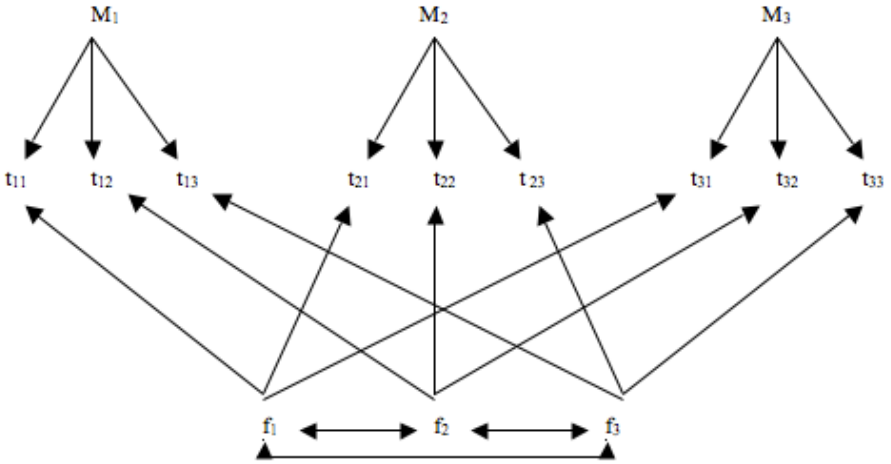
Thus, the observed correlation will only be equal to the correlation between the latent variables of interest (i.e. the correlation without measurement error), when reliability and validity are 1 (i.e. when random errors and method effects are 0), which is very unlikely. Besides, one cannot compare correlations or standardized estimates across countries without correction for measurement errors if the quality coefficients differ across countries. For more details on correction for measurement error, we refer to Saris and Gallhofer (2014).

In this report, we concentrate on the variation in measurement quality across different types of questions and across countries as far as this can be studied on the basis of the SB-MTMM experiments of the ESS Round 6.

The estimation procedure

The model presented in Figure 1 is not identified. Therefore, in order to be able to estimate reliability and validity coefficients, it is necessary to repeat several questions (called traits) using several methods (for instance, several scales: 2-point scale, 6-point scale, 11-point scale, etc.). This is the MTMM approach (Campbell and Fiske, 1959). Figure 2 illustrates the MTMM model for 3 traits each measured with 3 methods.

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



To avoid memory effects, 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 Split-Ballot MTMM approach was implemented in the ESS. It allows asking only two times the same respondent the same questions but still all reliability and validity coefficients can be estimated. It is possible to split the respondents in different numbers of groups. Table 1 presents an example of 3-group SB-MTMM design: in that case, respondents are randomly assigned to 3 different groups. Group 1 gets method 1 in the main questionnaire and method 2 in the supplementary questionnaire, group 2 gets respectively methods 2 and 3, and group 3 respectively methods 3 and 1.

Table 1: The 3-group SB-MTMM design

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

However, in this design, the respondents get different methods in the main questionnaire. To avoid this, a 2-group SB-MTMM design has been used in the ESS. Table 2 presents the current 2-group SB-MTMM design used so far in the ESS.

Table 2: The 2-group SB-MTMM design

	Main Q.	**Suppl. Q.**
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. However, the 2-group SB-MTMM design has a major disadvantage: the information between methods 2 and 3 is missing. This lack of information leads to problems in the estimation of the quality, as reported by Revilla and Saris (2011, 2013) and Saris, Satorra and Coenders (2004).

In ESS Round 6, the sample was divided in four groups in order to increase the number of variations of the formulations. The detailed design will be described hereafter.

The ESS Round 6 experiments

Since the beginning of the ESS in 2002, each ESS Round contained four to six SB-MTMM experiments to evaluate the quality of survey questions. In Round 6, the following four experiments were done:

- Attitudes towards immigration (“Immigration”)
- Engagement during everyday life (“Engagement”)
- Feelings about past week (“Feelings”)
- Evaluation of democracy (“Democracy”)

Each experiment contains three traits measured with several methods (three or five depending on the experiments). Table 3 presents the traits used in each experiment.

Table 3: ESS Round 6 SB-MTMM traits per experiment

Experiment	ID	Trait	Wording of the questions
Immigration	B32	Economy	- Bad or good for [country]'s economy that people come to live here from other countries
	B33	Culture	- [Country]'s cultural life is undermined or enriched by people coming to live here from other countries
	B34	Place	- [Country]'s made a worse or a better place to live by people coming to live here from other countries
Engagement			How much of the time would you generally say you are...
	D31	Interested	- ...interested in what you are doing
	D32	Absorbed	- ...absorbed in what you are doing?
	D33	Enthusiastic	- …enthusiastic about what you are doing?
Feelings			How much of the time during the past week…
	D5	Depressed	- …you felt depressed
	D7	Sleep	- …your sleep was restless
	D9	Lonely	- …you felt lonely
Democracy	E420	Opposition	- Opposition parties in [country] are free to criticise the government
	E215	Media	- The media in [country] are free to criticise the government
	E226	Information	- The media in [country] provide citizens with reliable information to judge the government

The "Feelings" and "Engagement" experiments were measured using a three traits per five methods design. The other two experiments, "Immigration" and "Democracy", were measured using the classic three traits per three methods design. Table 4 gives more information about the different methods analysed.

Table 4: ESS Round 6 SB-MTMM methods per experiment

	Method 1	Method 2	Method 3	Method 4	Method 5
Experiment	**Main Q.**	**Supplementary Questionnaire**			
	(all)	(SB-group 1)	(SB-group 2)	(SB-group 3)	(SB-group 4)
Immigration	11-point IS scale	7-point IS scale Intro	5-point IS scale Intro		
Engagement	11-point FR scale Bat	11-point IS scale Intro	7-point IS scale Intro	5-point IS scale Intro	3-point IS scale Intro
Feelings	4-point FR scale Bat	4-point FR scale Bat	4-point IS scale	10-point IS scale	6-point IS scale
Democracy	11-point IS scale Bat	11-point FR scale	11-point IS scale		

Note: IS: Item-Specific; FR: Frequency; Bat: Question in a battery of questions; Intro: Question with

introduction

The experiments presented in Table 4 allow us to compare the different number of points for item-specific (IS) scales and compare the quality of IS scales and frequency (FR) scales. 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). An example from the Depression experiment of an IS scale would be: *"To what extent did you feel depressed during the past week? – Not at all depressed – Extremely depressed"*. This means that if the question would instead had been asked about happiness the scale would go from "Not at all happy" to "Extremely happy". Besides, a FR scale is used to measure the number of times that an event occurs within a given period. The same example for a FR scale would be: *"How much of the time during the past week you felt depressed? – None or almost none of the time, some of the time, most of the time or all or almost all of the time"*.

Furthermore, in Table 4 it is illustrated that in the ESS Round 6 in order to cope with the five methods design, which increases the variation in the formulations (i.e. having more methods), the sample has been randomly split into four split-ballot groups (SB-groups). For the three method experiments, Immigration and Democracy, the questions were provided to two of the SB-groups, while the five method experiments, Feelings and Engagement, were provided to the four SB-groups.

Data and Methodology

In Round 6, the SB-MTMM experiments were conducted in the 29 participating countries. Because the language can affect the reliability and validity (Saris and Gallhofer, 2007 and Zavala-Rojas, 2015), the data was not only split by country but also by language in multilingual countries. Table 5 summarizes the countries and languages.

Table 5: ESS Round 6 SB-MTMM countries and languages available

Country	Language 1	Language 2	Language 3
Albania	Albanian [ALALB]		
Belgium	Dutch [BEDUT]	French [BEFRE]	
Bulgaria	Bulgarian [BGBUL]		
Switzerland	German [CHGER]	French*	Italian*
Cyprus	Greek [CYGRE]		
Czech Republic	Czech [CZCZE]		
Germany	German [ALALB]		
Denmark	Danish [DKDAN]		
Estonia	Estonian [EEEST]	Russian [EERUS]	
Spain	Spanish [ESSPA]	Catalan*	
Finland	Finnish [FIFIN]	Swedish*	
France	French [FRFRE]		
Great Britain	English [GBENG]		
Hungary	Hungarian [HUHUN]		
Ireland	England [IEENG]		
Israel	Hebrew [ILHEB]	Arabic*	Russian*
Iceland	Icelandic [ISICE]		
Italy	Italian [ITITA]		

Lithuania	Lithuanian [LTLIT]	Russian*	
Netherlands	Dutch [NLDUT]		
Norway	Norwegian [NONOR]		
Poland	Polish [PLPOL]		
Portugal	Portuguese [PTPOR]		
Russian Federation	Russian [RURUS]		
Sweden	Swedish [SESWE]		
Slovenia	Slovene [SISLV]		
Slovakia	Slovak [SKSLO]	Hungarian*	
Ukraine	Ukrainian [UAURK]	Russian [UARUS]	
Kosovo	Albanian [XKALB]	Serbian*	

*In brackets are the short names used for the country-language combinations for the rest of the report. The first two letters belong to the country ISO code and the last three letters belong to the corresponding language ISO code.

These cases with an asterisk (*) were not analysed because the sample size was too small (<150 cases per split-ballot group). Thus, taking into account the significant country-language combinations, we could analyse each of the experiments in the 32 country-language combinations presented in Table 5.

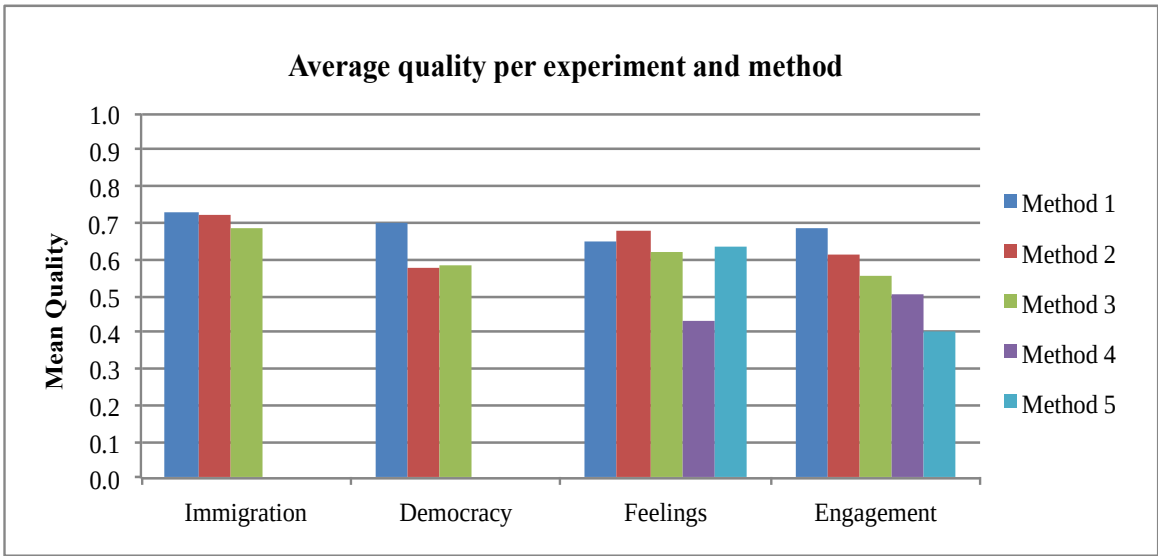
For each experiment and country, the estimates are obtained from LISREL by Maximum Likelihood (ML) estimation for multi-group analysis. In order to test if there are misspecifications, we use the JRule software (Van der Veld, Saris, Satorra, 2009) based on the procedure developed by Saris, Satorra and 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 do not test the model as a whole). This leads in many cases to the introduction of corrections with respect to the general model presented earlier (Figure 2). Principally, the changes consist in 1) adding a correlation between two methods when they are very similar; or 2) allowing unequal effects of one method on the different traits. Sometimes to solve cases of improper solutions (i.e. negative variances) or non-convergence we fix one of the method effects (respectively error variance) to zero if this method variance (respectively error variance) is not significantly different from zero. In order to be able to compare results across countries, we try to make the same corrections in all countries for one specific experiment. However, this is not always possible and sometimes we have to allow differences across countries.

Results

In this section, the results will be presented, first, looking at the overall picture and afterwards, focusing on each experiment and analysing the impact of the different methods. It has to be taken into account, that these results cannot be generalized nor extrapolated. The limits of these analyses and further research are presented in the next section.

Figure 3 presents the average quality in each of the four experiments for each of the methods used, i.e. mean of the quality of all the traits in all the countries.

Figure 3: Average quality of the questions in the different experiments by the methods used



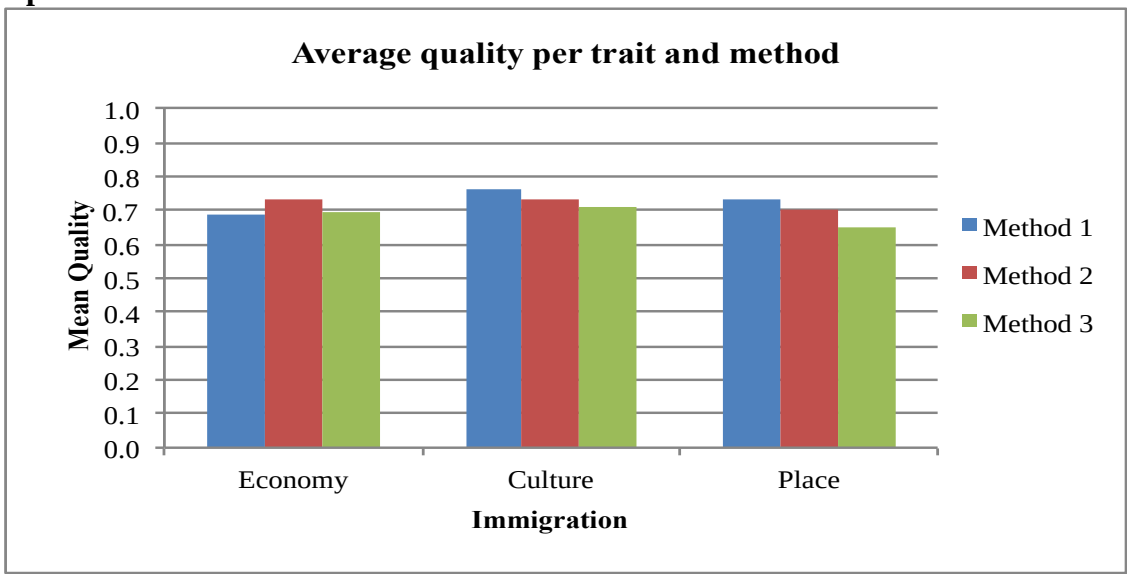
In general, the average quality among experiments is between 0.5 and 0.7, except for the experiments "Feelings" and "Engagement", which have qualities below 0.5 in Methods 4 and 5. For the "Feelings" experiment, the quality is especially low for Method 4, which corresponds to an IS scale with 10 response categories, thus, with no middle or neutral point. For the "Engagement" experiment, the quality seems to go down when the number of answer categories decreases.

However, Figure 3 gives only aggregated results. In order to study the differences across countries and across traits, results are presented for each experiment for the different traits and for the different countries, in the next section.

I. Immigration experiment

The Immigration experiment, allows studying the effect of the number of answer categories on the quality for IS scales. This can be done by comparing three IS scales with 11 points (Method 1), 7 points (Method 2) and 5 points (Method 3). The results per trait and method are presented in Figure 4.

Figure 4: Average quality of the questions per trait and method in the Immigration experiment



In Figure 4 we can see that the average quality decreases with the number of points for the traits "Culture" and "Place". For the trait "Economy", the 7-point scale (Method 2) has a higher average quality than the other two. Although the differences are small (0.69 for Method 1, 0.73 for Method 2 and 0.70 for Method 3), and it seems that Method 1 has a particularly lower quality for this trait. Thus, overall, the results suggest that there is a tendency that the quality is lower for shorter IS scales.

It is also interesting to compare countries. Indeed, standardized relationships across countries can only be compared if the quality is similar across countries. A comparison of the mean quality of the questions across methods for the different countries is presented in Figure 5.

Figure 5: Average quality of the questions per country and method in the Immigration experiment

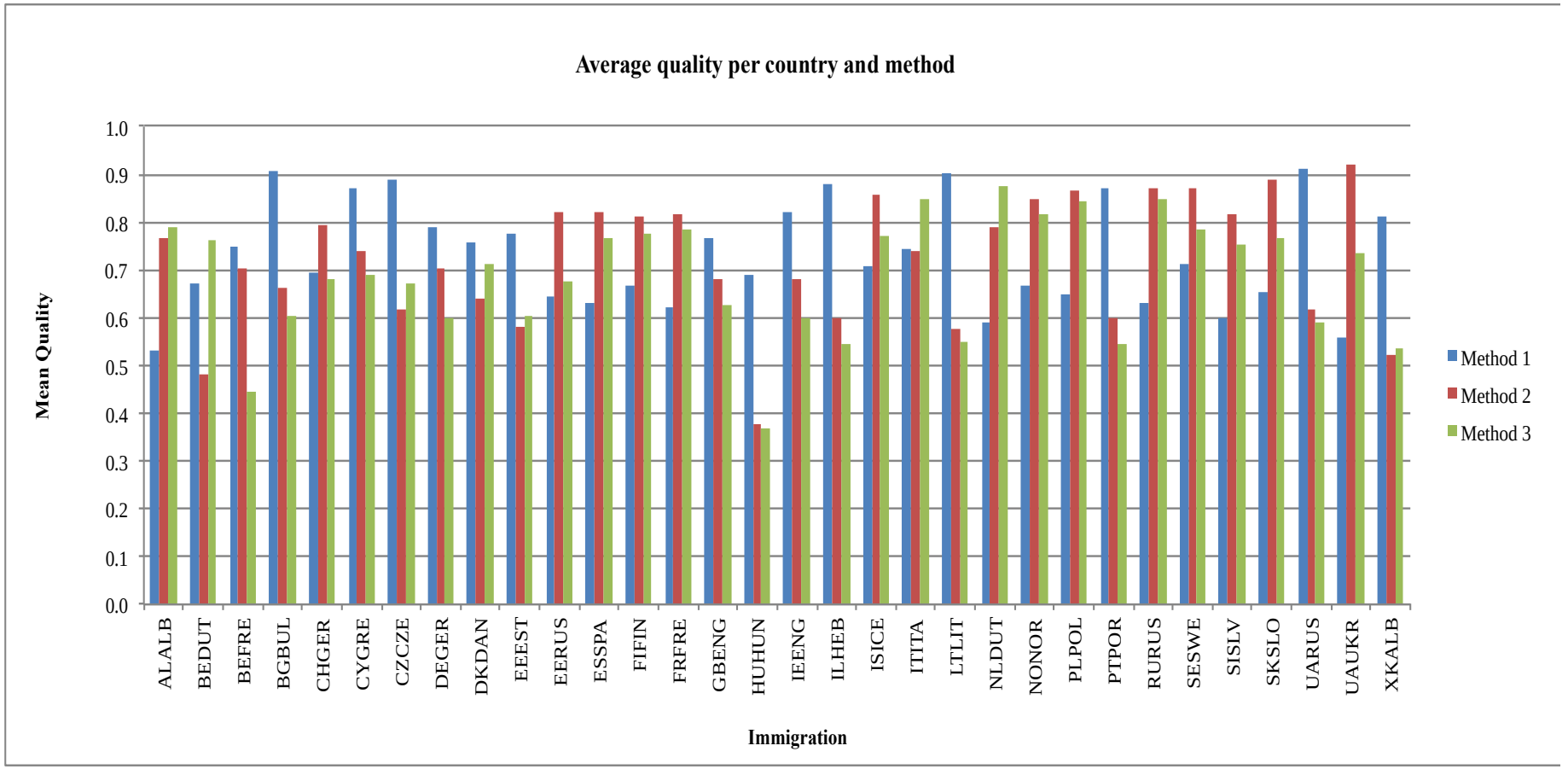
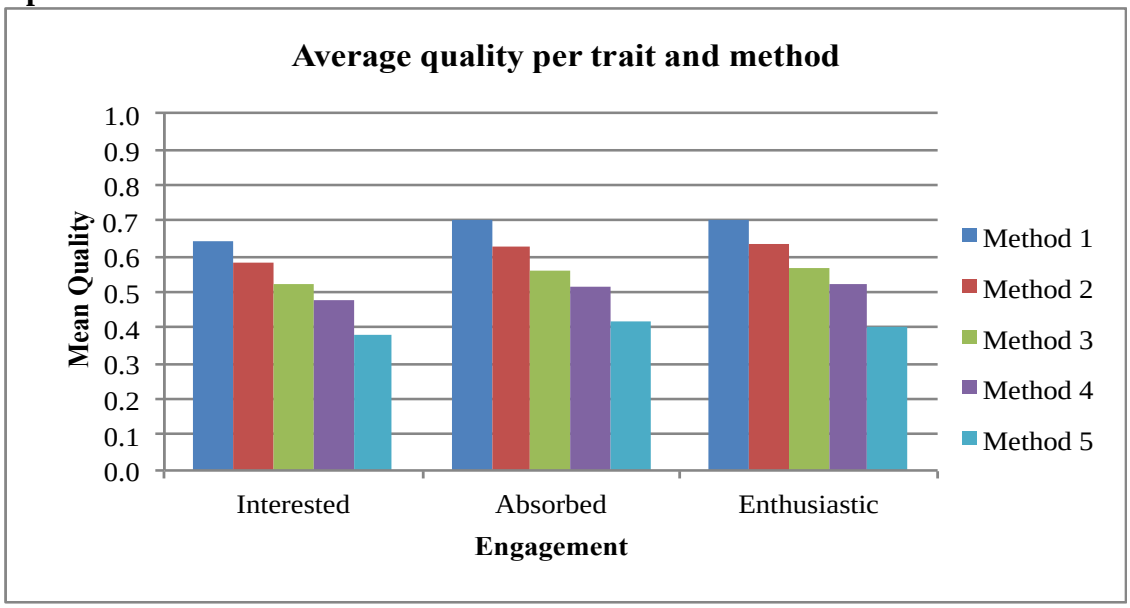


Figure 5 shows that there are differences in measurement quality across the countries. While for most of them, the methods with higher quality are Methods 1 or 2, there are some countries like Albania, Belgium (Dutch), Italy and the Netherlands for which Method 3 has higher quality. Therefore, it is necessary to correct for measurement errors before comparing standardized relationships across countries for these different traits. We can also notice that the highest quality estimates are found for Iceland, Italy and Sweden, which all have a mean quality above 0.7. On the other hand, we see that the questions in Hungary have the lowest quality.

II. Engagement experiment

Similar to the Immigration experiment is the purpose of the Engagement experiment, which allows observing the effect of the number of points in IS scales, comparing 11 points (Method 2), 7 points (Method 3), 5 points (Method 4) and 3 points (Method 5). Besides, this experiment allows also to compare 11-point frequency (FR) scales (Method 1) with 11-point IS scales (Method 2). The quality results are presented in Figure 6 per trait and method.

Figure 6: Average quality of the questions per method and trait in the Engagement experiment

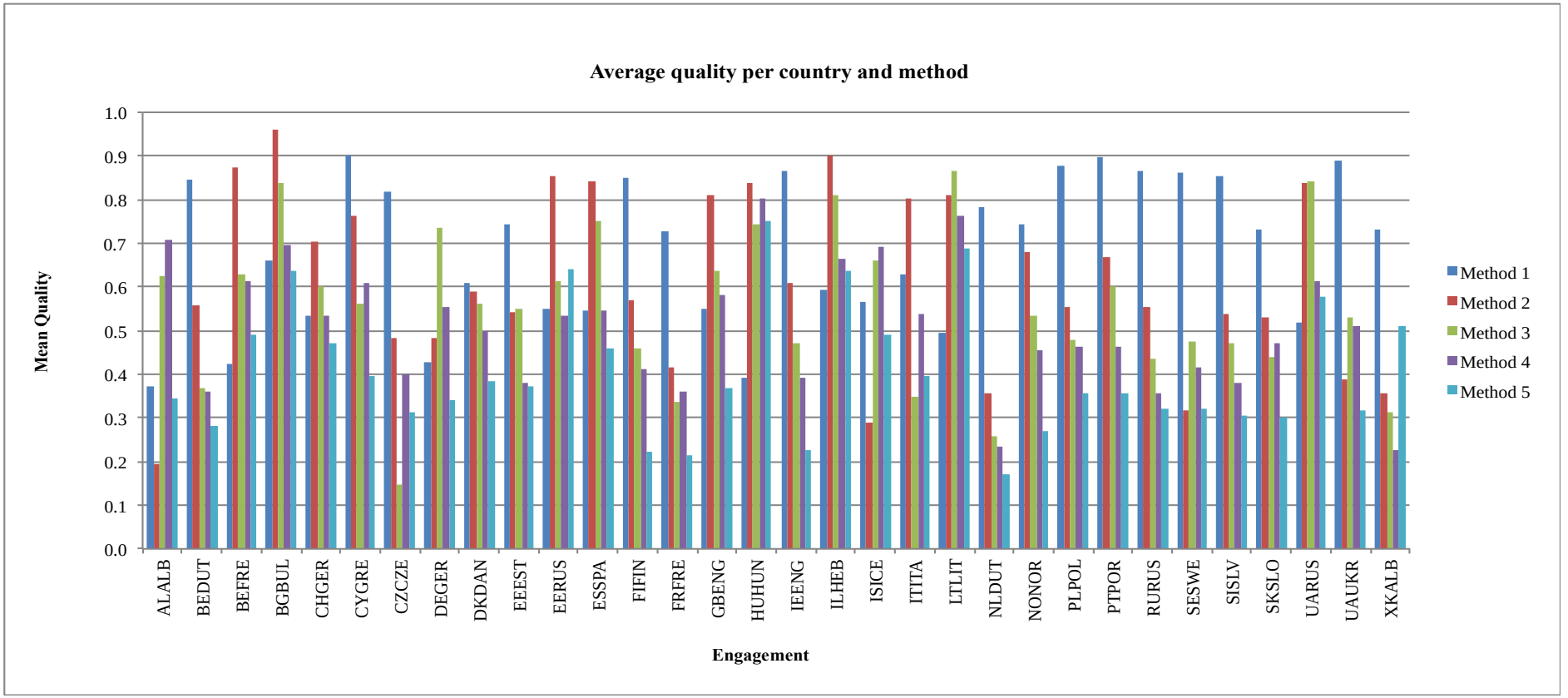


In line with the Immigration experiment, we can see in Figure 6 that the quality of the IS scales decreases with the number of answer categories. Across methods, the differences are large: the quality ranges from 0.38 to 0.70.

Furthermore, being Method 1 an 11-point FR scale, it can be compared to Method 2, an 11-point IS scale. The results show that frequency (FR) scales have higher quality. However, one should note that this result is only true for this experiment and cannot serve for generalization.

Next, the results per country and method are presented in Figure 7.

Figure 7: Average quality of the questions per country and method in the Engagement experiment



The information presented in Figure 7 is useful because countries can only be compared if the quality is similar across them. We can see from Figure 7 that Israel (Hebrew) is the country with higher quality overall and the lowest is Denmark.

Moreover, there are also deviations in terms of the general conclusions: 1) the quality increases with the number of points in IS scales, and 2) FR scales have higher quality than IS scales. In Albania, Germany, Iceland and Ukraine (Ukrainian) the 11-point IS scale has not the highest quality among the IS scales. Besides, in Belgium (French), Bulgaria, Switzerland (German), Germany, Estonia (Russian), Spain (Spanish), Great Britain, Hungary, Israel (Hebrew), Italy, Lithuania (Lithuanian) and Ukraine (Russian) IS scales (Method 2) have higher quality than FR scales (Method 1).

Taking into account the country deviations presented in Figure 7 by correcting for measurement errors these countries can be compared.

III. Feelings experiment

Similarly, the Feelings experiment also allows comparing the qualities between 4-point FR scales (Methods 1 and 2) and 4-point IS scales (Method 3). Moreover, using scales without explicit or implicit neutral point we can observe the impact of the number of response categories with 10-point (Method 4), 6-point (Method 5) and 4-point (Method 3) IS scales. The results per trait and method are presented in Figure 8.

Figure 8: Average quality of the questions per method and trait in the Feelings experiment

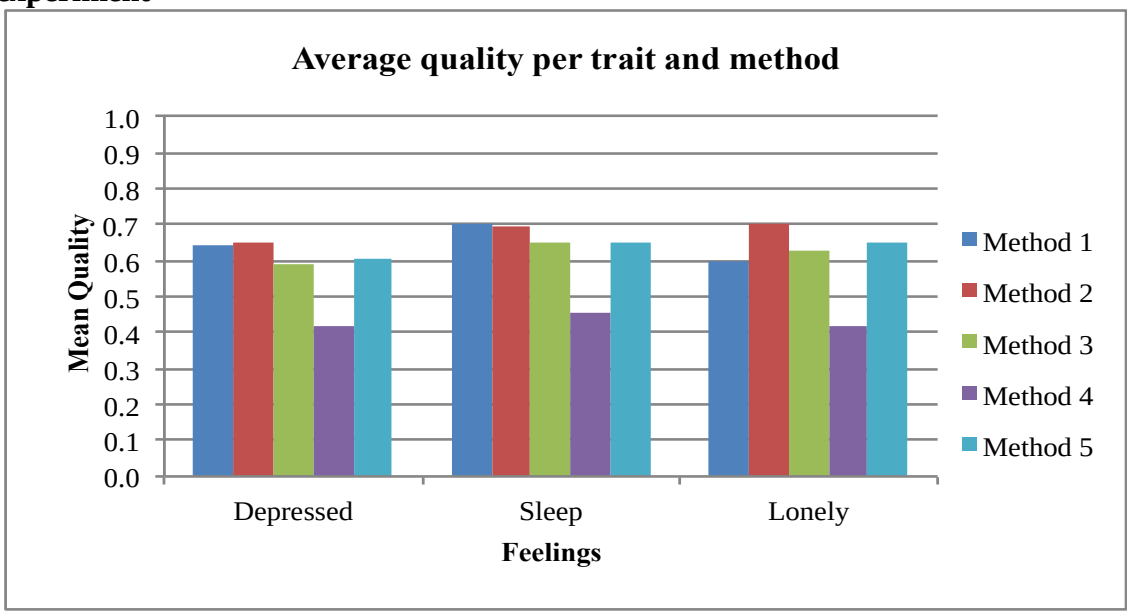


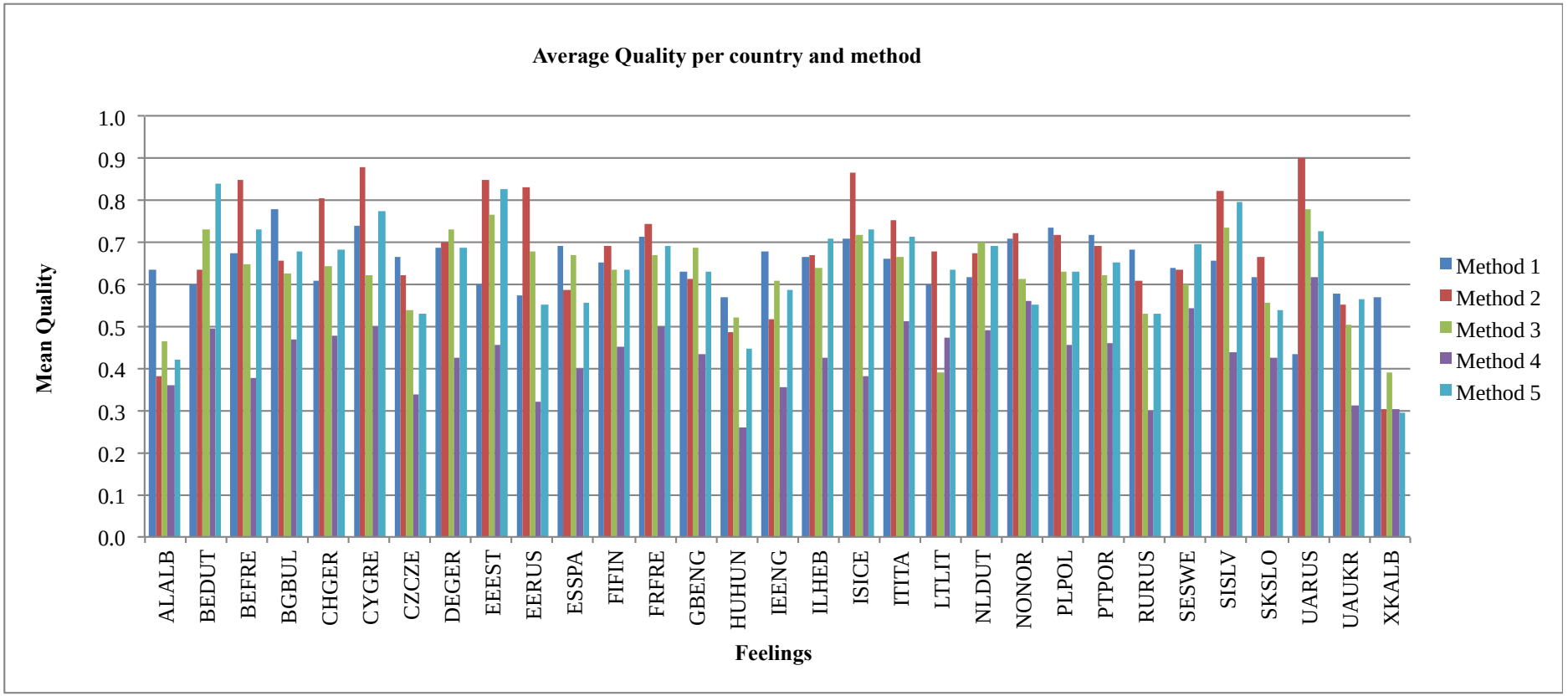
Figure 8 shows that, overall, the differences in the quality across methods is large, ranging from 0.41 to 0.7. This is mainly due to the fact that the average quality of Method 4 is very low, 0.43.

Thus, comparing the qualities of FR and IS scales we can conclude that Method 1 and 2 (both 4-point FR scales) perform slightly better than Method 3, a 4-point IS scale. This is in line with the finding from the "Engagement" experiment.

Figure 9 shows that the pattern observed in the previous experiments, i.e. “Immigration” and “Engagement”, does not apply for scales with an even number of response categories, i.e. without an implicit or explicit middle or neutral point. In this case, Method 5 (6-point IS scale) and Method 3 (4-point IS scale) perform better than Method 4 (10-point IS scale). Besides there is the possibility that larger IS scales with an even number of categories perform worse than shorter IS scales. Another possible reason why Method 4 has such a low quality could be due to design-dependency of the measurement quality estimates (Költringer, 1995). This means, larger scales (such as the 10-point scale), compared with the other shorter scales (i.e. 4-point and 6-point scales), could have a low quality because of design effects of this particular experiment.

In Figure 9, these same results are presented per country and method, in order to detect the countries that are more similar in terms of quality to be able to compare them.

Figure 9: Average quality of the questions per country and method in the Feelings experiment



*Note: Denmark was excluded from this experiment because they used an 11-point scale in Method 4 instead of a 10-point scale.

From Figure 9 it can be highlighted that Ukraine (Russian) is the country with higher quality, with a measurement quality for all methods higher than 0.6, and the countries with lower qualities are Ukraine (Ukrainian), Kosovo (Albanian) and again Hungary.

Furthermore, Figure 9 also shows the deviant countries in terms of the general conclusion (i.e. Method 1 and 2 perform better than Method 3). The countries for which Method 3 has a higher quality are Belgium (Dutch), Germany, Great Britain and the Netherlands.

IV. Democracy experiment

The Democracy experiment focuses on the comparison of IS scales (Methods 1 and 3) and frequency (FR) scales (Method 2), using 11 points scales. The results per trait and method are presented in Figure 10.

Figure 10: Average quality of the questions per method and trait in the Democracy experiment

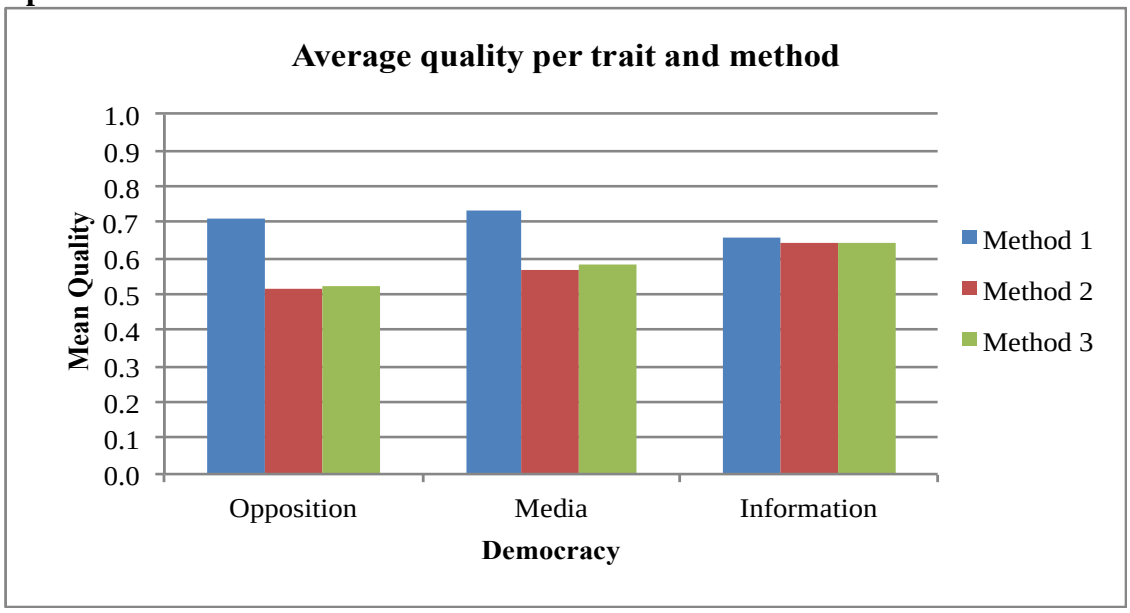
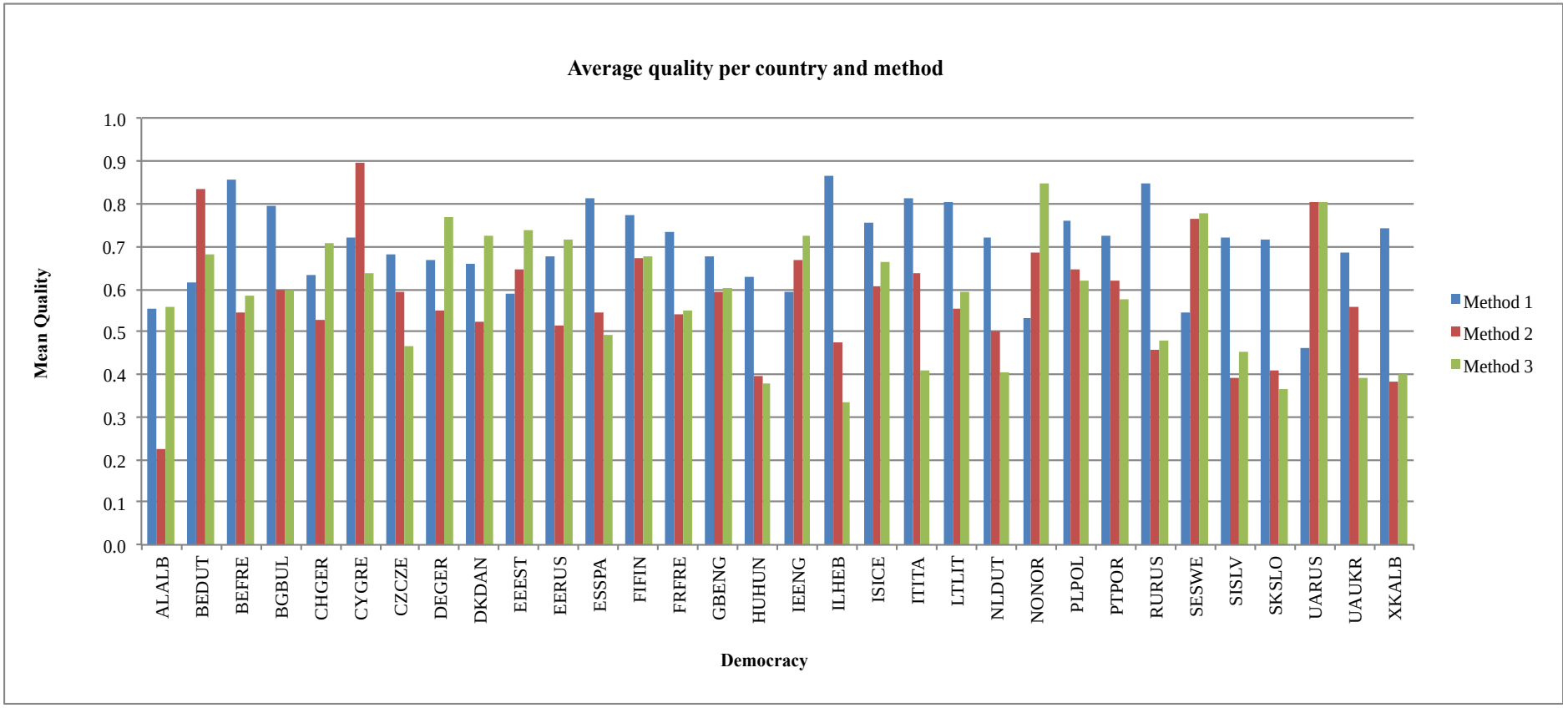


Figure 10 shows that the IS scales (Methods 1 and 3) have a higher quality than the FR scales (Method 2). This result is different from what has been found in the "Engagement" and "Feelings" experiments. This indicates that there is an interaction between the topic of the questions and the scale characteristics. Moreover, from this figure we can see that the differences over the quality are pronounced, as the average quality across methods ranges from 0.52 to 0.73.

In order to see how these effects are spread across countries, Figure 11 shows the differences on the average quality of each county and per method.

Figure 11: Average quality of the questions per country and method in the Democracy experiment



Countries with higher quality are Belgium (Dutch), Cyprus, Finland (Finish), Iceland and Poland which all have a mean quality above 0.6. On the other hand, we see that the questions in Albania and again Hungary have the lowest quality.

Figure 11 also allows detecting in which countries the quality of Method 2 (11-point FR scale) is higher than the other two. Method 2 performs better than Methods 1 and 3 in Switzerland (German), Germany, Denmark, Estonia (Estonian and Russian), Ireland and Norway.

This information is useful because countries can only be compared if the quality is similar across them or if these are corrected for 0measurement errors.

Conclusions

The four experiments analysed intended to measure the quality of different formulations of the same questions. The formulations chosen in the ESS Round 6 experiments allow, on the one hand, to observe the impact on the quality of the number of points in IS scales and, on the other hand, to compare frequency scales with IS scales.

First, from the two experiments, "Immigration" and "Engagement", designed with the purpose of observing the impact of the number of response categories on the quality of questions including IS response scale, comparing response categories with an uneven number of point (i.e. 11 points, 7 points, 5 points and 3 points), it can be concluded that for two topics the quality decreases with the number of response categories. More specifically, that 11-point IS scales perform better than 7 points, 5 points and 3 points, in the same way that 7-point IS scales perform better than 5 points, and 5-point IS scales perform better than 3 points.

Moreover, from the "Feelings" experiment we have seen that this result cannot be extrapolated to all topics or all number of points. It has been shown that for this specific topic and for scales with an even number of categories (i.e. without implicit or explicit middle or neutral point) the pattern does not hold. In this case, the 6-point IS scale has higher quality than the 4-point and 10–point IS scales, and the 4-point IS scale performs better than 10-point IS scale. The two possible explanations for this result could be that: 1) the effect of not having either an implicit or explicit middle point is much negative in larger than shorter scales; and that 2) design effects of comparing the number of categorical variables versus continuous variables have an effect on the results of such estimates.

Second, from the "Engagement" and "Democracy" experiment, which were both designed to compare 11-point FR scales with 11-point IS scales, we can conclude that which scale performs better depends on the topic. For the "Engagement" experiment the 11-point FR scale has slightly higher quality than the 11-point IS scale, while for the "Democracy" experiment the 11-point IS scale (from the Main Questionnaire) performs better than the 11-point FR scale (from the Supplementary Questionnaire). However, in the "Democracy" experiment an 11-point IS scale was also provided in the Supplementary Questionnaire, and the quality is very similar to the 11-point FR scale.

Similarly, the "Feelings" experiment also provided information about the quality of 4-point FR scale and 4-point IS scales, which results are in line with the Engagement

experiment. The 4-point FR scale has higher quality than the 4-point IS scale.

Besides, it has been illustrated that over all experiments the quality is quite low, being the maximum 0.73, and that there are large deviations not only across countries but also within countries for different languages. This shows how important it is to have these quality estimates in a cross-national survey that allow correcting for the quality and estimate the true correlations, i.e. the correlations corrected for measurement error, and, therefore, comparing the standardized relationships.

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

Limitations of the SB-MTMM approach and future research

Unfortunately the results of these experiments cannot be used to draw general conclusions for questionnaire designers about which type of scales or which number of categories should be used.

As presented in the first sections of this report, the quality estimates are obtained in the ESS trough SB-MTMM experiments. As most other approaches, the SB-MTMM approach has also some limitations. The first is that the results of the analyses cannot be generalized nor extrapolated out of the ESS context, nor of the ESS round or even the experiments themselves (e.g. it cannot be conclude that FR scales are, in general, better than IS scales, and that from now on FR scales would be preferred over IS scales). This means, that the quality information and, therefore, correction for measurement error, would be limited to those questions involved in SB-MTMM experiments.

The second limitation, which is closely related to the first, is that because the SB-MTMM design requires asking twice the same questions to the respondents, only a limited set of questions can be implemented in the questionnaire. With the current ESS questionnaire length and in the context of a face-to-face (CAPI or PAPI) interview, not all survey questions can be repeated to the respondents, as it would at least double the length of the questionnaire, its cost and it would also increase the cognitive burden for the respondents.

Thus, the aim is to provide the ESS and other researchers a tool to obtain the quality of new survey questions to be able, on the one hand, to generalize and extrapolate the results of the SB-MTMM experiments and, on the other hand, to correct for measurement errors and compare cross-country standardized relationships. With this purpose the results obtained from the ESS Round 6 SB-MTMM experiments, presented in Table 4, will be used to increase the Survey Quality Predictor (SQP) database of quality estimates and survey questions characteristics and to enrich its quality prediction meta-analysis.

Currently, SQP 2.0 is based on a meta-analysis with 3,726 MTMM questions, its quality estimates and its formal characteristics. These MTMM experiments were based on different formulations of survey questions used in ongoing survey research in the United States, the Netherlands, Belgium and Austria and more recently in the different

Rounds of the ESS. The alternatives for the questions in these studies were chosen in order to represent very common alternative formulations at that time. With these alternative formulations we have been able to detect a lot of factors that determine difference in quality of questions. For example in earlier research we detected that the number categories in agree-disagree scales have a negative effect on the quality (Revilla, Saris and Krosnick, 2013). In this report we have illustrated that in case of IS scales, at least for some topics, the number of categories has a positive effect on the quality. However in this report we also have mentioned that we have detected again that results cannot simply be generalized. Sometimes the topic of the questions changes the results quite a bit.

Besides that, in the recent years, there has been an increasing development in survey research towards new formulations and modes of data collection, that at the time SQP was developed these were only partially covered by the MTMM experiments. All these new developments require further research using MTMM experiments. Because the SQP quality prediction is as good as its meta-analysis and its meta-analysis is as good as its data, replication of different and new combinations of traits and methods using MTMM experiments is needed.

References

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

De Castellarnau, A. and Saris, W. E. (2014). A simple procedure to correct for measurement errors in survey research. *European Social Survey Education Net (ESS EduNet). Available:* http://essedunet.nsd.uib.no/cms/topics/measurement/

Kol̈tringer R. (1995). Categorization and measurement quality. *In W. E. Saris, and A. Muñnich (eds.), The Multitrait-Multimethod Approach to evaluate measurement instruments, Budapest: Eoïvos̈ University Press, 207─225.*

Revilla, M. and Saris, W. 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, M., Saris, W.E., and Krosnick, J.A. (2013). Choosing the number of categories in agree-disagree scales. *Sociological Methods and Research February 2014 43: 73-97, first published online on December 10, 2013, DOI: 10.1177/0049124113509605*

Revilla, M. and Saris, W. E. (2011). The split-ballot multitrait-multimethod approach: The importance of the third group. *Presented at the Meeting of the Working Group Structural Equation Modeling (Marburg, Germany, 25-26 March 2011)*

Saris, W.E., A. Satorra and W. van der Veld (2009), Testing Structural Equation Models or Detection of Misspecifications?, *Structural Equation Modeling, 16 pp. 561-582*

Saris, W. E. and Gallhofer, I. N. (2014). Design, evaluation and analysis of questionnaires for survey research. Second Edition. *Hoboken, Wiley*.

Saris, W. E. et al. (2010). Comparing questions with Agree/Disagree response options to questions with Item Specific response options. *Survey Research Methods Vol. 4, No. 1, pp. 61-79.*

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

Saris, W. E, Satorra, A. and Coenders, G. (2004). A new approach for evaluating the quality of measurement instruments: Split Ballot MTMM design. *Sociological Methodology, 34, 331-347*.

Saris, W. E. and Andrews, F. M. (1991). Evaluation of measurement instruments using a Structural Modeling Approach. *Pp. 575 – 99 in Measurement errors in surveys, edited by Biemer, P. P. et al. New York: Wiley*.

Zavala-Rojas, D. (2015). Cross cultural or cross national research? The role of language in a comparative survey. *RECSM Working Paper (forthcoming).*