European
Social
Survey

ESS-DACE

The European Social Survey - Data for a Changing Europe

Contract Number: 262208

Deliverable 4.9
MTMM Pilot Analysis Report (R6)

Start date of project: July 2010 Duration: 48 months

Organisation name of lead contractor for this deliverable: UPF

Dissemination level: PU (Public)

Due date of deliverable: June 2012

Actual submission date: August 2012

Analysis of ESS Pilot Round 6

Willem Saris
Paolo Moncagatta
Diana Zavala
RECSM – Universitat Pompeu Fabra

1. Analysis of measurement invariance

1.1 Equivalence test of items of well-being module

On request of the QDT we have tested several items on equivalence across countries. Equivalence of items is necessary for comparison of results of questions. If questions are interpreted or answered in different ways in different countries, the results of these questions cannot be compared. Equivalence can be tested by studying the similarity of the relationships between the variable one wants to measure and the observed variables. The variable one wants to measure is a latent variable, for example “depression”. This concept is measured in the module of round 6 by several indicators. We will use the items B6, B7 and B12 because the QDT wanted to know if these items were equivalent across the countries involved in the pilot study of the ESS. So we can specify a factor model with three indicators:

$B6 = a_6 + b_{61}F_1 + u_6$
$B7 = a_7 + b_{71}F_1 + u_7$
$B12 = a_{12} + b_{12,1}F_1 + u_{12}$

Where F_1 is the latent variable “depression”, a_i is the intercept, b_{ij} is the slope and u_i is the disturbance term in the equation.

If we assume that the errors are independent of each other and of the latent variable F_1 the intercepts and slopes of this one factor model can be estimated. This model should hold for both countries but in order to have equivalent measures in the different countries the slopes and intercepts should also be the same (Meredith 1993, Saris Gallhofer 2007). If only the slopes are invariant across countries (metric invariance) relationships with other variables can be compared a cross countries. If also the slopes are the same (scalar invariance) the means can be compared as well.

These requirements can be tested with multiple-group SEM programs assuming that the coefficients are the same. We have done these tests for several items for which sufficient information was available because one need at least two indicators per concept in order to be able to do the tests. Below we present the results of our analyses.

1.2 Equivalence of measures for feeling “depression” and “vitality”

The measurement model for the “depression” was indicated above. The measurement model for the “vitality” was similar but there were only two indicators B13 and B15. In case of two items a combined model with another related concept has to be made otherwise the model is not identified.

So we have added to the previous sets of equations
$B13 = a_{13} + b_{13,2}F_2 + u_{13}$
$B15 = a_{15} + b_{15,2}F_2 + u_{15}$

Where F2 is the latent variable “vitality.” The meaning of the other symbols remains the same and the requirements as well. Besides, we expect that F_1 and F_2 are correlated. In this way we have specified a two factor model with 5 observed indicators.

Testing for the invariance of the coefficients in this model across the countries GB and Russia we have detected only minimal differences. The result looks as follows for both countries

$B6 = 0.0 + 1.00 F_1 + u_6$
$B7 = 0.00 + 1.05F_1 + u_7$
$B12 = 0.00 + .97F_1 + u_{12}$
$B13 = 0.0 + 1.00F_2 + u_{13}$
$B15 = 4.84 + -1.10F_2 + u_{15}$

The only difference is that in Russia the intercept in the last equation is 5.21 instead of 4.84. This means that in the last equation the responses for B15 are always .4 higher in Russia than in GB.

In the general the “depression” items can be used for comparison of means and relationships across countries while the “vitality” can, strictly speaking, only be used for comparison of relationships and not for comparison of means across countries.

1.3 Equivalence of measures for the variables “control” and “Involvement in well-being promoting activities”

In the same way as above we specified a two factor model for the concepts “Control” and “Involvement in well-being promoting activities” with the items B22 and B37 as indicators for “Control”. From the last concept we only took the subconcepts “Awareness of internal world” and “take notice” (B43 and B44). Testing this model on the data of GB and Russia we had to conclude that the measures are not equivalent across countries. This can be seen in the difference in coefficients in the two countries.

For GB we obtained the following result:

$B22 = 0.0 + 1.00F_1 + u_{22}$
$B37 = -1.48 + 1.37F_1 + u_{37}$
$B43 = 0.00 + 1.00F_2 + u_{43}$
$B44 = 1.95 + 0.55 F_2 + u_{44}$

For Russia we got the following result:

$B22 = 0.0 + 1.00F_1 + u_{22}$
$B37 = -1.31 + 1.18F_1 + u_{37}$
$B43 = 0.00 + 1.00F_2 + u_{43}$
$B44 = .20 + 1.02 F_2 + u_{44}$

Because the latent variables have no fixed scale their scale is fixed by making their scales equal to the first item for each factor i.e B22 and B43. The choice of these two items is arbitrary. By fixing the intercept on zero and the slope on 1 the zero point of the observed and latent variable are the same (0.0) and by fixing the slope on 1 an increase of 1 point on the latent variable leads to an increase of 1 point on the observed variable.

Given this specification the relationship for the second item for each factor can be compared over countries. We see then that B37 has approximately the same relationship across countries. This suggests that these items are also rather equivalent across countries. However the item B44 has a rather different relationship with the latent variable in GB than in Russia. So in that case there is a problem with the equivalence of one or both item (B43 and B44). It is not clear what the problem is because fixing the scale of the latent variable was arbitrary. So our judgment is only a relative one. So the problem can be in B43 or B44 or in both items.

1.4 Conclusions

These analyses lead to the following conclusions:

1. The items for the feelings “Depression” are equivalent and can be used to compare means and relationships across countries
2. The items for the “vitality” can be used for comparison of relationships across countries and not for comparison of means (strictly speaking).
3. The items for the concept “Control” are also equivalent and can also be used for comparing means and relationships across countries
4. The items for “Awareness” and “take notice” are not equivalent and the means and relationships can be not be compared across countries

It is for us very difficult to indicate what went wrong in this case. Russian people should determine if they can compare their formulations with the British ones.

2. Analysis of SB-MTMM experiments in the Round 6 Pilot study.

The Pilot study of the ESS included two SB-MTMM experiments. One experiment on political trust and a second experiment on some well-being items.

2.1 SB MTMM experiment on political trust

The experiment on political trust was specified using three traits: trust in the government, trust in the legal system and trust in the police and three methods. The sample was divided in two groups. The experiment intended to test the effect of using a bipolar range for the scale phrased as trust-distrust, instead of the normally used “lack of trust” to “complete trust” range. It also analyse the effect of the used of the mid-point “neither trust nor distrust.” We wanted to test this new scale distrust-trust because there is a debate if trust is a unipolar or bipolar concept and because recent protests in Europe and other countries are explicitly speaking about distrust in the political system.

The design of the experiment is summarized in Table 1.

Table 1.Summary of political trust SB-MTM experiment

Formulation	Method 1	Method 2	Method 3
How much do you personally trust each of the institutions: - [Country]’s parliament - The legal system - The police	- 11-point scale - Battery - Labels: not trust at all – complete trust	- 11-point scale - Direct question - Labels: not trust at all – complete trust	- 11-point scale - Direct question - Labels: complete distrust – neither distrust nor trust - complete trust

For estimation, we specified a multi-group model to be invariant for the two countries, UK and Russia. We use JRule (Saris et. al. 2009) to detect misspecifications in the model and we free the parameters that were misspecified in the second country. The results of the SB-MTMM experiment shows that for both countries UK and Russia the first method: a unipolar scale in the form of 'not at all trust' to 'complete trust' has a higher total quality. The trait that has a higher quality is the one asking about trust in police this might be because people have more information about them for giving their opinion. In some sense the police is closer to the people than "parliament" or "legal system." In general there are larger method effects for the third method although all questions have an acceptable quality. The results of the SB-MTMM experiment are summarized in Table 2 below.

Table 2. SB-MTMM experiment on political trust

	United Kingdom				**Russia**			
Traits	**Validity**	**Reliability**	**Quality**	**Method effects**	**Validity**	**Reliability**	**Quality**	**Method effects**
Country's parliament – method 1	0,98	0,79	0,78	0.12	0,98	0,90	0,88	0.12
Legal System – method 1	0,98	0,81	0,79	0.11	0,98	0,90	0,88	0.11
Police- method 1	0,98	0,98	0,96	0.11	0,98	0,96	0,94	0.11
Country's parliament– method 2	0,88	0,77	0,68	0.33	0,92	0,86	0,80	0.29
Legal System – method 2	0,88	0,85	0,75	0.33	0,81	0,98	0,79	0.44
Police- method 2	0,88	0,92	0,81	0.34	0,88	0,90	0,80	0.34
Country's parliamen-– method 3	0,86	0,72	0,62	0.37	0,86	0,88	0,76	0.37
Legal System – method 3	0,86	0,83	0,72	0.38	0,86	0,94	0,81	0.38
Police- method 3	0,85	0,98	0,83	0.39	0,85	0,88	0,75	0.39

2.2 SB MTMM experiment on well-being

The experiment on well-being was specified using four traits and three methods. The sample was divided in two groups. The experiment intended to test the effect of a statement for the request with the labels: "Does not apply at all" to "Applies completely" in comparison with a direct question "To what extent do you feel…" and to a third direct question using "never to always." The traits selected were measures of the concepts and subconcepts respectively: "meaning and purpose – importance", "autonomy and control-control", "competence-feeling competent" and "engagement-interest in learning."

The design of the experiment is summarized in Table 3.

Table 3.Summary of well-being SB-MTM experiment	
Method	Formulation
Method 1 – -11-point scale - statement - labels: does not apply at all – applies completely	Say how much each of the following statements applies to you: - I regularly make time to do the things I really want to do - There are lots of things I feel I am good at. - Most of what I do feels unimportant to me. - My life involves learning new things.
Method 2 – -11-point scale - item-specific scale - direct question - labels: not at all – completely	To what extent do you… - make time to do the things you really want to do? - feel that there are a lot of things you are good at? - feel that what you do is unimportant? - does your life involve learning new things?
Method 3 – -11-point scale -frequencies - direct question - labels: never-always	How often, if at all, do you… - make time to do the things you really want to do? - feel that there are a lot of things you are good at? - feel that what you do is unimportant? - does your life involve learning new things?

The results show that, except for the negative item on unimportance, the second method performs better in terms of quality in both countries. It is better to use a direct question with an item specific scale. This is in line with the findings of Saris et. al.(2010). As there are differences in quality among countries when we use method 3, it is better to use “not at all” and completely” as ending points of the scale. Item 3 showed large method effects in methods 2 and 3 and the lowest reliability. We can argue that this is a reaction to the formulation of the request in a negative way, rather than a reaction to the scale used.

Table 4. SB-MTMM experiment on well being								
	United Kingdom				**Russia**			
Traits	**Validity**	**Reliability**	**Quality**	**Method effects**	**Validity**	**Reliability**	**Quality**	**Method effects**
Make time –method 1	0,92	0,66	0,60	0,26	0,92	0,72	0,67	0,26
Good at–method 1	0,92	0,79	0,73	0,29	0,92	0,72	0,67	0,29
Unimportant-method 1	0,94	0,69	0,65	0,24	0,94	0,72	0,68	0,24
Learning-method 1	0,94	0,67	0,63	0,25	0,94	0,76	0,71	0,25
Make time –method 2	0,98	0,81	0,79	0,15	0,98	0,81	0,79	0,15
Good at–method 2	0,98	0,83	0,81	0,17	0,98	0,79	0,78	0,17
Unimportant-method 2	0,96	0,34	0,32	0,22	0,96	0,29	0,28	0,22
Learning-method 2	0,98	0,88	0,87	0,13	0,98	0,64	0,63	0,16
Make time –method 3	0,92	0,62	0,58	0,27	0,92	0,67	0,62	0,27
Good at–method 3	0,92	0,86	0,80	0,29	0,92	0,74	0,68	0,29
Unimportant-method 3	0,86	0,30	0,26	0,36	0,86	0,34	0,29	0,36
Learning-method 3	0,94	0,85	0,80	0,23	0,94	0,92	0,87	0,23

2.3 Conclusions:

1. The SB-MTMM experiment on political trust suggests that the method with the range "not at all trust" to "complete trust" performs better in terms of quality.
2. It seems that trust in the police is a more reliable item. People may have more information when giving their opinion.
3. The SB-MTMM experiment on well-being shows that item specific questions perform better than statements.
4. Item 3, "feel that what you do is unimportant" had the lowest reliability. This can be a reaction to the negative formulation of the questions.

3. Democracy module analysis

The "Democracy rotating module" contains questions about a total of twenty one concepts related to democracy. Fourteen out of the total we will label in this analysis "simple" concepts, while the seven remaining we will label "complex" concepts. By "simple" we mean that the concept was evaluated by two questions: one that asked about the ***importance*** of the concept for a democratic system and another one that asked respondents to perform an ***evaluation*** of the current functioning of that concept in his/her country. For "complex" concepts one more question is added: using dichotomous items, respondents are asked to state a prior ***preference*** dealing with the concept at stake before the importance and evaluation questions are asked. Consequently, in this design the formulation of the importance question depends on the answer to the preference question. For example, if in a preference question a person expressed that "governments should only follow the demands of the majority" (over taking into account the demands of minority groups as well) (item C18), the item following will ask "how important do you think it is for a democracy that governments **do not take** into account the demands of minority groups?" instead of "how important do you think it is for a democracy that governments **take** into account the demands of minority groups?" (item C19).

In the following note we will perform two different kinds of analysis in order to evaluate the functioning of the questionnaire used in the pilot study in the United Kingdom and Russia. First, we will conduct an "importance-performance analysis" (IPA) for the "simple" concepts included in the questionnaire. The second part will look at the correlations between the importance and the evaluation questions, to distinguish if there are any signs of answers to the first influencing answers to the latter. A third part of the paper will refer to some problems we find with the "complex" items and finally we will express a few conclusions.

3.1 Importance – Performance Analysis

The technique of "importance-performance analysis (IPA)" was initially introduced in the field of market research (Martilla and James 1977), as a tool to evaluate customer satisfaction with products and services. Basically, the technique consists of analyzing both the *importance* customers give to the different attributes that make up a product or a service and the *evaluation* they make of those same attributes after having made use of them. The hypothesis behind this technique is that "consumer satisfaction is a function of both expectations related to certain important attributes and judgments of attribute performance" (Martilla and James 1977, pg. 77). This may be expressed through the following equation (Fishbein and Ajzen, 1975):

$$S_O = \sum_{i=1}^{n} I_i V_i$$

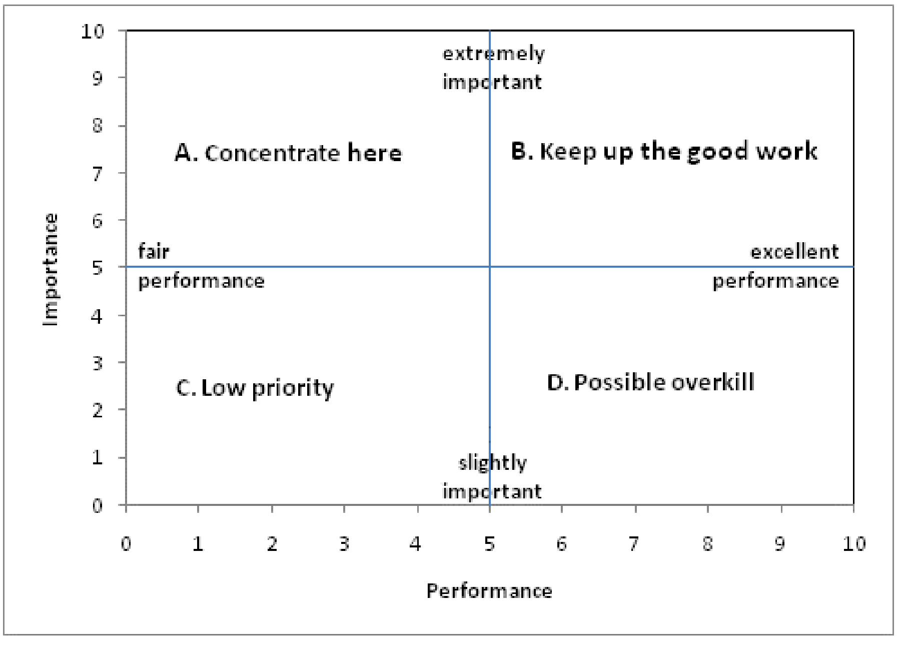
where S_0 is the total satisfaction with the product or service, I_i is the importance each attribute has for the respondents, V_i is the evaluation of the performance, and n is the total number of attributes that make up the product or service.

It is clear then, that not all attributes contribute the same to the final satisfaction a person has with a product, service, or as in our case, a concept such as democracy. Those attributes that respondents qualify as most important will be the most relevant towards the final satisfaction they feel, while the attributes that are seen as less important will be the ones that count the least towards the final satisfaction.

One of the features importance-performance analysis offers is the possibility to graphically display the results on a two-dimensional grid. A usual approach is to plot the points in a graph such as the one figure 1 displays. The four quadrants, labeled by the letters A, B, C and D, are indications of what market researchers call "marketing effort" (Martilla and James 1977, pg. 77). For example, "concentrate here" (quadrant A) denotes an area where attributes are important but performance is evaluated low (thus the need to "concentrate here"). Quadrant B is labeled "keep up the good work" and denotes an area where attributes are important and are evaluated positively. The two bottom quadrants of the graph denote areas of low importance for respondents, the difference being that in quadrant C the evaluation of the performance is low, and in quadrant D it is high, which could imply "possible overkill" of resources.

The positioning of the axes on these graphs is arbitrary. In fact, one of the controversies surrounding importance-performance analysis is the positioning of the vertical and horizontal axes on the grid. The advice from the original developers of the technique was that "positioning the vertical and horizontal axes on the grid is a matter of judgment…(as) the value of this approach lies in identifying relative, rather than absolute, levels of importance and performance" (Martilla and James, pg. 79). Different authors have argued for placement of the axes on arbitrary points that depend on the good judgment of the researcher, on the total means (or medians) of the importance and evaluations, or on the midpoint of the scale. For our analyses, in order to compare the two countries included in the pilot study on the same graph, we have opted for this last option.

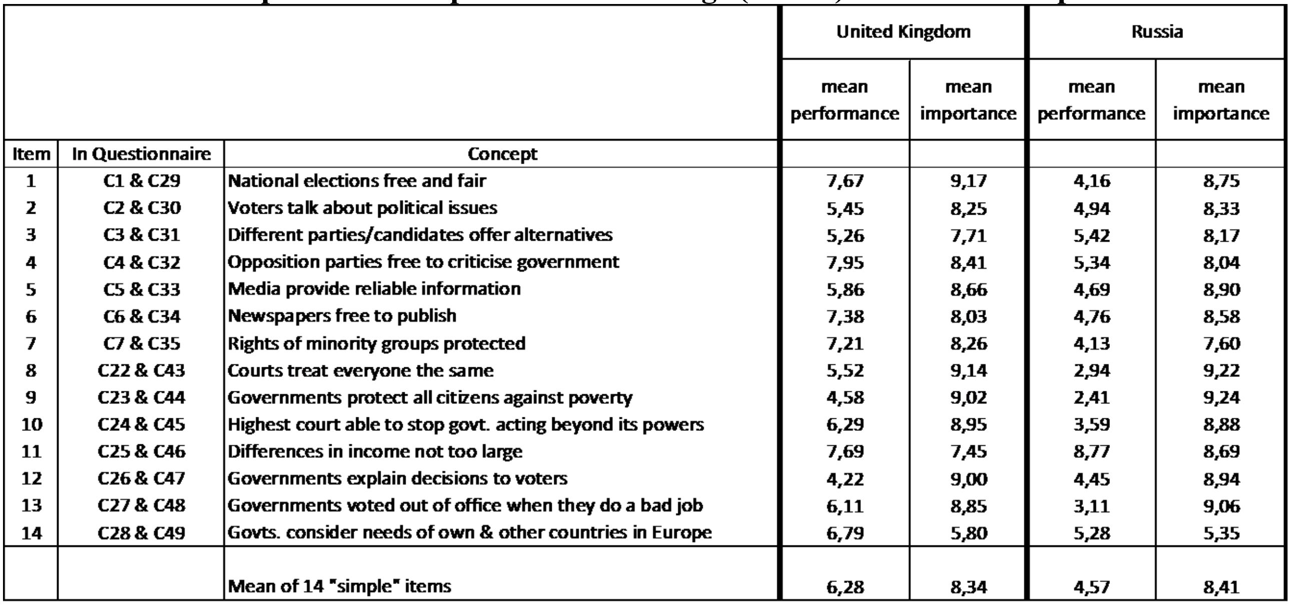
Figure 1.- Classical representation of the Importance-Performance Analysis (as presented by Martilla and James, 1977)



Following this approach, we performed “importance-performance analysis” for the “simple” concepts included in the survey. Table 1 presents the mean importance and performance ratings for these items for both countries. Figure 2 shows this same information in a graphical form.

Both table 1 and figure 2 allow a first analysis of citizens’ expectations, evaluations and general satisfaction with their democracies. It is seen that in both countries the importance that citizens give to the different attributes is similar, being above 7,45 for all items in both countries (the only exception is item #14 which has low importance ratings in both countries). The mean importance of these fourteen “simple” items is very similar too: 8,34 in the U.K. and 8,41 in Russia.

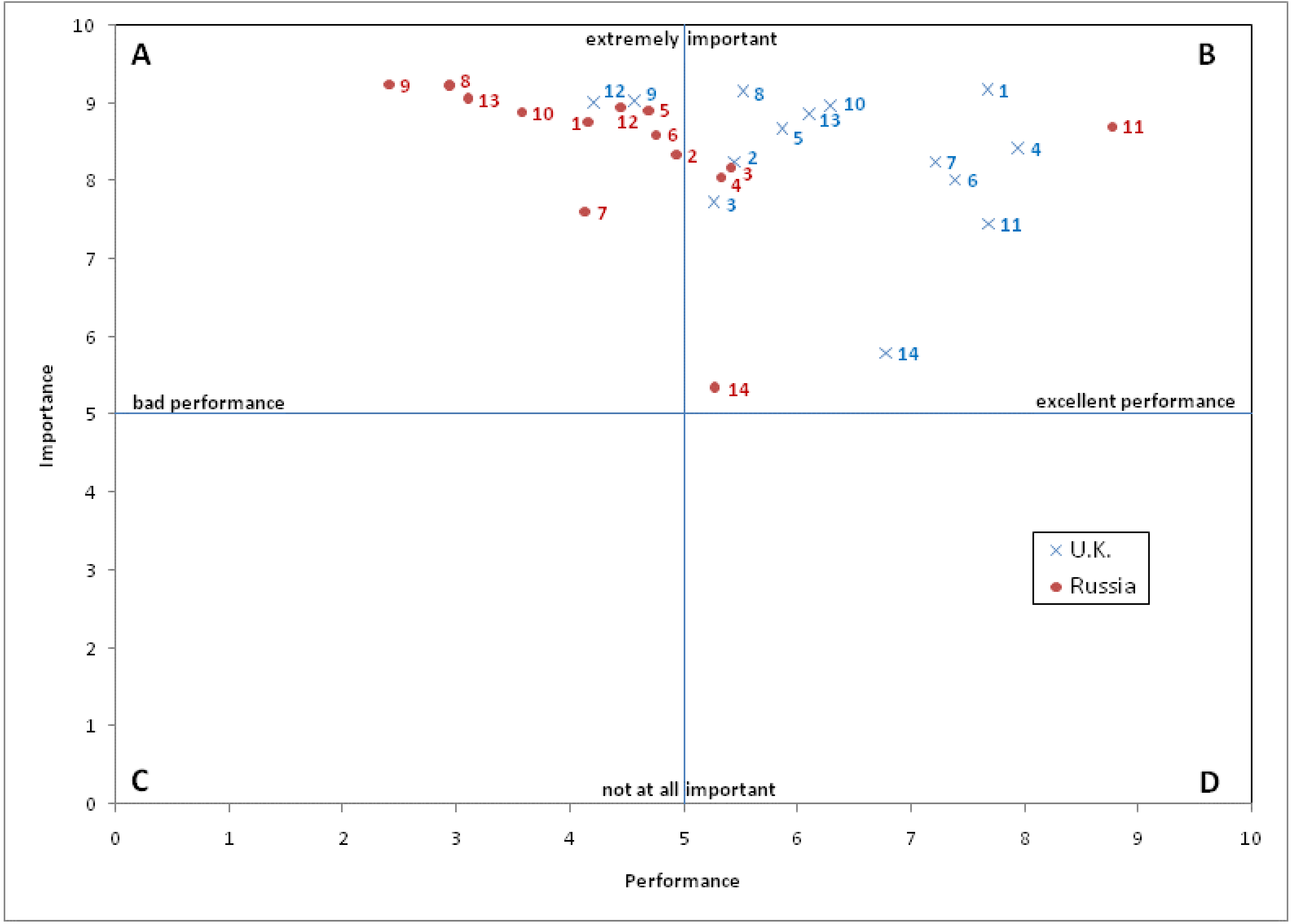
Table 1.- Importance and performance ratings (means) for the 14 “simple” items



			United Kingdom		Russia	
			mean performance	mean importance	mean performance	mean importance
Item	In Questionnaire	Concept				
1	C1 & C29	National elections free and fair	7,67	9,17	4,16	8,75
2	C2 & C30	Voters talk about political issues	5,45	8,25	4,94	8,33
3	C3 & C31	Different parties/candidates offer alternatives	5,26	7,71	5,42	8,17
4	C4 & C32	Opposition parties free to criticise government	7,95	8,41	5,34	8,04
5	C5 & C33	Media provide reliable information	5,86	8,66	4,69	8,90
6	C6 & C34	Newspapers free to publish	7,38	8,03	4,76	8,58
7	C7 & C35	Rights of minority groups protected	7,21	8,26	4,13	7,60
8	C22 & C43	Courts treat everyone the same	5,52	9,14	2,94	9,22
9	C23 & C44	Governments protect all citizens against poverty	4,58	9,02	2,41	9,24
10	C24 & C45	Highest court able to stop govt. acting beyond its powers	6,29	8,95	3,59	8,88
11	C25 & C46	Differences in income not too large	7,69	7,45	8,77	8,69
12	C26 & C47	Governments explain decisions to voters	4,22	9,00	4,45	8,94
13	C27 & C48	Governments voted out of office when they do a bad job	6,11	8,85	3,11	9,06
14	C28 & C49	Govts. consider needs of own & other countries in Europe	6,79	5,80	5,28	5,35
		Mean of 14 "simple" items	6,28	8,34	4,57	8,41

The differences come when looking at the evaluations of the performance of the different attributes. While there are items that behave similarly in both countries, such as numbers 2, 3 and 12, in general, great differences can be seen in the responses between Russian and British citizens. Russian citizens, in general, evaluate these aspects much lower than their British counterparts: in figure 2, Russians answers tend to cluster in the upper left hand part of the graph (quadrant A), while British responses are mostly spread through the upper right part (quadrant B). The mean performance of these fourteen items is also a reflection of this: 6,28 in the U.K. vs. 4,57 in Russia.
The distribution of the scores in figure 2 suggests that the “simple” items are functioning well. There is variation among the 2 axes. The greatest variation is found among the performance axis, which is a good sign. The ranges go from almost 2 to 6 in Russia and from 4 to 8 in the UK, implying that respondents are able to perform evaluations in different terms for the different concepts. In the importance items, there is less variation, with ranges going from 7 to 9 in both countries, but this is not surprising as we could have easily expected importance items to all have scores towards the top of the scale.

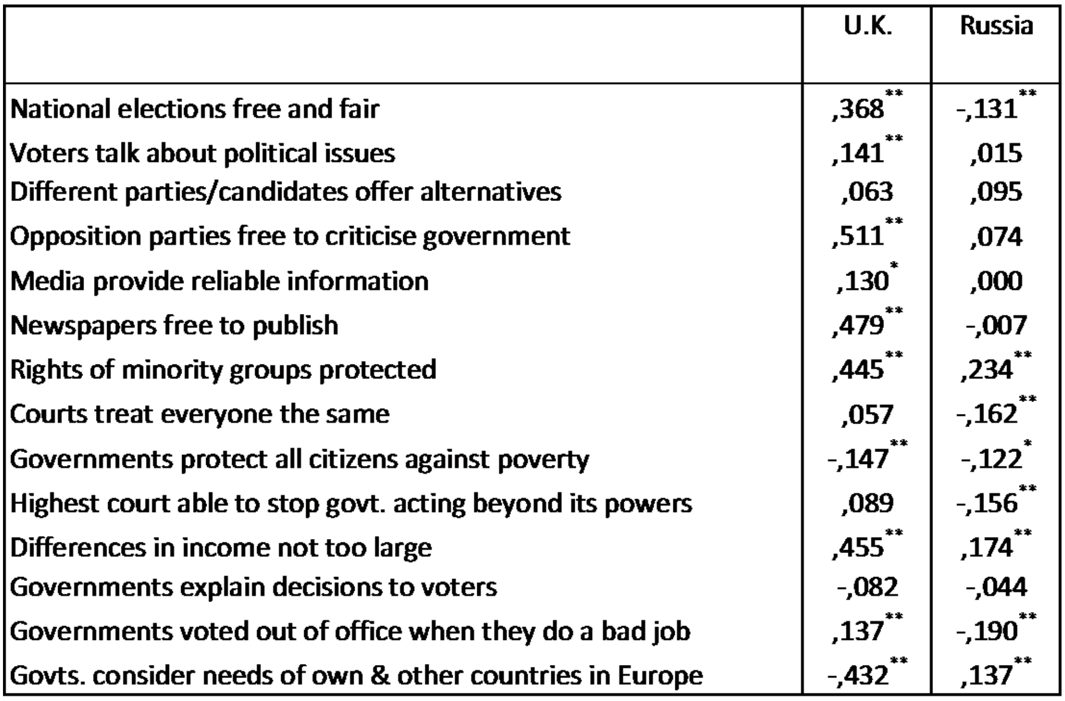
Figure 2.- Importance – Performance Analysis (14 “simple” items)



3.2- Correlations between importance and evaluation items

If the importance and evaluation items would correlate highly, one of these sets of judgments would be redundant. In order to check this we estimated the correlation between the importance and evaluation judgments by the respondents in both countries for each “simple” concept. The results can be seen in table 2. The table clearly indicates that the correlations are not very high. The highest correlation is found in the UK for “opposition parties free to criticize government” but even in this case the correlation is only .511 which means that the overlap of the concepts is not more than ($.511^2$ or) 27% . So the conclusion is clear that the correction of information of both aspects makes sense. We also see that the correlations in Russia are even lower than in the UK.

Table 2.- Correlations between importance and evaluation for the 14 “simple” items



	U.K.	Russia
National elections free and fair	,368**	-,131**
Voters talk about political issues	,141**	,015
Different parties/candidates offer alternatives	,063	,095
Opposition parties free to criticise government	,511**	,074
Media provide reliable information	,130*	,000
Newspapers free to publish	,479**	-,007
Rights of minority groups protected	,445**	,234**
Courts treat everyone the same	,057	-,162**
Governments protect all citizens against poverty	-,147**	-,122*
Highest court able to stop govt. acting beyond its powers	,089	-,156**
Differences in income not too large	,455**	,174**
Governments explain decisions to voters	-,082	-,044
Governments voted out of office when they do a bad job	,137**	-,190**
Govts. consider needs of own & other countries in Europe	-,432**	,137**

3.3. The problems of the complex items

We did the analysis of the simple questions first because there is a problem with the complex questions. The people can have different preferences and indicate the importance of the preferred option but they cannot evaluate their preferred option.

An illustration of the above problem is the group of items C16, C17 and C40, which ask about the concept “freedom of expression”. First, C16 asks every respondent if “Everyone should be free to express their political views openly, even if they are extreme”, or if “Those who hold extreme political views should not be free to express them openly”. The frequency distribution of this question is as the following:

Table 3.- Item C16 (frequencies)

FREE TO EXPRESS EXTREME POLITICAL VIEWS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Everyone should be free to express their political views ope	477	58,0	61,5	61,5
	Those who hold extreme political views should not be free to	229	27,8	29,5	91,0
	Neither of these / it depends	70	8,5	9,0	100,0
	Total	776	94,3	100,0	
Missing	Don't know	47	5,7		
Total		823	100,0		

The table shows that there are indeed different preferences. Next, the items C17a and C17b ask about the importance of freedom of expression, making the distinction between those who think everyone should be free to express political views and those who think not everyone should be free. This creates 2 different variables: those 477 people belonging to the first group answer item C17a and the 229 people belonging to the second group answer item C17b. Table 4 presents the descriptive statistics for variables C17a and C17b.

Table 4.- Items C17a and C17b (descriptive statistics)

Statistics

		C17a. HOW IMPORTANT, FREE TO EXPRESS EXTREME POLITICAL VIEWS	C17b. HOW IMPORTANT, EXPRESSION OF EXTREME POLITICAL VIEWS NOT FREE
N	Valid	471	222
	Missing	352	601
Mean		7,55	7,55
Std. Deviation		2,025	2,284

Both groups show exactly the same mean importance in their questions, but the two are concepts completely opposite to each other! C17a asks about how important is that everyone is free and it is assumed that freedom of expression here is seen as a positive value. On the other hand, C17b asks how important it is that freedom of expression is not free and we can thus assume that for this group of people freedom of expression is something negative.

Later in the questionnaire, in item C40, again the whole sample is asked about freedom of expression, but this time in terms of an evaluation. The question asks "to what extent do you think people in (country) are free to express their political views openly, even if they are extreme?" It is not clear here if they consider this as a good thing or a bad thing. It is just a description of the situation. It is clear that using this item (C40) by itself as an evaluation variable would be a mistake. So the question is how to get an evaluation?

We do not see how this can be done using the present questions. One solution to this problem would be to include another evaluation item for each one of the “complex” items. This way, each importance item would be paired with its own evaluation item. If these questions are asked directly after each other no extra questions are needed. If they are separated in space one has first to ask the preference again or use the previous obtained preferences to formulate the evaluation questions. However we fear that this approach would lead to a lot of errors given the number of questions and the complexity of the process.

3.4 Conclusions

1. The “simple” items seem to be functioning well: citizens are able to differentiate between levels of importance of the different attributes, and are able to give different evaluations to the different attributes as well.
2. There seems not to be any problems with importance questions influencing responses to evaluation questions. These sets of items measure really different judgments.
3. It is not clear what the “direction” of the “evaluation” items is in some of the complex items. This makes it difficult to perform a proper analysis, and most likely.