ESS Mode Change: Response Scale Report

Ranjit K. Singh (GESIS)

06/11/2023

Table of contents

Background and Objectives

- Ranjit Singh (GESIS) is studying the transformation of response scales in the context of mode switch to provide harmonised data to users in collaboration with Diana Zavala-Rojas (CST/UPF)
- The idea to collect new data for an experiment was dropped at the 26th CST meeting.
- Instead, the CST decided to look into the database of experiments on different response option formulations (MTMMs) to gather some information of the scope of this project.

Objectives

1. Gathering evidence on the impact that response scale changes during the mode change have on quality and comparability:
   - Reducing the number of response scale points (e.g., from 11-point to 5-point scales)
   - Pivoting horizontal to vertical response scales.
   - Fully labeling the new (shorter) response scales as opposed to the former partially labelled scales.
2. Exploring ways to increase comparability with ex-post harmonization techniques:
   - How variable are scale design effects across countries? The more variable the effects, the greater the need for country specific experiments / solutions.
   - What error do we incur if we simply use linear stretching?

Chosen Methods Experiments

- All three experiments involve the MTMM experiments from the supplemental ESS questionnaires.
- Each experiment involves three questions which represent one concept. This triplet structure allows us to calculate measures of internal consistency to assess the items' reliability across the method conditions.
- Experiments 1 and 3 are from round 7, experiment 2 is from round 6 of the ESS.

MTMM Experiment Designs

- Note that in the ESS MTMM experiments in rounds 1 to 7, all respondents saw a specific question version in the source questionnaire and then one randomly chosen question from a set of alternative versions later in the MTMM experiment section.
- This is important, because for experiments 1 and 3 we compare a condition from the source questionnaire with one from the MTMM variants. This allows us to compare responses not only between respondents, but also within respondents.
- For experiment 2, we only look at the four randomly varied MTMM scale versions. This means we can only compare responses between but not within respondents.

Experiment 1

`ESS Round 7`

Experiment 1 varies two response scale versions:

- **11-point, horizontal, partially labelled**
  (response scale: 00 Not at all able / — / 10 Completely able)
- **5-point, vertical, fully labelled**
  (response scale: 1 Not at all able / 2 A little able / 3 Quite able / 4 Very able / 5 Completely able)

The three items were:

- How able do you think you are to take an active role in a group involved with political issues?
- And using this card, how confident are you in your own ability to participate in politics?
- Using this card, how easy do you personally find it to take part in politics?

Experiment 2

`ESS Round 6`

Experiment 2 varies four response scale versions. All are partially labelled scales, but with varying numbers of response options in between. The endpoints were labelled "Not at all [adjective]" and "Fully [adjective]", where [adjective] was replaced with "interested", "absorbed" or "enthusiastic", depending on the item (see below).

- **11-point**
- **7-point**
- **5-point**
- **3-point**

The three items were:

Please use CARD 31 for the next three questions. How much of the time would you generally say you are...

- Interested in what you are?
- Absorbed in what you are doing?
- Enthusiastic on what you are doing?

Experiment 3

`ESS Round 7`

Experiment 1 varies two response scale versions:

- 11-point, **horizontal**, partially labelled
  (response scale: 00 Extremely unimportant / — / 10 Extremely important)
- 11-point, **vertical**, partially labelled
  (response scale: 00 Extremely unimportant / — / 10 Extremely important)

The three items were:

- Please tell me how important you think being able to speak [country's official language(s)] should be in deciding whether someone born, brought up and living outside [country] should be able to come and live here.
- And how important do you think being white ...
- Now, how important do you think being committed to the way of life in [country] ...

Experimen order

Please note that Experiment 1, 2, and 3 were numbered according to earlier presentations on our analysis plans. To make this report consistent with previous presentations we kept the order as is. However, in the results section, we will move through the experiments in whatever order makes the most important results easiest to understand. The result section also has brief reminders about the respective experiment in the rightmost column of this report.

Results

The following analyses tackle two broad issues in response scale comparability:

1. Changes in reliability. Response scale changes may lead to changes in measurement precision.
2. Changes in scaling, that is the numerical mapping of respondent attitudes onto response scales. Different response scale designs may result in different response distributions and thus in a break in the ESS time series.

Reliability differences

Scale point reduction (partially labelled)

In this analysis we look at the impact of the number of response options on reliability. Note that experiment 2 featured partially labelled scales.

Experiment 2: *11-point*, *7-point*, *5-point*, and *3-point* scales; all partially labelled.

Reliability estimates are depicted by country and response scale condition. Values are Cronbach's Alpha as a measure of internal consistency.

Each dot in the plot represents Cronbach's Alpha for a specific country in one of the response scale conditions. The shape behind the points is a violin plot, illustrating the density of points for each condition (i.e., where most countries lie). The orange line in front is a linear trend across the conditions.

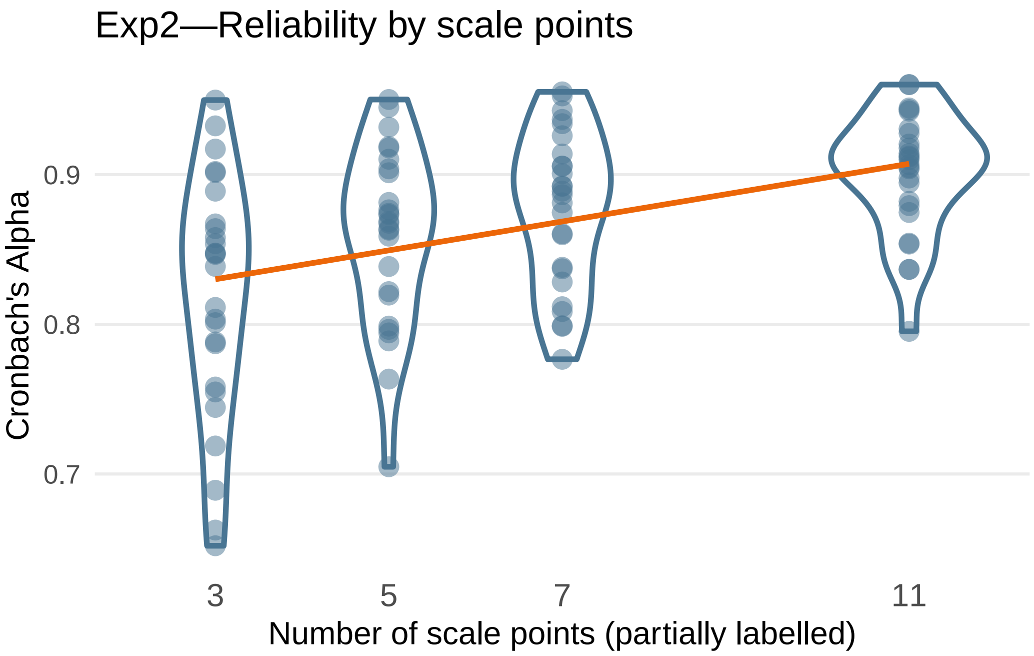


Table 1: Exp2—Reliability by scale points and country

country	A (11-point)	B (7-point)	C (5-point)	D (3-point)
AL	0.89	0.81	0.76	0.66
BE	0.92	0.89	0.86	0.85
BG	0.94	0.94	0.92	0.90
CH	0.84	0.81	0.82	0.79
CY	0.96	0.96	0.93	0.90
CZ	0.91	0.89	0.87	0.85
DE	0.80	0.80	0.79	0.65
DK	0.85	0.84	0.87	0.79
EE	0.90	0.87	0.86	0.84
ES	0.93	0.93	0.90	0.92
FI	0.84	0.80	0.80	0.72
FR	0.88	0.83	0.82	0.75
GB	0.91	0.90	0.86	0.80
HU	0.96	0.96	0.92	0.88
IE	0.94	0.94	0.92	0.89
IL	0.91	0.89	0.87	0.85
IS	0.85	0.78	0.70	0.69
IT	0.92	0.86	0.95	0.95
LT	0.94	0.92	0.90	0.85
NL	0.93	0.89	0.89	0.80
NO	0.87	0.86	0.79	0.74
PL	0.91	0.91	0.88	0.86
PT	0.96	0.95	0.94	0.93
RU	0.91	0.89	0.88	0.86
SE	0.88	0.84	0.80	0.76
SI	0.94	0.93	0.91	0.85
SK	0.90	0.91	0.84	0.81
UA	0.93	0.91	0.90	0.87
XK	0.90	0.88	0.87	0.80

Table 2: Exp2—Reliabilities by scale points averaged across countries

3-point	5-point	7-point	11-point
0.82	0.86	0.88	0.90

Interpretation

We see a trend with a greater number of response options leading to higher reliability. This is consistent with the idea, that more response options can capture more finely grained information. However, note that all conditions in experiment 2 feature partially labelled scales.

11-point partially versus 5-point fully labelled scales

Next, we look at reliability differences between 11-point partially labelled versus 5-point fully labelled scales.

Experiment 1: *11-point, horizontal, partially labelled* versus *5-point, vertical, fully labelled*

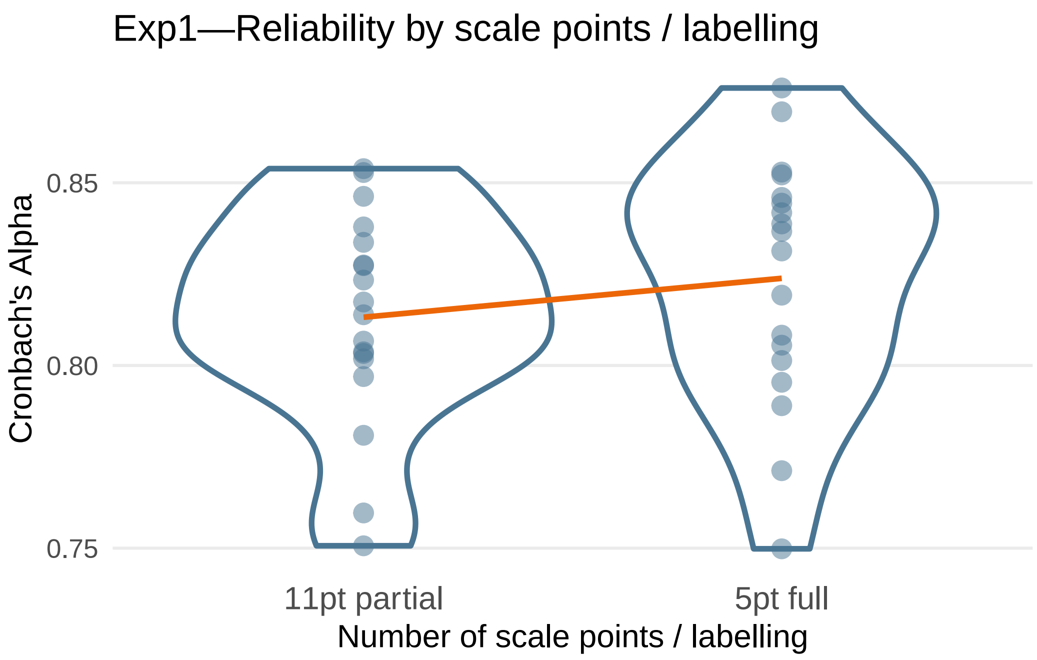


Table 3: Exp1—Reliabilities by scale points / labelling and country

country	main11pt	mtmm5pt
AT	0.82	0.81
BE	0.80	0.80
CH	0.78	0.79
CZ	0.81	0.83
DE	0.76	0.77
DK	0.81	0.84
EE	0.84	0.82
ES	0.82	0.84
FI	0.80	0.80
FR	0.75	0.75
GB	0.83	0.85
HU	0.78	0.76
IE	0.85	0.88
IL	0.85	0.87
LT	0.93	0.92
NL	0.86	0.82
NO	0.83	0.84
PL	0.80	0.84
PT	0.83	0.85
SE	0.85	0.85
SI	0.80	0.81

Table 4: Exp1—Reliabilities by scale points / labelling averaged across countries

main11pt	mtmm5pt
0.82	0.83

Interpretation

Unlike with experiment 2, we see no difference in reliability between a fully labelled 5-point response scale and a partially labelled 11-point response scale. It is possible, that a fully labelled scale increases ease of response and thus counteracts the information loss.

Horizontal versus vertical scales

Lastly, we explore reliability differences between horizontal and vertical response scales.

Experiment 3: *horizontal* versus *vertical* scale; both 11-point, partially labelled

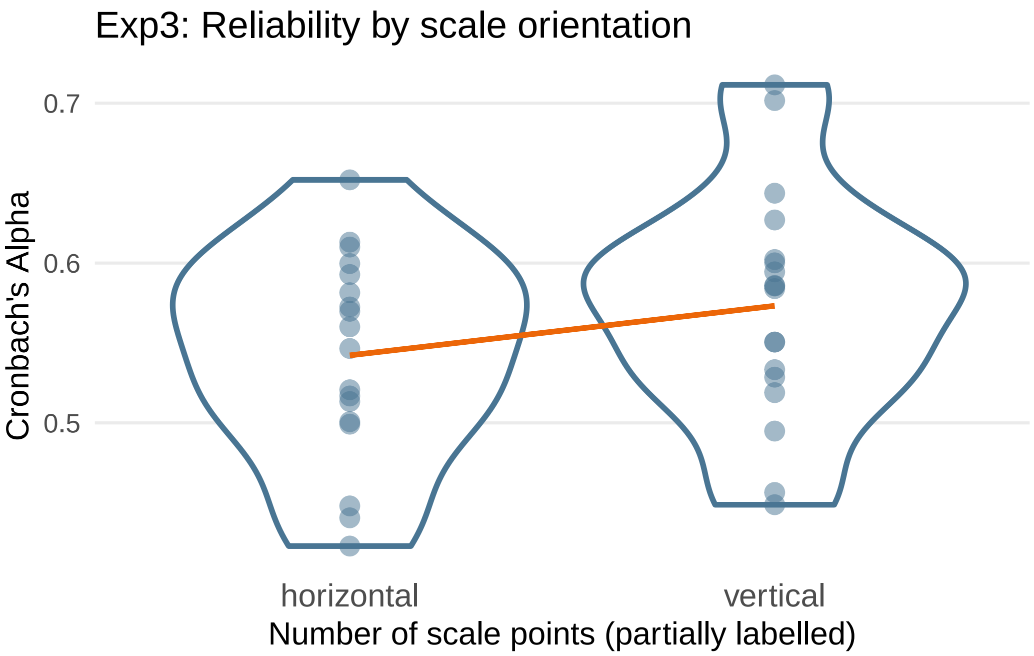


Table 5: Exp3—Reliabilities by scale points / labelling and country

country	horizontal	vertical
AT	0.57	0.52
BE	0.51	0.55
CH	0.42	0.49
CZ	0.60	0.59
DE	0.45	0.46
DK	0.61	0.70
EE	0.50	0.53
ES	0.57	0.59
FI	0.58	0.63
FR	0.55	0.60
GB	0.56	0.55
HU	0.64	0.78
IE	0.44	0.58
IL	0.52	0.45
LT	0.68	0.76
NL	0.46	0.46
NO	0.65	0.71
PL	0.61	0.64
PT	0.50	0.60
SE	0.59	0.59
SI	0.52	0.53

Table 6: Exp3—Reliabilities by scale points / labelling averaged across countries

mainhorizontal	mtmmvertical
0.55	0.59

Interpretation

Reliabilities are very similar on average between the horizontal (Alpha = .55) and vertical (Alpha = .59) 11-point response scales.

Reliability summary

The plot below summarises the results on reliability from experiment 2, 1, and 3. Each row in the plot refers to a comparison of two design features and their impact on reliability. Specifically, each dot represents the difference in Cronbach's Alpha in one country. For example, `11pt partial - 5pt partial` compares the 11-point partially labelled scale with the 5-point partially labelled scale. As the title suggests, the values are the difference between the 11pt reliabilities and the 5pt reliabilities, which means that positive values imply greater reliability of the 11-point partially labelled scale.

To make the plot easier to understand, two clarification features were added. First, a horizontal, blue line which represents the average reliability difference between the two question design features across all countries. The delta Alpha value on each row is the corresponding numerical difference value. Second, the curve in each row is a density plot. It illustrates where most of the points (i.e., intra-country differences) are located. The density curve thus shows the position and spread of differences across countries.

(In the first plot, for example, it means that dots to the right of zero are countries where 11pt partial scales led to higher reliability than 5pt fully labeled. Left of zero, we observe the opposite.)

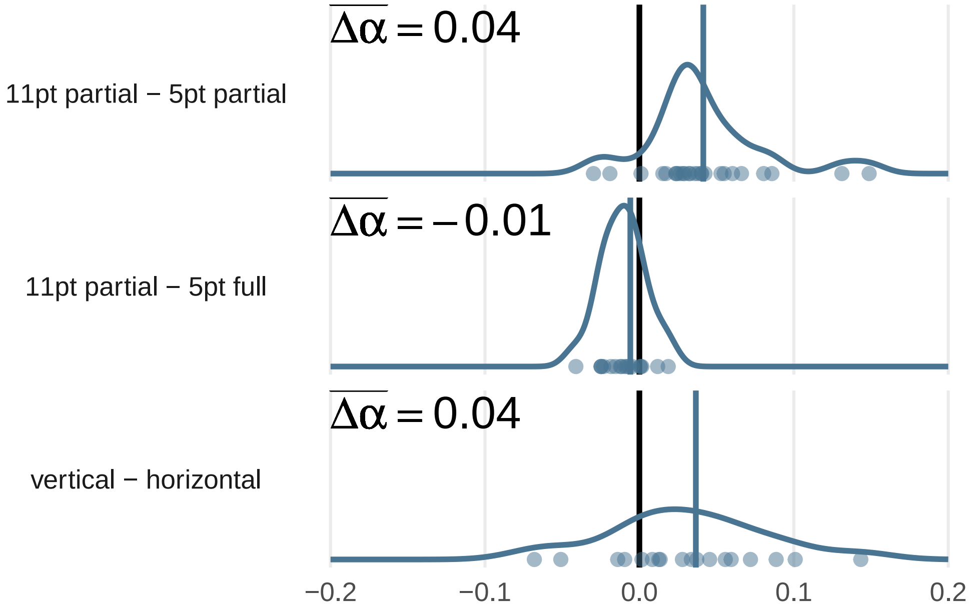


Table 7: Reliability differences between selected conditions

Country	Experiment	Comparison	Delta Alpha
AL	Exp. 2	11pt partial - 5pt partial	0.13
BE	Exp. 2	11pt partial - 5pt partial	0.05

Table 7: Reliability differences between selected conditions *(continued)*

Country	Experiment	Comparison	Delta Alpha
BG	Exp. 2	11pt partial - 5pt partial	0.02
CH	Exp. 2	11pt partial - 5pt partial	0.02
CY	Exp. 2	11pt partial - 5pt partial	0.03
CZ	Exp. 2	11pt partial - 5pt partial	0.04
DE	Exp. 2	11pt partial - 5pt partial	0.00
DK	Exp. 2	11pt partial - 5pt partial	-0.02
EE	Exp. 2	11pt partial - 5pt partial	0.04
ES	Exp. 2	11pt partial - 5pt partial	0.02
FI	Exp. 2	11pt partial - 5pt partial	0.04
FR	Exp. 2	11pt partial - 5pt partial	0.06
GB	Exp. 2	11pt partial - 5pt partial	0.05
HU	Exp. 2	11pt partial - 5pt partial	0.03
IE	Exp. 2	11pt partial - 5pt partial	0.03
IL	Exp. 2	11pt partial - 5pt partial	0.04
IS	Exp. 2	11pt partial - 5pt partial	0.15
IT	Exp. 2	11pt partial - 5pt partial	-0.03
LT	Exp. 2	11pt partial - 5pt partial	0.04
NL	Exp. 2	11pt partial - 5pt partial	0.04
NO	Exp. 2	11pt partial - 5pt partial	0.09
PL	Exp. 2	11pt partial - 5pt partial	0.04
PT	Exp. 2	11pt partial - 5pt partial	0.02
RU	Exp. 2	11pt partial - 5pt partial	0.03
SE	Exp. 2	11pt partial - 5pt partial	0.08
SI	Exp. 2	11pt partial - 5pt partial	0.03
SK	Exp. 2	11pt partial - 5pt partial	0.07
UA	Exp. 2	11pt partial - 5pt partial	0.03
XK	Exp. 2	11pt partial - 5pt partial	0.02
AT	Exp. 1	11pt partial - 5pt full	0.01
BE	Exp. 1	11pt partial - 5pt full	0.00
CH	Exp. 1	11pt partial - 5pt full	-0.01
CZ	Exp. 1	11pt partial - 5pt full	-0.02
DE	Exp. 1	11pt partial - 5pt full	-0.01
DK	Exp. 1	11pt partial - 5pt full	-0.02
EE	Exp. 1	11pt partial - 5pt full	0.02
ES	Exp. 1	11pt partial - 5pt full	-0.02
FI	Exp. 1	11pt partial - 5pt full	0.00
FR	Exp. 1	11pt partial - 5pt full	0.00
GB	Exp. 1	11pt partial - 5pt full	-0.02
HU	Exp. 1	11pt partial - 5pt full	0.02
IE	Exp. 1	11pt partial - 5pt full	-0.02
IL	Exp. 1	11pt partial - 5pt full	-0.02
LT	Exp. 1	11pt partial - 5pt full	0.01
NL	Exp. 1	11pt partial - 5pt full	0.04
NO	Exp. 1	11pt partial - 5pt full	-0.01
PL	Exp. 1	11pt partial - 5pt full	-0.04
PT	Exp. 1	11pt partial - 5pt full	-0.01
SE	Exp. 1	11pt partial - 5pt full	-0.01
SI	Exp. 1	11pt partial - 5pt full	-0.00
AT	Exp. 3	vertical - horizontal	-0.05
BE	Exp. 3	vertical - horizontal	0.04
CH	Exp. 3	vertical - horizontal	0.07
CZ	Exp. 3	vertical - horizontal	-0.01

Table 7: Reliability differences between selected conditions *(continued)*

Country	Experiment	Comparison	Delta Alpha
DE	Exp. 3	vertical - horizontal	0.01
DK	Exp. 3	vertical - horizontal	0.09
EE	Exp. 3	vertical - horizontal	0.03
ES	Exp. 3	vertical - horizontal	0.01
FI	Exp. 3	vertical - horizontal	0.05
FR	Exp. 3	vertical - horizontal	0.06
GB	Exp. 3	vertical - horizontal	-0.01
HU	Exp. 3	vertical - horizontal	0.13
IE	Exp. 3	vertical - horizontal	0.14
IL	Exp. 3	vertical - horizontal	-0.07
LT	Exp. 3	vertical - horizontal	0.08
NL	Exp. 3	vertical - horizontal	-0.00
NO	Exp. 3	vertical - horizontal	0.06
PL	Exp. 3	vertical - horizontal	0.03
PT	Exp. 3	vertical - horizontal	0.10
SE	Exp. 3	vertical - horizontal	0.00
SI	Exp. 3	vertical - horizontal	0.01

Interpretation

On average, the average reliability differences between the scale design conditions are very minor in all three experiments (*delta alpha* <= 0.04).
Let us consider the potential impact on data analysis.Reductions in reliability of a measurement reduce all empirical correlations with that variable through *attenuation*. However, this effect is rather small with such minor changes in reliability. Imagine two variables measured with reliabilities of *alpha* = 0.8 each. If we now reduce both their reliabilities by *alpha* = 0.04, the resulting reductions in empirical correlations are negligible. Assuming a true population correlation of $r = .5$, empirical correlations would only drop by *delta* $r = .02$ (from $r = .40$ to $r = .38$). If the true population correlation was $r = .3$, empirical correlations would drop even less by $r = 0.01$ (from $r = .24$ to $r = .23$).

> Charles, E. P. (2005). The Correction for Attenuation Due to Measurement Error: Clarifying Concepts and Creating Confidence Sets. *Psychological Methods*, *10*(2), 206–226. https://doi.org/10.1037/1082-989X.10.2.206

Measurement units

In this section we explore the impact on response distributions and thus the potential for breaking the ESS time-series.

Harmonizing response scales

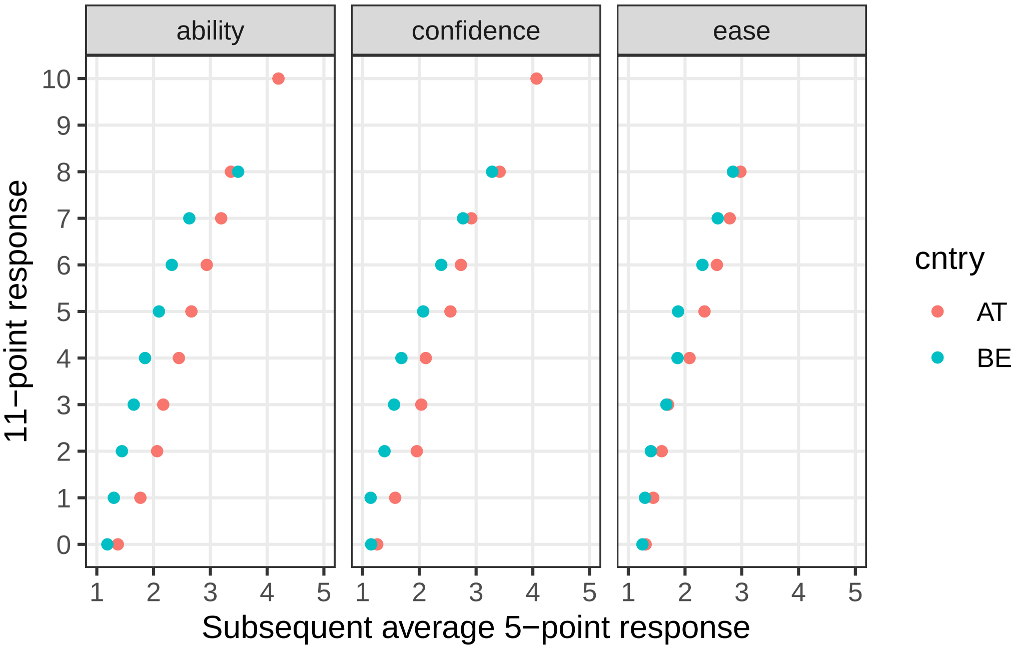
Here we explore how the 5-point fully labelled and 11-point partially labelled scales in experiment 1 relate to each other across different countries. In experiment 1, respondents answered the 11-point scale in the main ESS questionnaire and the 5-point scale later on in the supplemental questionnaire. Thus, we can divide the sample into people who chose different responses in the 11-point scale and then explore which response they subsequently chose on the 5-point scale. The plot shows this separately for each country. A point can be

Experiment 1: *11-point, horizontal, partially labelled* versus *5-point, vertical, fully labelled* interpreted as “respondents who chose y on the 11-point scale chose x on the 5-point scale” in that specific country.

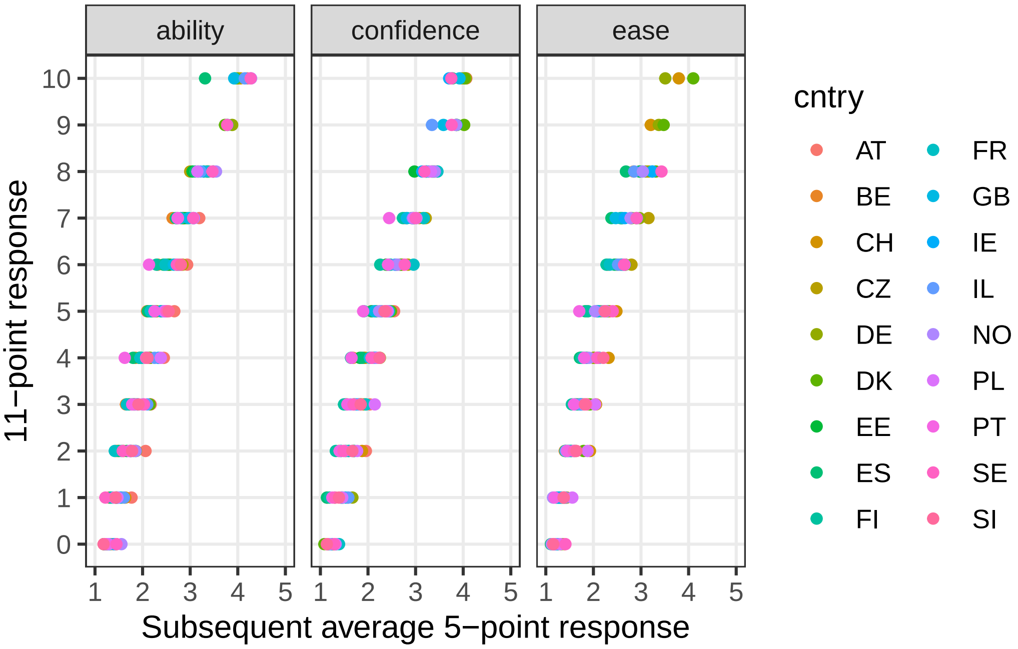
Illustrative expample: Belgium vs. Austria

Here we have chosen an illustrative example: A comparison between Belgium and Austria. If we focus on the left facet “ability”, we see that respondents from Belgium consistently chose lower responses on the five-point scale than respondents from Austria, even if they had chosen the very same response in the 11-point scale.

Let us take a specific example. If we look at “5” on the y-axis, then we have respondents who chose a “5” on the 11-point scale. Here we clearly see that these respondents then chose a lower response on the 5-point scale in Belgium than they did in Austria.



Plot for all countries



Interpretation

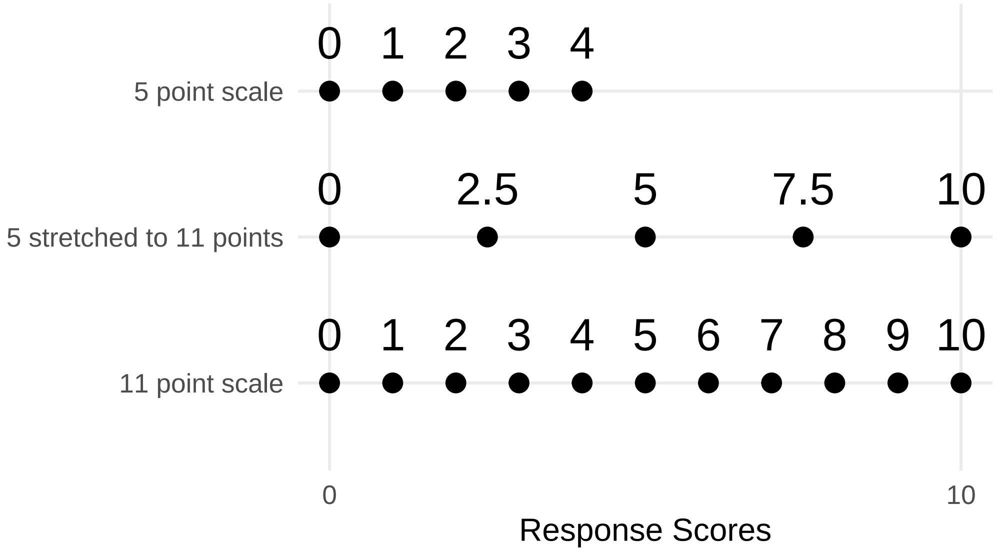
The plot shows that respondents who chose a specific response option in the 11-point scale chose markedly different responses on the 5-point scale (on average) in different countries. **In other words, the difference in response behavior caused by the two response scales is not stable across countries.**
A pertinent question is whether we can harmonize response data across such a change in response scale. The results in the previous section mean that **we cannot easily generalize response effects across countries**. Consequently, we would require parallel runs in all countries to apply empirical harmonization methods such as observed score equating.

Linear Stretching

Experiment 1: *11-point, horizontal, partially labelled* versus *5-point, vertical, fully labelled*

Aside from more sophisticated harmonization methods, such as observed score equating, there is the so called **Linear Stretching** method. The method only takes the number of response options into account. It is usually not recommended, since it ignores issues such as differences in item difficulty, scale label wording, or scale orientation.

The linear stretching formula simply sets the extreme values as equal and all other response options in between as equidistant. The schematic diagram below illustrates this process for stretching a 5-point response scale towards an 11-point response scale.



The following analysis illustrates the methodological cost of linear stretching with data from experiment 1. Here, we used linear stretching to harmonize the 5-point and 11-point responses. Then we calculated the standardized mean difference as Cohen's d. Due to the experimental design, we would expect a mean difference of zero if harmonization resulted in perfect comparability.

Mean bias after linear stretching

Points represent the remaining standardized mean differences (Cohen's d) between the two scale designs even after linear stretching has been applied. Again, we have added two clarification features. First, a blue, vertical line to denote the position of the average mean difference. Second, a density curve, which makes the position and spread of the points easier to see.

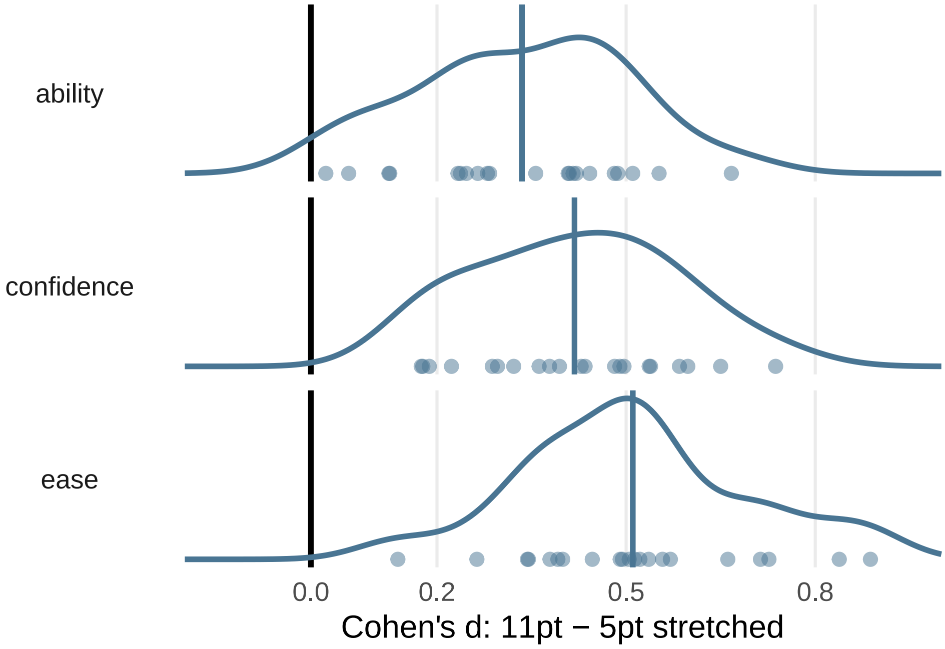


Table 8: Mean bias by country and item

Country	Item	Mean bias (Cohen's d)
AT	ability	0.06
BE	ability	0.48
CH	ability	0.44
CZ	ability	0.25
DE	ability	0.49
DK	ability	0.42
EE	ability	0.28
ES	ability	0.41
FI	ability	0.67
FR	ability	0.51
GB	ability	0.28
HU	ability	0.02
IE	ability	0.23
IL	ability	0.12
LT	ability	0.26
NL	ability	0.55
NO	ability	0.41
PL	ability	0.24
PT	ability	0.36
SE	ability	0.42
SI	ability	0.13
AT	confidence	0.30
BE	confidence	0.58
CH	confidence	0.49
CZ	confidence	0.22
DE	confidence	0.48
DK	confidence	0.43
EE	confidence	0.36
ES	confidence	0.50
FI	confidence	0.65
FR	confidence	0.32
GB	confidence	0.43

Table 8: Mean bias by country and item *(continued)*

Country	Item	Mean bias (Cohen's d)
HU	confidence	0.38
IE	confidence	0.39
IL	confidence	0.19
LT	confidence	0.29
NL	confidence	0.74
NO	confidence	0.60
PL	confidence	0.18
PT	confidence	0.54
SE	confidence	0.54
SI	confidence	0.18
AT	ease	0.40
BE	ease	0.66
CH	ease	0.49
CZ	ease	0.38
DE	ease	0.56
DK	ease	0.54
EE	ease	0.45
ES	ease	0.57
FI	ease	0.84
FR	ease	0.71
GB	ease	0.49
HU	ease	0.51
IE	ease	0.39
IL	ease	0.26
LT	ease	0.35
NL	ease	0.89
NO	ease	0.73
PL	ease	0.14
PT	ease	0.52
SE	ease	0.50
SI	ease	0.34

Interpretation

We see that linear stretching results in substantial bias in mean estimation. Cohen's d effect sizes are conventionally interpreted as small > 0.2, medium > 0.5, and high > 0.8. We see a large group of countries in the area of medium effects. If unresolved, this issue results in breaks in the time series of the same magnitude. Also note again how different the method effect plays out in different countries.

Harmonizing horizontal and vertical scales

Lastly, we also explored distribution differences between horizontal and vertical response scales. Due to the experimental design, we would expect no mean difference between responses if both horizontal and vertical scale measure comparably. The standardized mean differences between the two scales could also be seen as the magnitude of a break in the time series after a scale change that is attributable to that design choice.

Experiment 3: *horizontal* versus *vertical* scale; both 11-point, partially labelled

Plot

Points represent the mean differences (Cohen's d) between the two scale designs. Again, we have added two clarification features. First, a blue, vertical line to denote the position of the average mean difference. Second, a density curve, which makes the position and spread of the points easier to see.

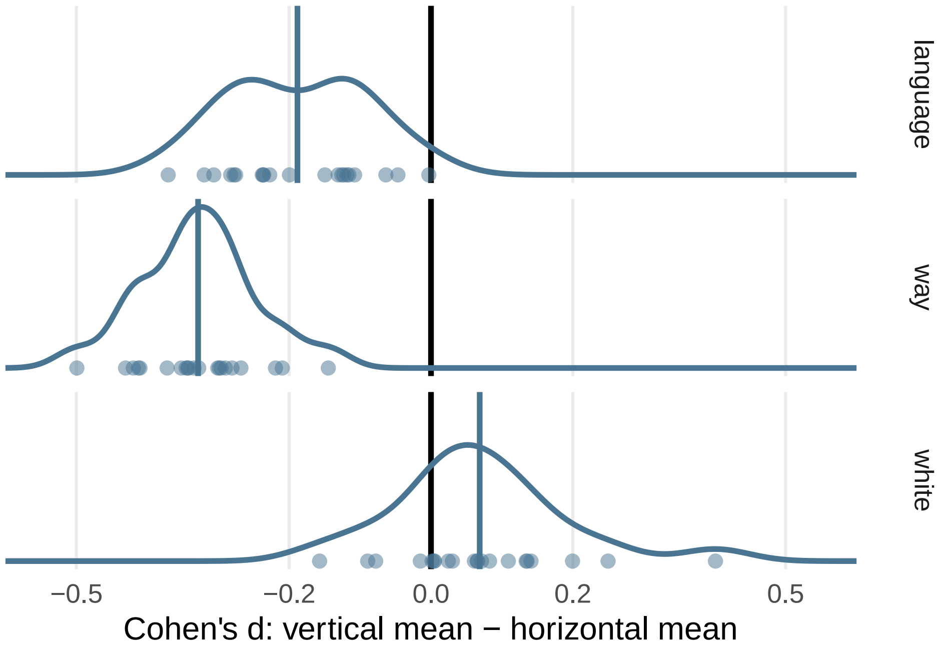


Table 9: Mean bias by country and item

Country	Item	Mean bias (Cohen's d)
AT	language	-0.24
BE	language	-0.23
CH	language	-0.13
CZ	language	-0.24
DE	language	-0.11
DK	language	-0.20
EE	language	-0.00
ES	language	-0.12
FI	language	-0.32
FR	language	-0.31
GB	language	-0.28
HU	language	-0.28
IE	language	-0.28
IL	language	-0.06
LT	language	-0.37
NL	language	-0.12
NO	language	-0.15
PL	language	-0.12
PT	language	-0.13
SE	language	-0.24
SI	language	-0.05
AT	way	-0.29
BE	way	-0.43
CH	way	-0.30
CZ	way	-0.33

Table 9: Mean bias by country and item *(continued)*

Country	Item	Mean bias (Cohen's d)
DE	way	-0.33
DK	way	-0.14
EE	way	-0.41
ES	way	-0.30
FI	way	-0.50
FR	way	-0.27
GB	way	-0.37
HU	way	-0.34
IE	way	-0.35
IL	way	-0.28
LT	way	-0.41
NL	way	-0.42
NO	way	-0.22
PL	way	-0.21
PT	way	-0.35
SE	way	-0.30
SI	way	-0.34
AT	white	0.02
BE	white	0.14
CH	white	0.03
CZ	white	-0.02
DE	white	0.08
DK	white	0.07
EE	white	-0.08
ES	white	0.07
FI	white	0.00
FR	white	0.11
GB	white	0.13
HU	white	0.01
IE	white	0.20
IL	white	-0.09
LT	white	-0.16
NL	white	0.40
NO	white	0.14
PL	white	0.06
PT	white	0.25
SE	white	0.00
SI	white	0.07

Interpretation

We see slight shifts in mean scores induced by different scale orientations. However, please note that these effects are rather small. In fact, most effects are small or even less than small ($|d| < 0.2$). Again, we find pronounced differences between the different countries. Also note that the third item "white" behaves differently from the rest. This is because the item's content is rather extreme and thus the response distribution is very different from the other two.

Executive summary

The planned mode change in the ESS will likely involve changes in response scale design characteristics, such as scale orientation, the number of response options, or the response labels. To assess the impact of such response scale changes, we chose pertinent experiments from among the ESS MTMM experiments.

Main findings

1. **Reliability differences were insubstantial in all three experiments.** Even where some differences occurred, they would not lead to substantive changes in analysis results. The response scale effects were also far smaller than the already present cross-country differences.
2. **Harmonizing the time series after switching from an 11-point to a 5-point response scale** would be challenging for the following reasons:

- We found that respondents from **different countries reacted quite differently** to such response scale changes. This means that we cannot generalize scale effect findings from one country to all others.
- This variability in scale effects across countries also makes it hard to apply robust harmonization procedures such as **Observed Score Equating** to maintain the time series across such a response scale change. Observed Score Equating would consequently require **parallel runs in all countries**.
- **Linear stretching** is a frequently used harmonization approach, but as our analyeis show it is **inadvisable**. Even after applying linear stretching the 5-point and 11-point scales, substantial mean bias, and thus breaks in the time series, persisted.

3. Lastly, there were concerns that **horizontal versus vertical response scale orientations** might reduce comparability. However, we **only found comparably small shifts in response distributions** due to scale orientation. If we apply the conventional effect sizes for Cohen's *d*, then scale orientation effects were at worst small or even less than small.

> ⚠ **Limitations**
>
> - These results are based on the three experiments we selected from the MTMM experiment pool. Different questions or topics may lead to different effects.
> - Note that the majority of ESS MTMM Experiments were conducted face-to-face. This means that response scale effects may turn out differently when combined with an actual mode change. This is especially true for the scale orientation of 11-point scales, where respondents on smartphone screens are the most likely bottleneck.

Recommendations for further experiments

1. The overarching finding was that all **response scale effects varied strongly between countries**. This means we cannot generalize findings from some countries to

all other countries, when assessing or mitigating comparability issues. Consequently, **experiments in several, or perhaps even all countries are advisable**. If resources permit, then **parallel runs** would be optimal.

2. Findings should ideally be corroborated with a greater **variety of concepts and question wordings**.
3. If the **ESS decides to shorten response scales** to 5-point or 7-point scales, we strongly recommend **parallel runs** in most or ideally all countries. This would allow us to harmonize the time series more robustly afterwards.
4. If the **ESS decides against shortening the response scale**, we recommend experiments that specifically **validate the performance of 11-point scales** on paper, on larger computer screens, and on smartphone screens (and in several or ideally all countries).