Appendix A - Loevinger's H

To test the assumption of *unidimensionality*, we can perform a test of homogeneity based on Loevinger's H coefficient, which measures whether the responses of the items in the scale are consistent (Wheatley 2016). Loevinger's H coefficient equals 1 when the items form a perfect Mokken scale, and 0 when there is no association between the answers of the users on the items (Stochl et al. 2012). For different items X_i and X_j The item scalability coefficient H_i is defined as:

$$H_i = \frac{\sum_{j \neq i} Cov(X_i, X_j)}{\sum_{j \neq i} Cov_{max}(X_i, X_j)} \tag{1}$$

while the coefficient H for all k items is:

$$H = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} Cov(X_i, X_j)}{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} Cov_{max}(X_i, X_j)} \tag{2}$$

Generally, items are said to form a Mokken scale if the coefficient H is larger than 0.3 and each item-specific H_i is also larger than 0.3 (Germann and Mendez 2016). H values between 0.30 and 0.40 form a weak scale, values between 0.40 and 0.50 a medium scale, and all values ≥ 0.50 a strong scale (Mokken 1971, p.185).

Appendix B - The Latent Class Reliability Coefficient

The Latent Class Reliability Coefficient (LCRC) is a new measure of reliability based on latent class models (van der Ark et al. 2011). These models relate a set of observed categorical variables (or items) to a set of latent unobserved latent categorical variables or items - which is the same as the goal of IRT.

To see how this works, we take $\pi_{x(i)}$ to be the probability that a user drawn at random gives a correct answer to item i. Note that a correct answer in IRT context means that the user answers the item as expected based on their position on the latent dimension we are trying to measure. For example, given an item on leaving the European Union, the correct answer for someone who is a proponent of leaving the European Union would be to *agree* with the statement. An incorrect answer would be when they *disagree*. Going back to the formula, we also take $\pi_{x(j)}$ to be the probability that a user drawn at random answers item j correct. Thus, $\pi_{x(ij)}$ is the probability that a user drawn at random answers both items i and j correct, and $\pi_{x(ii)}$ the probability that they answer item i correct in two independent repetitions - the quantity we are after but do not know as the questionnaire has not been repeated. Furthermore, we let σ_X^2 denote the variance of an item, and $\rho_{XX'}$ the correlation between the original response (X) and the response in a possible re-test (X'). Note that $\rho_{XX'}$ is the same measure as we discussed earlier with classical test theory. Applied to IRT, Molenaar and Sijtsma (1988) show that it can be stated as:

$$\rho_{XX'} = \frac{\sum\limits_{i \neq j}\sum\limits_{x}\sum\limits_{y}\sum[\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]}{\sigma_X^2} + \frac{\sum_i \sum_x \sum_y \pi_{x(i),y,(i)} - \pi_{x_i}\pi_{y_i}}{\sigma_X^2} \tag{3}$$

where the nominator is like the term σ_T^2 in CTT. The denominator in both the left and right hand part of the equation represents — in CTT terms — the variance of the true score that we are trying to measure. Yet, while we can measure the left hand side of the equation, the right hand side is problematic, as $\pi_{x_i}\pi_{y_i}$ represents the probability of obtaining sore x and y on two independent instances of the same item for the same respondent - which, as we saw, is problematic. To get around this, we can estimate the right hand side of the equation using latent class models. These latent class models are unrestrictive, in the sense that they do not assume any single underlying latent factor. The formula then becomes:

$$\begin{aligned} LCRC = & \frac{\sum\limits_{i \neq j}\sum\limits_{x}\sum\limits_{y}\sum[\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]}{\sigma_X^2} \\ & + \frac{\sum\limits_{i}\sum\limits_{x}\sum\limits_{y}[\sum\limits_{u=x}^{m}\sum\limits_{v=y}^{m}\sum\limits_{k=1}^{K} P(\zeta = k)P(X_i = u|\zeta = k)P(X_i = v|\zeta = k) - \pi_{x(i)}\pi_{y(i)}]}{\sigma_X^2} \end{aligned} \tag{4}$$

where ζ is the underlying number of latent classes of the series of items. The number of latent classes has thus to be defined first, which can be done through confirmatory factor analysis, in the way described earlier.

As the LCRC does not assume a single underlying factor, as α does and allows for differences in item difficulties, which ω lacks, van der Ark et al. (2011) find that the LCRC is the least biased and should come closest to the true reliability of the model. Moreover, the LCRC does not assume monotonicity and non-intersecting item response functions.

To calculate the LCRC in **R** I used the **mokken** package (van der Ark 2007, 2012), and calculated the number of latent classes using the **poLCA** package (Linzer and Lewis 2011, 2013) with 5000 iterations and 5 repetitions.

Appendix C - The Dirty Data Index

Blasius and Thiessen (2012) give the following set of steps to calculate the DDI. Note that to calculate the quantification values I used the CATPCA package provided in SPSS (Meulman et al. 2004). Other options to calculate these values are the **homals** or **gifi** packages in **R** (de Leeuw and Mair 2009; Mair and de Leeuw 2017). First, one calculates the mid-points of the item(s), starting with the mass of the first category and dividing it by 2:

$$g_{1k} = m_{1k}/2 \quad (\text{for } j = 1) \tag{5}$$

Here, g_jk is the (first) mid-point of category 1 for item k and m_1 the mass of category 1 (being the percentage of cases in item 1) for item k. The masses for each category are given by $m_j = f_{jk}/N$, where m_j is the mass for category j, f_{jk} is the percentage of cases for category j in item k, and N is the total number of cases for all the items (which is thus the same for all the categories).

For the second mid-point, m_{1k} is added to the half the mass of the second category $(m_{2k}/2)$:

$$g_{2k} = g_{1k} + (m_{2k}/2) \quad (\text{for } j = 1) \tag{6}$$

This procedure is repeated for each of the masses of the later categories, so that in effect the first of the (J_{K-1}) masses $(c_{(1,j-1)})$ plus half the mass of the last category is added, with the number of thresholds being the same as the number of categories J_K. Here, $c_{(1,j)k}$ represents the cumulative masses of the categories j of item k, which is calculated by $c_{(1,j)} = m_j + c_{(1,j-1)}$ if $(j = 1, c_{(1,j-1)} = 0)$. Thus, for $j = 1$ to J_k (for each item k):

$$g_{jk} = (g_{jk} - 1) + (m_{jk}/2) \quad (\text{with } g_0 = 0) \tag{7}$$

Here J_K are the number of categories in each item, j the specific category, K the total number of the item and k the specific item. Next, one calculates the area-under-the-curve to the left of the quantification value q_jk (the quantification of category j of item k) by finding the area that corresponds to that value on the standard normal function $N(\mu, \sigma^2)$. This gives the quantification area q_{jk} for category j of item k. Then the difference areas d_jk between the quantification area and the thresholds for each category j is calculated and added to get the total of the areas of difference d_k:

$$d_k = \sum_{j=1}^{j} g_j - q_j \tag{8}$$

Then, d_k is standardized by an upper bound, which is $l/(l-1)$ with l being the number of categories. This gives the DDI for a single item k. This procedure is then repeated for all the other items k which, when added and divided by the total number of items K, gives DDI for a given scale. Blasius and Thiessen (2012) advise to interpret values smaller than 0.3 and 0.15 as indicating data of good and exceptional quality and values exceeding 0.5 as indicating data of bad quality.

Appendix D - EUVox Data Cleaning

Table 1 Number of users in the EUVox data-set. Each column shows the number of users with the specific characteristics mentioned.

		Number of Users with:								
Country	Original	Mobile	Return	Ans. $<$ 2 sec.	$>$ 3 Ans. $< 3s$	$> 5399s$	$< 121s$	$>$ 10 Similar	$>$ 10 NA	Final
AT	10,669	1,538	951	398	215	6	22	5	7	7,527
BG	7,214	364	636	287	63	19	10	31	28	5,776
CY	5,176	671	613	225	87	6	3	20	9	3,542
CZ	28,630	1,059	1,345	849	1,098	48	72	33	42	24,084
DE	9,658	1,201	379	408	370	18	45	10	19	7,208
DK	126,261	23,558	5,552	2,936	832	145	124	72	409	92,633
EE	18,172	577	628	3,432	1,168	29	0	19	52	12,267
ES	159,552	41,716	16,855	3,478	1,981	124	159	139	396	94,704
FI	8,274	535	411	296	225	12	24	8	34	6,729
FR	8,704	736	690	346	229	16	9	15	7	6,656
GR	63,687	7,020	7,907	1,525	698	81	91	230	37	46,098
HR	7,666	863	948	285	238	7	17	19	8	5,281
HU	6,711	209	580	227	119	6	15	5	14	5,536
IE	9,523	1,894	458	369	517	11	63	5	8	6,198
IT	36,614	6,034	2,691	849	600	27	89	56	33	26,235
LT	9,072	643	900	265	161	20	17	7	9	7,050
PL	73,521	4,201	6,845	1,951	1,621	94	162	141	77	58,429
PT	54,165	4,461	4,925	1,404	817	96	101	73	89	42,199
RO	9,508	556	616	310	53	13	7	11	20	7,922
SI	3,842	236	443	181	168	5	11	5	10	2,783
SK	7,238	525	389	247	129	11	12	16	4	5,905
UK†	100,897	6,730	5,425	3,782	6,406	75	803	117	156	77,403

†Includes only England

Appendix E - EUVox Core Items

Table 2 Overview of the EUVox core items

Code	Item
EU1	Country X should exit the Euro (Eurozone countries)/ never adopt the Euro (non-Eurozone countries)
EU2	A single member state should be able to block a treaty change, even if all the other member states agree to it
EU3	The right of EU citizens to work in Country X should be restricted
EU4	There should be a common EU foreign policy even if this limits the capacity of Country X to act independently
EU5	The EU should redistribute resources from richer to poorer EU regions
EU6	Overall, EU membership has been a bad thing for the Country X
EU7	EU treaties should be decided by [name of national parliament] rather than by citizens in a referendum
EC1	Free market competition makes the health care system function better
EC2	The number of public sector employees should be reduced
EC3	The state should intervene as little as possible in the economy
EC4	Wealth should be redistributed from the richest people to the poorest
EC5	Cutting government spending is a good way to solve the economic crisis
EC6	It should be easy for companies to fire people
EC7	External loans from institutions such as the IMF are a good solution to crisis situations
CU1	Immigrants must adapt to the values and culture of Country X
CU2	Restrictions on citizen privacy are acceptable in order to combat crime
CU3	To maintain public order, governments should be able to restrict demonstrations
CU4	Less serious crimes should be punished with community service, not imprisonment
CU5	Same sex couples should enjoy the same rights as heterosexual couples (FR) / to marry (remaining countries)
CU6	Women should be free to decide on matters of abortion
CU7	The recreational use of cannabis should be legal

Appendix F - Original EUVox Scales

Table 3 Items included in the original EUVox Scales. Questions with a ⋆ belong to EU scales, questions with a ○ to economic scales, and questions with a ● to cultural scales.

	EU							EC							CU							AD									
Cty.	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	8	9	10
AT	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●		⋆		●	⋆	⋆	○	○		
HR	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	○	●	●	○	○	○	○			
CZ	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	○	●		⋆	⋆	⋆	○			
DK	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	⋆	⋆	⋆	●	⋆	⋆	⋆			
EE	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	⋆	⋆	○		○	●				
FI	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●										
FR	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●		○		●	⋆	⋆	○	○	●	●
DE	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●				⋆	⋆					
GR	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●		○	○	●	●	●	⋆		○	
HU	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	⋆	⋆						●	●	
IE	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	○		○	●	⋆	⋆	○	○	○	
IT	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	●	●	○	○	○	○				
LT	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	⋆	○		⋆	○		●			
PL	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	●	●	●	⋆	○	●				
PT	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	⋆	○	○	○	○	○	○	●	○	
SL	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	○	●		⋆	⋆			⋆	○	
UK	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	○	⋆	●	●	●	⋆	⋆	○	○	

Appendix G - DSV EUVox Scales

Table 4 Items included in the Final EUVox Scales. Questions with a ⋆ belong to EU scales, questions with a ○ to economic scales, and questions with a ● to cultural scales.

	EU							EC							CU							AD									
Cty.	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	8	9	10
AT	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○		●	●	●		●	●	●				●	⋆		○			
HR	⋆	⋆	⋆	⋆		⋆			○	○			○		●				●	●	●	○	●	●							
CZ	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○		○		●				●			○	●	●	⋆	⋆	⋆	○			
DK	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○		●	●	●		●			⋆	⋆	⋆	●	⋆		⋆	○		
EE	⋆	⋆	⋆	⋆					○	○		○	○						●	●	●										
FI	⋆	⋆	⋆	⋆		⋆	⋆	○	○	○	○	○	○		●				●	●										●	
FR	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○		●	●	●		●	●	●		●	●			⋆	○			●
DE	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○		●	●	●		●	●	●				⋆						
GR	⋆	⋆	⋆	⋆		⋆		○	○	○	○		○	○	●				●	●	●		○	○	●	●	●	⋆			
HU	⋆	⋆	⋆	⋆		⋆		○	○	○				○	●	●			●	●	●	⋆	⋆	⋆							
IE	⋆	⋆	⋆	⋆		⋆		○	○	○	○	○	○	○		●	●		●	●	●			○					○		
IT	⋆	⋆	⋆	⋆		⋆	⋆	○	○	○		○	○		●		●		●	●	●	●	●								
LT	⋆	⋆	⋆	⋆		⋆		○		○	○		○		●		●		●	●	●	⋆	○		⋆						
PL	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○		○		●				●	●		●	●	●	⋆			●			
PT	⋆	⋆		⋆		⋆		○	○	○	○	○	○	○	●	●	●		●	●	●	⋆	○	○				○		○	
SL	⋆	⋆	⋆	⋆		⋆		○	○	○			○		●				●	●	●	○	●		⋆			⋆	⋆		
UK	⋆	⋆	⋆	⋆	⋆	⋆			○	○	○	○	○		●			●	●			○	⋆	●	●	●		⋆			

Appendix H - Quasi-Inductive EUVox Scales

Table 5 Items included in the Quasi-Inductive EUVox Scales. Questions with a ⋆ belong to EU scales, questions with a ○ to economic scales, and questions with a ● to cultural scales.

	EU							EC							CU							AD									
Cty.	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	8	9	10
AT	⋆	⋆	⋆	⋆	⋆	⋆		○		○	○	○	○		⋆	●	●		⋆		●	⋆	⋆	⋆	⋆			○			
HR	⋆	⋆	●	⋆		⋆			○				○		●				●	●	●	○	●	●							
CZ	⋆	⋆		⋆	⋆	⋆		○	○	○	○		○	⋆	⋆							○		⋆	⋆	⋆	⋆				
DK	⋆	⋆	○	⋆	○	⋆		○	○	○	○	○			⋆							⋆	⋆	⋆	⋆	⋆		⋆	○		
EE	⋆	⋆		⋆	○	⋆	⋆				○				●								●			○	●				
FI	⋆	⋆	⋆	⋆	○	⋆	⋆	○	○	○	○	○	○		⋆	●	●		⋆		●	○		○						○	
FR	⋆	⋆	○	⋆	○	⋆		○	○	○	○	○	○		○		○		⋆	⋆				⋆	○		⋆	○	○	○	○
DE	⋆	⋆	⋆	⋆	⋆	⋆	●	○	○	○	○	○	○	●	⋆				⋆			⋆		⋆	⋆	●			●	●	
GR	⋆		●			⋆	⋆	⋆	⋆	⋆		⋆	⋆	⋆		⋆	⋆		●	●	●	⋆	⋆	⋆	●	●	●			⋆	
HU	⋆	⋆	⋆	⋆		⋆		⋆		⋆				⋆	⋆				⋆	⋆		⋆	⋆	⋆							
IE	⋆	⋆	●	⋆	○	⋆	⋆	○	○	○	○	○	○		●					●	●	⋆	⋆	○	●						
IT	⋆		⋆			⋆		○	○	○		○	○		⋆		●		⋆	●/⋆	●	⋆	⋆			⋆					
LT	⋆					⋆	⋆	○		○			○						⋆			⋆	⋆	⋆	⋆		⋆				
PL	⋆	⋆		⋆	○	⋆		○	○	○	○		○		⋆				⋆	⋆		⋆	⋆					○			
PT	⋆					⋆		⋆	⋆	⋆			⋆	⋆			⋆		●	●	●		⋆	⋆				⋆			
SL	⋆		⋆	⋆		⋆		○	○	○			○		⋆				⋆	●	●	○	⋆	⋆	⋆					●	
UK	⋆	⋆	⋆	⋆	⋆	⋆			⋆	⋆	⋆	⋆	⋆		⋆				⋆			⋆	⋆	⋆	⋆	⋆		⋆		⋆	

Appendix I - Procedure

The procedure to generate the quasi-inductive scales is carried out here using the **mokken** package (van der Ark 2007, 2012) as implemented in **R**. To generate the scales, one should take the following steps:

1. Load the data into R. The resulting data frame should have the variables in the columns and the user responses in the rows. Only response values are allowed in the cells. All NA values should be classified as such.
2. Generate reverse responses for each variable. Thus, a response value of 2 in variable A on a five-point scale is copied as a response value of 4 in a new variable A-reverse. The resulting data-set has twice the number of variables as the original data-set.
3. Load the **mokken** package
4. Run the **aisp** (Automated Item Selection Procedure). Specify the data-frame with the original and reversed values as the data input, set the option *search* to **ga** (referring to the genetic algorithm used here) and the option *lowerbound* to 0.3.
5. The resulting object has the number of scales in duplicate (as each of the items is included both in its original or reversed form).
6. Construct two new matrices based on the variables occurring in the scales that were the output of the **aisp** procedure
7. Run the **coefH** procedure on each of the matrices to get the values for Loevinger's H
8. Run the **check.monotonicity** procedure using each of the matrices and with the *minsize* set to 100 to find the crit values.
9. If any of the crit-values are > 80 or H-values are < 0.3, remove that item from the scale and the matrix and re-run both the **coefH** and **check.monotonicity** until no such values remain

Appendix J - Loevinger's H

Table 6 H Values based on the original and DSV scales for EUVox

	Original			DSV			Quasi-Inductive		
Country	EC	EU	CU	EC	EU	CU	EC	EU	CU
Austria	0.26	0.39	0.30	0.34	0.45	0.33	0.37	0.41	0.37
Croatia	0.15	0.28	0.32	0.37	0.39	0.50	0.41	0.45	0.45
Czech Republic	0.27	0.43	0.20	0.37	0.43	0.47	0.39	0.40	–
Denmark	0.30	0.20	0.27	0.36	0.21	0.35	0.41	0.39	–
Estonia	0.15	0.20	0.13	0.26	0.33	0.23	0.44	0.38	0.36
Finland	0.33	0.42	0.28	0.40	0.46	0.36	0.40	0.33	0.39
France	0.39	0.37	0.40	0.45	0.42	0.42	0.39	0.38	–
Germany	0.28	0.41	0.28	0.36	0.46	0.32	0.36	0.43	0.37
Greece	0.36	0.27	0.34	0.42	0.36	0.45	–	0.38	0.45
Hungary	0.20	0.42	0.29	0.41	0.52	0.41	–	0.48	–
Ireland	0.24	0.24	0.26	0.34	0.38	0.31	0.36	0.38	0.34
Italy	0.21	0.33	0.32	0.33	0.42	0.41	0.33	0.39	0.36
Lithuania	0.17	0.27	0.19	0.30	0.35	0.30	0.33	0.38	–
Poland	0.19	0.40	0.19	0.41	0.40	0.46	0.40	0.46	–
Portugal	0.30	0.26	0.25	0.36	0.40	0.33	–	0.37	0.48
Slovakia	0.21	0.27	0.22	0.37	0.32	0.37	0.37	0.35	0.36
United Kingdom*	0.32	0.52	0.38	0.46	0.60	0.54	–	0.46	–

* Only includes England

Appendix K - Latent Class Reliability Coefficient

Table 7 LCRC Values based on the original and DSV scales for EUVox

	Original			DSV			Quasi-Inductive		
Country	EC	EU	CU	EC	EU	CU	EC	EU	CU
Austria	0.74	0.83	0.77	0.76	0.83	0.77	0.75	0.89	0.58
Croatia	0.67	0.70	0.82	0.67	0.74	0.85	0.72	0.85	0.67
Czech Republic	0.78	0.87	0.68	0.79	0.87	0.74	0.77	0.88	–
Denmark	0.76	0.81	0.73	0.80	0.81	0.72	0.80	0.89	–
Estonia	0.63	0.65	0.54	0.56	0.60	0.50	0.67	0.69	0.55
Finland	0.77	0.80	0.72	0.78	0.82	0.65	0.83	0.83	0.60
France	0.84	0.82	0.85	0.85	0.83	0.85	0.90	0.83	–
Germany	0.73	0.85	0.71	0.75	0.85	0.73	0.75	0.88	0.74
Greece	0.85	0.74	0.84	0.84	0.76	0.84	0.84	0.89	–
Hungary	0.67	0.86	0.76	0.72	0.89	0.76	–	0.92	–
Ireland	0.78	0.70	0.73	0.80	0.71	0.71	0.78	0.81	0.71
Italy	0.73	0.76	0.81	0.69	0.80	0.83	0.68	0.85	0.61
Lithuania	0.63	0.73	0.63	0.63	0.77	0.66	0.54	0.82	–
Poland	0.69	0.81	0.77	0.75	0.81	0.85	0.81	0.89	–
Portugal	0.85	0.72	0.72	0.85	0.75	0.73	–	0.85	0.69
Slovakia	0.69	0.75	0.72	0.72	0.77	0.76	0.71	0.81	0.60
United Kingdom*	0.83	0.91	0.85	0.82	0.92	0.84	–	0.94	–

* Only includes England

Appendix L - Dirty Data Index

Table 8 Overview of the DDI Scores for the Original, DSV and Quasi-Inductive Scales.

	Original			DSV			Quasi-Inductive		
Country	EC	EU	CU	EC	EU	CU	EC	EU	CU
Austria	0.22	0.38	0.26	0.16	0.25	0.21	0.13	0.23	0.17
Croatia	0.47	0.16	0.37	0.14	0.27	0.22	0.12	0.22	0.22
Czech Republic	0.25	0.26	0.44	0.20	0.26	0.11	0.18	0.26	–
Denmark	0.11	0.18	0.18	0.17	0.17	0.17	0.19	0.19	–
Estonia	0.36	0.40	0.52	0.14	0.19	0.60	0.07	0.33	0.39
Finland	0.13	0.20	0.21	0.12	0.17	0.22	0.10	0.17	0.21
France	0.20	0.47	0.16	0.15	0.41	0.17	0.18	0.38	–
Germany	0.26	0.33	0.16	0.29	0.17	0.17	0.29	0.19	0.31
Greece	0.29	0.27	0.12	0.28	0.29	0.17	–	0.26	0.14
Hungary	0.24	0.20	0.34	0.18	0.26	0.32	–	0.24	–
Ireland	0.30	0.39	0.27	0.25	0.26	0.22	0.20	0.28	0.24
Italy	0.32	0.17	0.26	0.13	0.19	0.15	0.13	0.17	0.12
Lithuania	0.42	0.16	0.41	0.27	0.24	0.16	0.31	0.15	–
Poland	0.49	0.37	0.33	0.44	0.37	0.18	0.42	0.23	–
Portugal	0.28	0.23	0.32	0.21	0.23	0.14	–	0.17	0.20
Slovakia	0.28	0.32	0.35	0.17	0.32	0.18	0.17	0.22	0.30
United Kingdom*	0.25	0.32	0.22	0.17	0.29	0.22	–	0.19	–

* Only includes England

Appendix M - Dot plots for the Correlations

Fig. 1 Correlations for the Original Scales

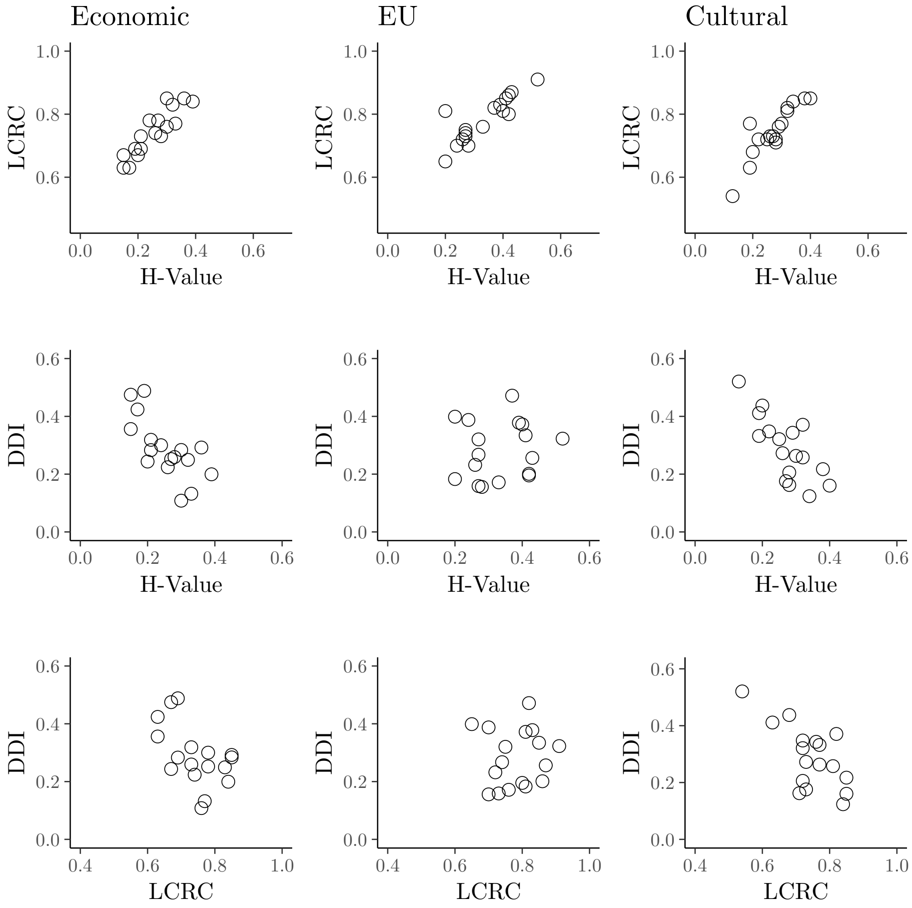
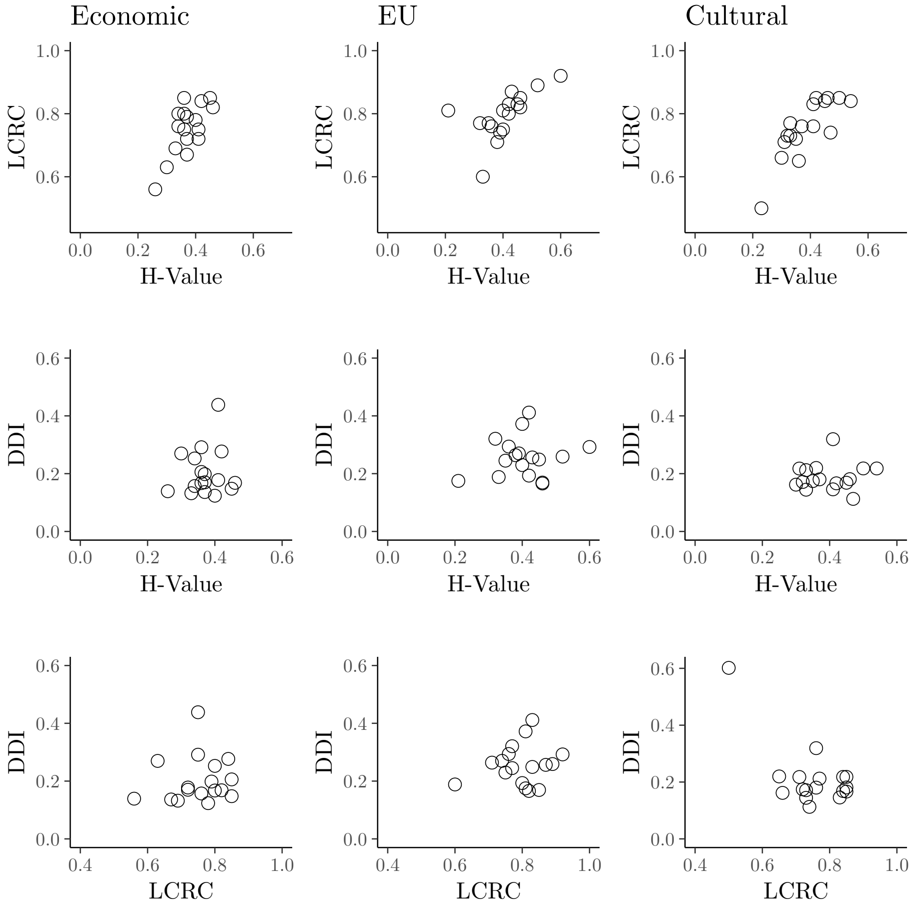


Fig. 2 Correlations for the DSV Scales

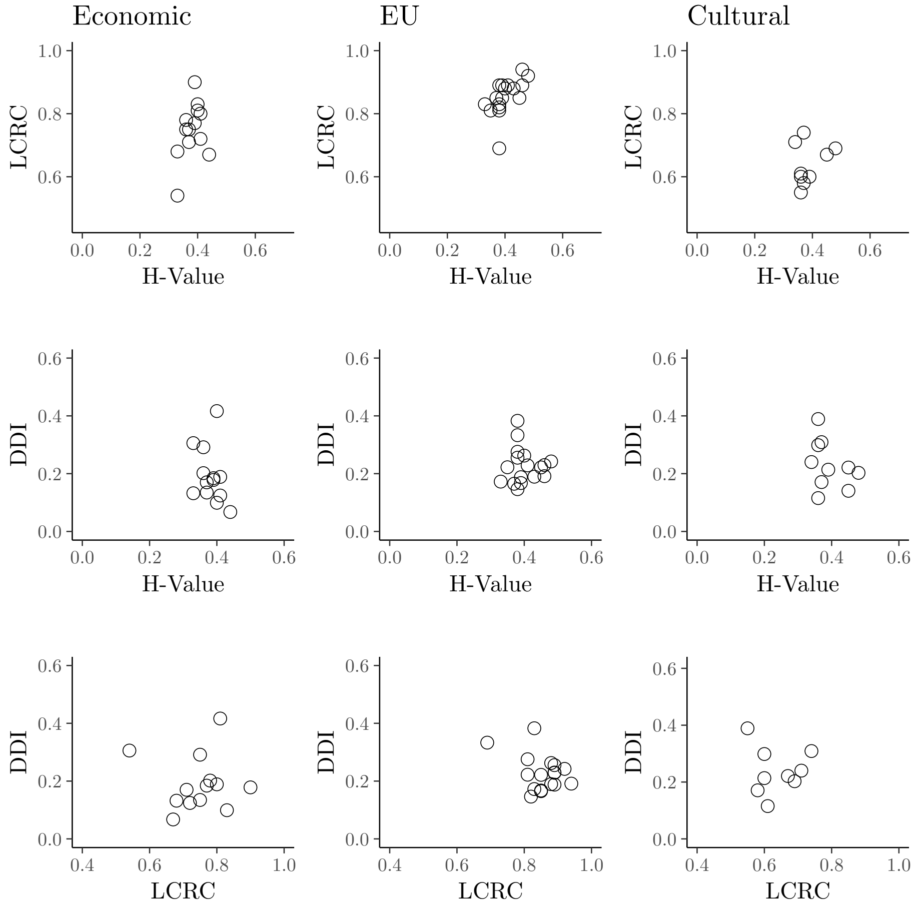


Appendix N - Robustness

One of the main problems with the analysis as given in the main paper is that it depends on a single data-set - the EUVox 2014 dataset. This data-set was specifically designed for the elections for the European Parliament in 2014. To guarantee that the recommendations as given in the main article do not depend solely on this data-set, I here re-run the analysis on a very similar data-set. I do so using data from the 2014 *euandi* VAA (Trechsel, Alexander H. and Garzia, Diego and De Sio, Lorenzo 2015). This was a VAA launched during the same elections as the EUVox VAA, with a similar aim to provide a VAA for almost all countries in the EU. As such, I expect to reach similar conclusions for the scales in this VAA as I did for those in the EUVox VAA.

Here, I use the data for four countries: Germany, Estonia, Italy and Poland. I do so as these countries had a high number of users, as in other countries the number was often too low. After cleaning, which was similar to the cleaning for the EUVox data except for the step that removed mobile phone users, the step that removed users who responded to any

Fig. 3 Correlations for the Quasi-Inductive Scales



items in less than 2 second, and the step that removed users who responded to three items in 3 seconds or less, as timer data for the individual items was not available. As returning users were already removed from the data-set it was unnecessary to remove them as well. In total, this resulted in 35114 users for Germany, 8219 users for Estonia, 135030 users for Italy, and 16292 users for Poland.

The *euandi* VAA assigned each of its 30 items to three scales, the composition of which was equal for all countries: an economic left-right scale (EC), a traditional-liberal values scale (SO), and a pro-anti EU integration scale (EU). In contrast to the EUVox scale, some items loaded on more than a single scale.

Figure 9 shows the results. There are a few points of interest here. First, when using the quasi-inductive algorithm, only a single scale remained in all countries that combined all the other three scales, with a primary focus on EU issues. Given that the VAA was designed for the EU Parliament elections, this is not unexpected. For the DDI, Germany and Italy show a good performance on all scales, while Estonia and Poland fare worse, with the 0.51 for the EC scale in Poland being especially high. This is remedied in the DSV scales, as the EC

Table 9 DDI, Loevinger's H, and LCRC values for the *euandi* VAA

Type	Country	DDI			Loevinger's H			LCRC		
		EC	EU	SO	EC	EU	SO	EC	EU	SO
Original	Germany	0.18	0.17	0.18	0.28	0.27	0.24	0.79	0.79	0.79
	Estonia	0.40	0.46	0.32	0.11	0.14	0.17	0.58	0.61	0.69
	Italy	0.28	0.18	0.29	0.18	0.19	0.29	0.68	0.72	0.82
	Poland	0.51	0.36	0.27	0.21	0.34	0.27	0.73	0.83	0.82
DSV	Germany	0.19	0.18	0.17	0.36	0.46	0.39	0.76	0.82	0.79
	Estonia	-	0.25	0.12	-	0.37	0.36	-	0.57	0.60
	Italy	0.11	0.17	0.22	0.33	0.43	0.42	0.57	0.74	0.79
	Poland	0.37	0.39	0.27	0.37	0.49	0.36	0.77	0.86	0.81
Quasi-Inductive	Germany	-	0.21	-	-	0.40	-	-	0.88	-
	Estonia	-	0.20	-	-	0.35	-	-	0.78	-
	Italy	-	0.21	-	-	0.40	-	-	0.88	-
	Poland	-	0.32	-	-	0.48	-	-	0.91	-

scale for Estonia is dropped, though Poland still has all scales higher than (or close to) 0.30. With the quasi-inductive scales, the single scales all have values around 0.20, with again the exception of Poland. For Loevinger's H, we find that none of the scales in their original form passed the 0.30 threshold. That they all did in the DSV scales is due to the algorithm, do this did lead to the EC scale for Estonia being dropped. Here, especially the EU scales score high values, which are not very dissimilar from the values they receive in their quasi-inductive version. For the LCRC, we find that none of the scales passes the criterion of 0.9, except for the quasi-inductive version of the EU scale in Poland. Two of the other scales of that type - in Germany and Italy - come close with a 0.88 value. Otherwise, the values lie between 0.50 and 0.80, with the Estonian scales scoring poorest.

The picture for this VAA is thus quite like the one formed by the EUVox VAA: the original scales under-perform on all three criteria, the DSV scales improve, though sometimes disappear for a lack of items, and the quasi-inductive scales are longer and have the highest values on the criteria, there is often only a single scale per country. Also, for (almost) none of the scales does the LCRC reach a value higher than 0.90. Therefore, also here the criteria show their usefulness in that they show that the original scales are insufficient and that the two potential replacements score better in almost all regards, with the choice being up to the VAA designer which scales they require for their VAA.

Appendix O - Design Exercise

This appendix describes how designers could carry out the validation steps as proposed in the main section of the text. As is common with most VAAs, designers will be able to get a data-set in an appropriate format, with the variables in the columns and the responses of the users in the rows.

1. **Decide on timing**. Depending on the country and the election, this is most often done after the election ends. Doing this allows the designer to check afterwards whether the VAA did indeed live up to its expectations and - if not- can take lessons for the next time. Alternatively, one can download the data earlier, as suggested by Germann et al. (2015) and Germann and Mendez (2016) for Dynamic Scale Validation. Here, the aim is different as the aim is not to assess the VAA afterwards (in full) but during its running (in part) to improve the VAA. While the next steps are the same for both, the main difference is that running the analysis at an earlier stage means that one desires to arrive at a certain number of scales (most often to construct a political map). This means that the results of the quasi-inductive scales might not always be useful.
2. **Clean the data**. Which cleaning steps are taken depends on the wishes of the designer, the expectation of which kinds of problems might occur, and whether too many users will be removed due to cleaning. As for which users to exclude, see Andreadis (2014) and Mendez et al. (2014) for a discussion on this topic. Besides, it is advisable to drop any unnecessary variables as well, so the resulting data-set will consist of the variables containing the items and their responses.
3. **Construct the scales**. At this step, one should make three versions of the data-set. One includes the scales as were originally intended by the designer, another the scales as a result of DSV, and another the scales as a result of the quasi-inductive process. For the original scales, one can select the appropriate variables from the data. For the DSV and quasi-inductive scales, one should follow the procedure as explained by Germann et al. (2015) and Germann and Mendez (2016) for the DSV and Wheatley (2016) for the quasi-inductive scales. An overview for the latter procedure is given in Appendix M. For DSV, the procedure is the same, with the difference that in step 4 the **aisp** procedure should not be run on the whole data-set, but only on the items of a certain scale.
4. **Calculate Unidimensionality**. For each of the three sets of scales, calculate the unidimensionality (values for Loevinger's H). This can be done using the **coefH** procedure in the **mokken** package.
5. **Calculate Reliability**. Calculate the Latent Class Reliability Coefficient for each of these scales using the *check.reliability* procedure form the **mokken** package. For this procedure, it is necessary to specify the number of latent classes expected for the scales. These can be calculated for each scale using the *poLCA* command from the **poLCA** package.
6. **Calculate the quality**. Calculate the DDI scores for each of the scales using the procedure set out in Appendix B. To calculate the quantification values required, one can use the CATPCA package provided in SPSS (Meulman et al. 2004) or the **homals** or **gifi** packages in **R** (de Leeuw and Mair 2009; Mair and de Leeuw 2017).
7. **Assess the results**. Depending on the initial goal of the designer, various options are possible. If the goal was to see whether the original scales were sufficient, one can look at the quality, unidimensionality and reliability of those scales to see if for the next time different decisions need to be made on which variables to include in which scales, or where the main problems in the current scales occurred. If the goal was to improve the

scales during the run of the VAA, one can look at the results of the DSV scales, assess whether they meet the criteria and implement them as a replacement for the original scales in the VAA. If the aim was to study the political space in general, one can look at the quasi-inductive scales to see if the political space was as expected, and which number of dimensions best explains the political space.

8. **Implement the changes**. Depending on the aim of the exercise, designers should implement any changes suggested by the assessment. This might mean updating the scales, removing or adding items, or rethinking the number of scales required in later VAAs.

References

Andreadis I (2014) Data Quality and Data Cleaning. In: Garzia D, Marschall S (eds) Matching Voters with Parties and Candidates: Voting Advice Applications in a Comparative Perspective, ECPR Press, Colchester, pp 79–92

Blasius J, Thiessen V (2012) Assessing the Quality of Survey Data. SAGE, London

de Leeuw J, Mair P (2009) Gifi Methods for Optimal Scaling in R: The Package homals. Journal of Statistical Software 31(4):1–20, DOI 10.18637/jss.v031.i04

Germann M, Mendez F (2016) Dynamic scale validation reloaded - Assessing the psychometric properties of latent measures of ideology in VAA spatial maps. Quality & Quantity 50(3):981–1007, DOI 10.1007/s11135-015-0186-0

Germann M, Mendez F, Wheatley J, Serdült U (2015) Spatial maps in voting advice applications: The case for dynamic scale validation. Acta Politica 50(2):214–238, DOI 10.1057/ap.2014.3

Linzer DA, Lewis JB (2011) poLCA: An R Package for Polytomous Variable Latent Class Analysis. Journal of Statistical Software, Articles 42(10):1–29, DOI 10.18637/jss.v042.i10

Linzer DA, Lewis JB (2013) poLCA: Polytomous Variable Latent Class Analysis. URL http://dlinzer.github.com/poLCA, r package version 1.4

Mair P, de Leeuw J (2017) Gifi: Multivariate Analysis with Optimal Scaling. URL https://CRAN.R-project.org/package=Gifi, r package version 0.3-8

Mendez F, Gemenis K, Djouvas C (2014) Methodological Challenges in the Analysis of Voting Advice Application Generated Data. In: Ninth International Workshop on Semantic and Social Media Adaptation and Personalization, IEEE Computer Society, Los Alamitos, CA, pp 142–148, DOI 10.1109/SMAP.2014.32

Meulman JJ, Heiser WJ, SPSS Inc (2004) SPSS Categories 13.0. SPSS Inc., Chicago, IL

Mokken RJ (1971) A Theory and Procedure of Scale Analysis - With Applications in Political Research. Mouton, The Hague

Molenaar IW, Sijtsma K (1988) Mokken's approach to reliability estimation extended to multicategory items. Kwantitatieve Methoden 9(28):115–126

Stochl J, Jones PB, Croudace TJ (2012) Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. BMC Medical Research Methodology 12(1):74–90, DOI 10.1186/1471-2288-12-74

Trechsel, Alexander H and Garzia, Diego and De Sio, Lorenzo (2015) euandi (General Population Survey) - User Profiles in the 2014 European Elections. DOI 10.4232/1.12246

van der Ark LA (2007) Mokken Scale Analysis in R. Journal of Statistical Software 20(11):1–19, DOI 10.18637/jss.v020.i11

van der Ark LA (2012) New Developments in Mokken Scale Analysis in R. Journal of Statistical Software 48(5):1–27, DOI 10.18637/jss.v048.i05

van der Ark LA, van der Palm DW, Sijtsma K (2011) A latent class approach to estimating test-score reliability. Applied Psychological Measurement 35(5):380–392, DOI 10.1177/0146621610392911

Wheatley J (2016) Cleavage Structures and Dimensions of Ideology in English Politics: Evidence From Voting Advice Application Data. Policy & Internet 8(4):457–477, DOI 10.1002/poi3.129