Quality & Quantity
https://doi.org/10.1007/s11135-025-02436-9

Between two minds: the influence of numerical labels on survey responses in a cross-national study

Gianmaria Bottoni[1] · **Eva Aizpurua**[2]

Received: 14 January 2025 / Accepted: 1 October 2025

Abstract
Previous research highlights the impact of nonverbal cues and visual layouts on respondents' interpretations and answers to survey questions (Schwarz et al. in Pub Opin Q 55(4):570–582, 1991; Tourangeau et al. in Pub Opin Q 71(1):91–112, 2007; Toepoel and Dillman in Soc Sci Comput Rev 29(2):193–207, 2011). In this study, we investigate the effect of numerical values on responses to five questions measured using 11-point bipolar scales. These questions asked for evaluations of the general impact of climate change and its effects on different groups. Respondents were randomly assigned to scales ranging from 0 to 10, or from −5 to +5. The scales were presented horizontally and had identical verbal endpoints anchored as "extremely negative" and "extremely positive". The experiment was embedded in a cross-national face-to-face survey conducted in Hungary, Portugal, and Great Britain. Contrary to prior research (Schwarz et al. in Pub Opin Q 55(4):570–582, 1991; O'Muircheartaigh et al. J Off Stat 11:295–308, 1995; Tourangeau et al. in Pub Opin Q 71(1):91–112, 2007), our analyses show an opposite trend: the means tend to be consistently and significantly lower for the −5/+5 scale. The discrepancy between our findings and previous studies suggests that the interpretations of the 0/10 and −5/+5 scales could be item specific. The implications of these findings and how they fit with previous work are discussed.

Keywords Survey methodology · Questionnaire design · Visual cues · Response scales · Numerical labels

✉ Gianmaria Bottoni
ginbottoni@gmail.com

Eva Aizpurua
eva.aizpu@gmail.com

[1] City St George´s, University of London, Northampton Square, London EC1V 0HB, UK

[2] Center for Social and Behavioral Research, University of Northern Iowa, 2304 College Street, Cedar Falls 50614-0402, USA

Published online: 31 October 2025

1 Introduction

Rating scales with verbal labels, or numeric rating scales with labelled endpoints, are arguably the most widely used measurement instruments in the social sciences. These scales comprise a set of response options that range from one extreme to another, thereby forming a measurement continuum. The assumption that attitudes can be measured (with varying degrees of measurement error) on a single continuum is widely accepted (Thurstone 1928). Responding to questionnaires involves considerable cognitive effort. First, respondents must interpret the meaning of the question stem, along with any accompanying instructions. Second, they need to recall the necessary information to provide an optimal answer. This information must then be integrated to form a coherent judgment. Finally, respondents must interpret the meaning of the response options to indicate their position on the attribute in question (Tourangeau et al. 2000). This article focuses on this last component – the selection and recording of the response option. This process is related to the effect that a particular scale format can have on response behaviour.

Throughout this complex response process, several factors may influence outcomes. These factors include the wording of the question, accompanying instructions, the positioning of the question within the survey, the number of scale categories, whether response options are partially or fully labelled, the terms used to label the endpoints, and more (Wildt and Mazis 1978; Krosnick and Fabrigar 1997; Scherpenzeel and Saris 1997; Weng 2004; Saris and Gallhofer 2007; Menold 2020).

Couper et al. (2004) distinguish between task elements and style elements. Task elements refer to features that are essential for completing a survey, such as question wording, instructions, and response options. Style elements, in contrast, pertain to characteristics that determine the overall "look and feel" of the survey. These elements are typically, though not necessarily, visual features (e.g., font style, logos, survey organisation name).

While this distinction holds heuristic validity, it may be misleading, since all elements influence task completion. For example, using a clear font style or designing a mobile-first layout that minimises scrolling can affect survey completion. Similarly, the inclusion of a recognisable logo may encourage respondents to engage in optimal behaviour, especially when associated with highly reputed institutions.

Respondents tend to employ five basic heuristics when interpreting the visual characteristics of survey questions (Tourangeau et al. 2004). These heuristics are: (1) Middle means typical: the middle option is seen as the most typical; (2) Left and top means first: the leftmost or top option is interpreted as the first in conceptual terms; (3) Near means related: options placed closely together are seen as conceptually related; (4) Up means good: the top option is viewed as the most desirable; (5) Like means close: visually similar options are deemed conceptually related.

In addition, Tourangeau et al. (2007) proposed a hierarchy of features that respondents attend to, with verbal labels taking precedence over numerical ones, and numerical labels taking precedence over purely visual cues (see also Toepoel and Dillman 2011).

The purpose of this paper is to investigate whether, in the context of 11-point rating scales administered both aurally and visually, using negative numerical labels (ranging from −5 to +5) as opposed to positive numerical labels (0–10) affects the distributions and interpretation of survey questions. The paper implements an experiment that isolates numeric labels as the experimental variable and provides evidence of their impact on distributions.

By adopting a cross-national perspective, the study contributes to the methodological literature which has often focused on wording or the number of points rather than the value assigned to each point, and has been primarily conducted in the US.

2 Background

Amongst the factors influencing response behaviour, one that is often neglected in research practice is the numerical labels used to anchor the scale categories in rating scales.

Krosnick and Fabrigar (1997) stated that the numbers in numeric rating scales carry no inherent meaning other than suggesting equal distance between answer categories. Some features can be underestimated when designing a survey, and some characteristics can be considered secondary elements. However, these elements can inadvertently introduce additional meaning to the questions. Schwarz et al. (1991) demonstrated that respondents use the numeric values provided in rating scales as cues for interpreting the meaning of the labels (see also Schwarz and Hippler 1995). In particular, numeric values can alter the meaning of endpoint labels, resulting in systematic variations in responding as well as different interpretations of the scale, which in turn affects variable distributions. Importantly, this finding holds even when numeric values do not include negative numbers (Schwarz et al. 1998; Evans 2019; Aybar et al. 2024). This suggests that variations in response behaviour do not solely reflect respondents' hesitation to assign negative values when assessing an item. Instead, respondents rely on all features of scales to determine the meaning they assign to a particular item.

Schwarz et al. (1991) demonstrated this effect with the item "How successful have you been in life, so far?". A scale with numeric values ranging from 0 to 10 was not equivalent to the same scale ranging from −5 to +5, even though both had endpoints labelled "not at all successful" (0 and −5) and "extremely successful" (10 and +5). When the scale ranged from 0 to 10, 34% of respondents chose values between 0 and 5. In contrast, only 13% of respondents selected formally equivalent values between −5 and 0 with the scale ranging from −5 to +5. The authors argue that on a 0 to 10 scale, respondents interpret 0 as the absence of the characteristic in question (i.e., lacking any success). However, on a − 5 to +5 scale, -5 is not only seen as the absence of success, but as an indication of failure. In other words, the use of different numeric values (0 to 10 versus −5 to +5) shifts the perception from a unipolar scale (0/10 – not successful to successful) to a bipolar scale (− 5/+5 – unsuccessful to successful).

O'Muircheartaigh et al. (1995) replicated Schwarz et al.'s (1991) experiment using a different item, arguing that the lower verbal anchor ("not at all successful") in Schwarz et al.'s (1991) study could be ambiguous. This ambiguity could lead respondents to interpret the scale as either unipolar (absence of success) or bipolar (being a failure). To address this, they selected the item "How entertaining do you think the adverts on television are, compared to the programmes?", and used explicit bipolar verbal anchors: "much more" and "much less". Their study confirmed the results of Schwarz et al. (1991), indicating that different numeric labels can produce shifts in response distributions. In particular, when negative numeric values were assigned to the low end of the scale, responses tended to gravitate towards the high end of the scale, thereby increasing the mean scores.

Further evidence supporting this finding was provided by Tourangeau et al. (2007), who showed the impact of negative numeric lsabels as opposed to positive numeric labels. In

two experiments, they demonstrated that the effects of negative numeric labels on response distributions persist regardless of whether the scale is unipolar or bipolar.

Several studies (O'Muircheartaigh 1995; Schaeffer and Presser 2003; Kennedy 2008; Krebs 2012; Menold 2021; Höhne et al. 2022; Stefkovics and Höhne 2024) have shown that responses to bipolar scales are typically skewed toward the positive end of the scale compared to unipolar scales. This means that when respondents are presented with bipolar scales rather than unipolar scales, they tend to select response options clustered toward the positive end of the scale.

2.1 Current study

Building on Schwarz et al.'s seminal work showing that, although numeric values are chosen for administrative purposes, they shape how respondents assign meaning to questions, we test whether two formally equivalent 11-point rating scales, one labelled −5 to +5 and one 0 to 10, produce different response distributions.

Classicexperiments (Schwarz et al. 1991; Schwarz and Hippler 1995; O'Muircheartaigh et al. 1995; Tourangeau et al. 2007)consistently demonstrate that respondents are less likely to select options at the lower end of a scale when these options carry negative numerical values. As a result, scales labelled from −5 to +5 are likely toyield higher means compared to formally equivalent scales labelled from 0 to 10. A more recent experiment by Pew Research showed that renumbering a left-right ideology item from 0 to 6 to 1–7 reduced midpoint selection and shifted choices toward the lower end, even though wording and scale length were identical (Evans 2019).

In addition, considering the positivity bias (Tourangeau et al. 2000) – which suggests that respondents lean towards positive answers, avoiding the negative end of a rating scale – we anticipate that scales labelled from −5 to +5 will produce more negatively skewed distributions than ones labelled from 0 to 10.

Building on Krosnick's satisficing principles (Krosnick 1991), it is expected that respondents will more frequently select middleoptions on the −5 to +5scale. Because negativenumbers are more difficult toendorse due to positivity bias, respondents are likely to choose the first "acceptable" non-negative number, producing additional clustering towards the upper end of the scale. While removing verbal labels from rating scales can be attractive to reduce the potential interplay between both verbal and numeric labels, recent studies caution that doing so may increase cognitive load (Gummer and Kunz 2021). Given that verbal and numeric cues routinely coexist in practice, examining their interaction remains essential. To summarise, the hypotheses are as follows:

H_1 Items measured with an 11-point scale labelled from −5 to +5 will exhibit more negatively skewed distributions and higher means than items measured with the same scale labelled from 0 to 10.

H_2 Items measured using a – 5 to +5 scale will show a higher proportion of responses in the midpoint of the scale compared to items measured on a 0 to 10 scale.

3 Method

We contribute to the literature by analysing the results of a randomised experiment examining the impact of numerical labels on survey responses in a cross-national study. Specifically, we investigate how numerical values influence responses to five questions, each measured using an 11-point bipolar scale with labelled endpoints. These questions assessed the impact of climate change in general, as well as its impact on various groups (e.g., people in your country, people in developing countries). Details of these items are provided in Table 1.

Respondents were randomly assigned to scales ranging from 0 to 10 or from −5 to +5. The scales were presented horizontally on showcards (see Fig. 4 in the Appendix). Both scales had identical verbal endpoints anchored as "extremely negative" and "extremely positive", making the scales clearly bipolar.

The experiment was embedded in a cross-national face-to-face survey (Computer Assisted Personal Interviewing and Pen-and-Paper Personal Interviewing) conducted in Hungary ($n=1181$), Portugal ($n=1258$) and Great Britain ($n=906$) as part of the European Social Survey Round 8 Omnibus testing (2015–2017). For 79 respondents, it was not possible to identify which scale order they were assigned to[1]; therefore, they were removed from the analysis (see Table 2 for sample composition). Interviews were conducted in Hungarian, Portuguese, and English during May and June 2015. In terms of sampling approaches, Great Britain implemented a quota sampling approach (gender, working status, and children in the household), Hungary used a multi-stage random sample, and in Portugal respondents were proportionally selected by group (gender, age, education, employment, region, and settlement size). No incentives were provided to respondents. The questionnaire was drafted in English and then translated into Hungarian and Portuguese.

To test our hypotheses, we first compared the item distributions between the two experimental conditions, harmonising responses so that both scales spanned from 0 to 10. Independent samples t-tests were then used to evaluate potential differences in means between the experimental groups, with effect sizes assessed using Cohen's d.

To further analyse the impact of the different numerical values on response behaviour, linear regression models were estimated. An "attitude towards climate change" index was

[1] The assignment question was empty for those 79 cases, preventing us from assigning them to experimental conditions. 42 cases were missing in GB, 19 in Hungary, and 18 in Portugal.

Table 1 Questions included in the experiment

Items	Response scale
How positive or negative do you think the impact of climate change will be on people in the world?	0/10 – 5/+5
Please tell me how negative or positive you think the impact of climate change will be on each of the following:	
I2 …you personally	
I3…people in [Country]?	
I4…people in developing countries?	
I5…nature and the environment?	

Table 2 Descriptive statistics for all study variables

	Overall		0–10		– 5/+5	
Variable	%	*N*	%	*N*	%	*N*
Country						
Portugal	37.6	1258	39.1	648	36.2	610
Hungary	35.3	1181	34.5	573	36.1	608
Great Britain	27.1	906	26.4	438	27.8	468
Sex						
Female	56.5	1889	57.3	950	55.7	939
Male	43.5	1456	42.7	709	44.3	747
Age M (SD)	48.5	(18.3)	48.6	(18.2)	48.4	(18.3)
15–29	18.5	617	18.2	302	18.7	315
30–39	16.3	544	16.2	268	16.4	276
40–49	17.6	587	18.2	301	17.0	286
50–59	16.8	560	16.2	269	17.3	291
60–69	15.9	531	16.0	265	15.8	266
70+	15.1	504	15.3	253	14.9	251
Item nonresponse	0.1	2	0.1	1	0.1	1
Years of education M (SD)	12	(4.7)	11.9	(4.6)	12.1	(4.8)
Item nonresponse	1.2	40	1.2	20	1.2	20

M: Mean, SD: Standard Deviation

regressed on the experimental condition while controlling for the country where the survey was conducted, as well as respondents' age, sex, and education level. Finally, to assess whether the effect of the experimental condition was moderated by respondents' education level—considered a proxy for cognitive sophistication or ability—and to determine if the effect differed across countries, we estimated two additional models including interaction terms between the experimental condition and these two variables. Similar models were estimated to assess interactions between the experimental variable, sex, and age to explore potential moderation effects.

Before constructing the "attitude towards climate change" index, we conducted a reliability test by computing Cronbach's alpha and testing the equivalence of Cronbach's alpha within each country using the Cocron package in R (Diedenhofen 2016).

4 Results

4.1 Description of the sample

As expected, due to random allocation and relatively large sample sizes (see Table 2), there were no significant differences in the proportions of respondents allocated to each experimental condition by country (– 5 to +5 condition: 48.5% Portugal, 51.5% Hungary, 51.7% Great Britain; $X^2 = 2.961,\ p = .228$). Respondents' ages ranged from 15 to 97 (M=48.5,

SD = 18.3), and 57% of the sample were female. In terms of education, respondents had an average of 12 years of education (*SD* = 4.7).

4.2 Distributions and means by experimental condition

The distributions of the five items were compared between the two experimental groups. Generally, both conditions showed pronounced positively skewed distributions, with respondents indicating that climate change will have a negative impact. However, a significant difference emerged when comparing the proportions of responses for the first five categories between the two experimental groups $(X^2 = 73.382,\ p < .001)$.

Specifically, for the first item, 82% of respondents who were given the −5 to +5 scale selected a value between −5 and −1. In contrast, only 71% of respondents presented with the 0 to 10 scale selected a value within the equivalent range of 0 to 4. A similar pattern was observed for the other items (see Fig. 1), with larger proportions of respondents selecting a value between −5 and −1 compared to those selecting a value within the equivalent range of 0 to 4 (Item 2: 72% vs. 64%, Item 3: 78% vs. 67%, Item 4: 83% vs. 75%, Item 5: 86% vs. 77%).

In general terms, when considering equivalent values between 0 and 4, there is a 9-percentage-point difference between the two scales, with respondents more likely to lean toward the lower end of the scale in the −5 to +5 condition. The item with the largest discrepancy (11% points) was "Please tell me how negative or positive you think the impact of climate change will be on people in [Great Britain/Hungary/Portugal]". These findings are further reflected by the skewness of the items, with the −5 to +5 items consistently more positively skewed than the 0 to 10ones (see Table 5 in the Appendix for the full set of descriptive statistics). Additionally, the medians and modes are generally higher for the 0 to 10 scale. For example, for Item 1, the median and mode are 3 and 2, respectively, on the 0 to 10 scale, while they are 2 and 0 on the −5 to +5 scale (see Table 5 in the Appendix).

Respondents treat the midpoint differently between the two experimental conditions. Those assigned the 0–10 scale consistently select the midpoint more often than those who received the −5 to +5 scale (see Fig. 1). Thegap isgreatest for the item assessing the impact of climate change on the respondent's own country. While personal or global consequences-can elicit strong emotional responses – either positive or negative – evaluating the national-level effects may be harder or less salient, prompting greater use of a midpoint.

The means for all items were significantly higher on the 0 to 10 scale than on the −5 to +5 scale at an alpha level of 0.001 (Fig. 2). This pattern runs contrary to earlier research (Schwarz et al. 1991; O'Muircheartaigh et al. 1995; Tourangeau et al. 2007). This suggests that, in the context of climate change – a topic often associated with strong negative emotions – the explicit inclusion of negative numbers may have activated more critical evaluations. Respondents may have used the negative side of the −5 to +5 scale to signal the perceived seriousness of the issue. The discrepancy with previous studies suggests that the relative effect of the 0 to10 and −5 to +5 labels could be item or domain- specific.

Even though the differences between the means were statistically significant, the effect sizes were small, with Cohen's d ranging from 0.2 to 0.3 (see Table 5 in the Appendix).

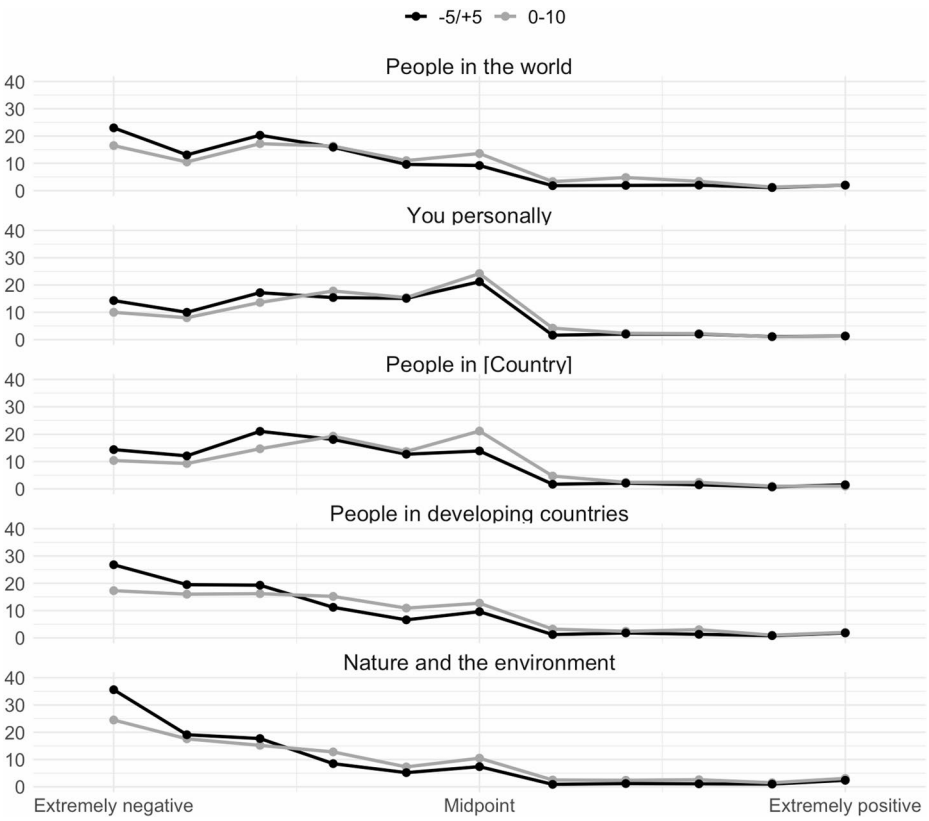

Fig. 1 Distribution of the items by experimental condition

We subsequently tested the differences in the means between the experimental conditions by country. The results, disaggregated by country, remained consistent, with means consistently and significantly higher for the 0 to 10 condition across all items and countries (see Fig. 3 in the text and Table 6 in the Appendix for the t-test and Cohen's d results). Only one item in Great Britain showed comparable means (Item 2: "Please tell me how negative or positive you think the impact of climate change will be on you personally"). While effect sizes for the pooled data were small, noticeable differences across countries emerged. In general, effect sizes were larger in Hungary, with Cohen's d values ranging from 0.3 to 0.4, and lower in Great Britain, ranging from 0 to 0.2 (see Table 6). The items with the largest effect sizes are Item 4 (impact on people in developing countries) and Item 5 (impact on nature and the environment).

To further assess the effect of different numerical labels on response behaviour while controlling for demographic variables[2], we estimated a linear regression model. The dependent variable in this model is an index derived by computing the average of the 10 items after harmonising the scales (five measured on the 0 to 10 scale and five on the −5 to +5 scale). Before constructing this index, we conducted internal consistency tests by comput-

[2] Even though as shown in Table 2 there are no differences in any of these variables, controlling for demographic variables helps with the precision of the estimates.

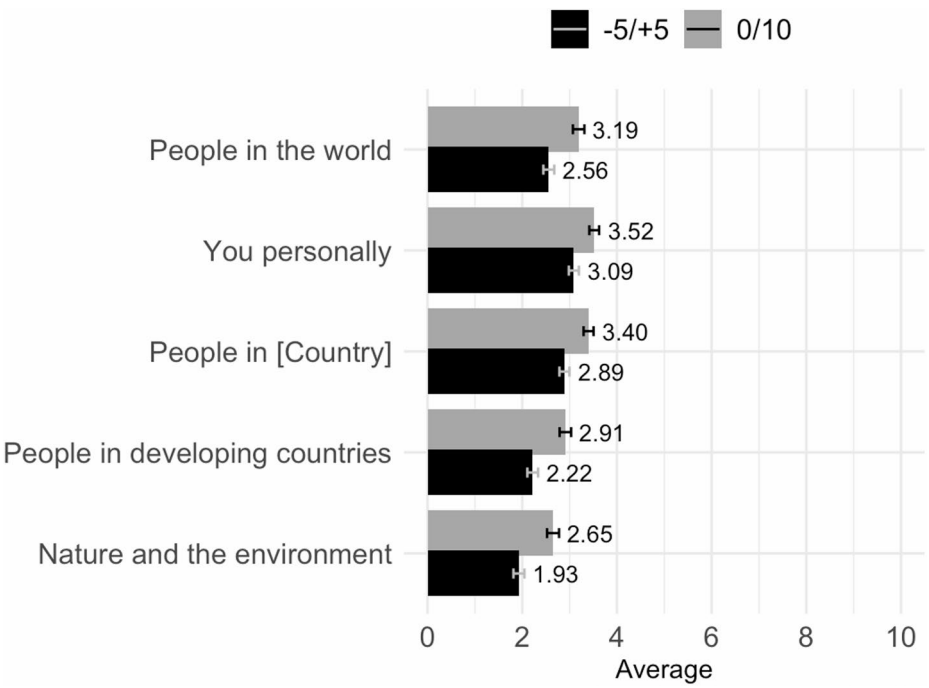

Fig. 2 Mean values and confidence intervals by experimental conditions

ing Cronbach's alphas and statistically tested the equivalence of the estimates within each country (Table 3). The null hypotheses were retained, indicating that Cronbach's alpha does not differ across the two experimental groups within each country. The internal consistency of the climate change items is high across the two scale formats and countries. As noted, there were no statistically significant differences in reliability between the −5/+5 and 0–10 groups within each country. This suggests that while numeric labels affect means and distributions, they do not appear to compromise measurement quality. In addition, we performed an exploratory factor analysis which highlighted the existence of a single factor, with factor loadings ranging from 0.68 to 0.92 (Table 3). The extracted factor explained 69% of the variance. In light of these results, we proceeded to create the index by computing the average of the items.

We then estimated a linear regression model with the climate change index as the dependent variable and the experimental condition as the main independent variable, controlling for country, sex, age, and education. To explore possible moderators, we estimated four models that included interaction effects. These models tested whether the effect of different numerical labels varied across countries, sexes, and ages, and whether respondents' education level, which is considered a proxy for cognitive ability, moderated the effect of the experimental manipulation. The results from these models are presented in Table 4.

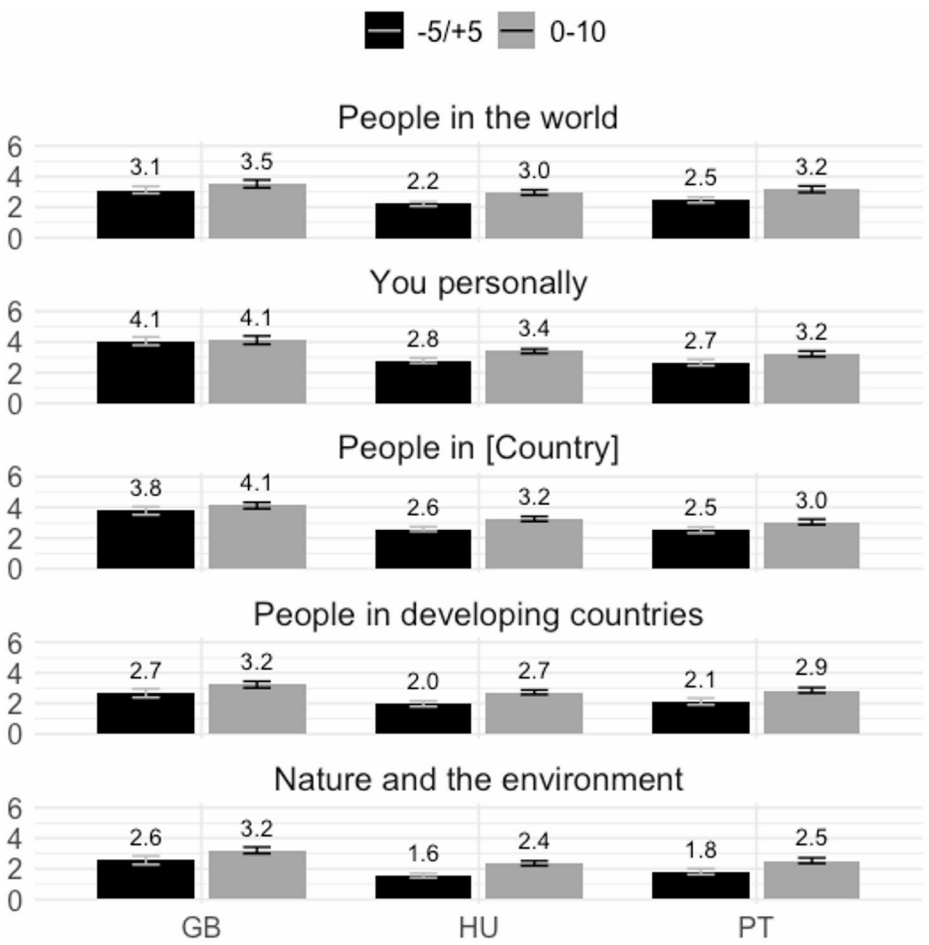

Fig. 3 Means and confidence intervals by experimental condition, item and country

Table 3 Reliability and exploratory factor analysis

Country	Cronbach's Alpha		Differences	*p*-value
	0–10	– 5+5	X^2	
GB	0.867	0.887	1.725	0.189
HU	0.921	0.920	0.013	0.906
PT	0.924	0.924	0.000	1.000
				Factor loadings
Item 1				0.677
Item 2				0.841
Item 3				0.916
Item 4				0.862
Item 5				0.830

Extraction Method: Maximum Likelihood

Variance explained: 69%

Table 4 Regression models exploring the impact of numerical labels on responses to climate change items

			Exp*Country		Exp*Gender		Exp*Age		Exp*Education	
	B	S.E.	B	S.E.	B	S.E.	B	S.E.	B	S.E.
(Constant)	2.969	0.154	2.983	0.159	2.968	0.158	2.918	0.179	2.920	0.179
Exp (– 5 + 5)	−0.594***	0.069	−0.628***	0.112	−0.592***	0.104	−0.595***	0.069	−0.594***	0.069
Hungary	−0.037	0.085	0.020	0.116	−0.037	0.085	−0.036	0.085	−0.038	0.085
Great Britain	0.877***	0.095	0.739***	0.128	0.877***	0.095	0.876***	0.095	0.877***	0.095
Female	−0.127	0.069	−0.130	0.069	−0.125	0.098	−0.128	0.069	−0.127	0.069
Age	0.004*	0.002	0.004*	0.002	0.004*	0.002	0.005	0.003	0.004*	0.002
Years of Education	−0.014	0.008	−0.014	0.008	−0.014	0.008	−0.014	0.008	−0.010	0.011
Exp*HU			−0.110	0.160						
Exp*GB			0.274	0.174						
Exp*Female					−0.004	0.138				
Exp*Age							−0.002	0.004		
Exp*Education									−0.008	0.015

N = 3206. The adjusted R-squared values are as follows: 0.058 for the main effects model and models with interactions involving gender, age and education. 0.059 for the interaction model involving the country variable.

The main effects model indicates that even after controlling for demographics, the means of the −5 to +5 scale remain significantly lower ($b = -0.594$, p-value < 0.001). Examining the interaction models, we see that the effect of the experimental manipulation does not differ significantly across the three countries, which is consistent with the bivariate findings. Similarly, the coefficients of the interaction terms involving sex, age, and education are not statistically significant. This suggests that the tendency to provide scores closer to the lower end of the scale when using the −5 to +5 scale is consistent across sexes, age groups, and education levels.

5 Discussion

Contrary to other aspects of scale design, such as the number of scale points, there has been limited recent research on the impact of the numerical labels attached to rating scales (see the literature review provided by DeCastellarnau 2018). This study contributes to addressing this gap by investigating the effect of numerical values on responses to five questions measured using 11-point bipolar scales, with respondents randomly assigned to scales ranging either from 0 to 10 or from −5 to +5. In addition, our study incorporates a cross-national component, while previous research has been predominantly conducted in the US. Including Hungary, Portugal and Great Britain allows us to capture variation along three relevant dimensions: survey climate, political context, and climate-change salience. The countries differ in survey literacy and exposure to surveys. In addition, Hungary represents a Central European country with a more recent democratic tradition and lower political trust, while Great Britain (Western Europe) and Portugal (Southern Europe) are longer-standing democracies with different histories of civic engagement and institutional trust. This variation allows us to explore whether the effect of numeric labels operates similarly across diverse democratic contexts and regions of Europe. Finally, climate change has different levels of salience in public discourse across the three countries: concern and politicisation are relatively high in Great Britain, lower in Hungary (in ESS Round 8, 61.7% of people in the UK believed that the climate is "definitely changing," compared to 48.4% in Hungary. Similarly, 60.2% of people in the UK felt it was their personal responsibility to reduce climate change, compared to just 32.9% in Hungary). This allows us to assess whether the effects of numeric labels remain consistent across countries with differing attitudes toward, and perceived salience of climate change topics.

Schwarz et al. (1991) showed that the item "How successful have you been in life, so far?" yielded different results depending on the scale used. In their study, a scale with numeric values ranging from 0 to 10 was not equivalent to an identical scale with values ranging from −5 to +5. Notably, respondents tended to prefer response categories on the positive side of the scale, resulting in higher means for the −5/+5 scale compared to the 0/10 scale. The authors attributed this shift in means to a change in the scale's interpretation: the 0/10 scale was perceived as unipolar, while the −5/+5 scale was perceived as bipolar.

Our results corroborate that the two scales are not equivalent. However, contrary to prior results, our analyses reveal an opposite trend: the means tend to be consistently and significantly lower for the −5/+5 scale.

Hypothesis 1 posited that items measured with an 11-point scale labelled from −5 to +5 would show more negatively skewed distributions and higher means than items measured on the same scale but labelled from 0 to 10. Our data do not support this hypothesis.

Schwarz et al. (1991) (see also Schwarz and Hippler 1995) attributed the non-equivalence of means between the two scales to a shift in scale polarity - from unipolar to bipolar. In our study, the items were unambiguously bipolar, with the endpoints anchored as "extremely negative" and "extremely positive". O'Muircheartaigh et al. (1995) argued that the verbal anchors used in Schwarz et al. (1991) could lead to interpreting the scale as either unipolar (with 0/10) or bipolar (with −5/+5). They replicated the experiment using explicit bipolar verbal anchors, and their results confirmed those of Schwarz et al. (1991), showing that even with clear bipolar scales, respondents tend to avoid responses on the negative side of the scale when the labels range from −5 to +5.

In contrast, our results do not corroborate these earlier findings. Across the three countries, respondents were more negative about the detrimental effects of climate change when using the −5/+5 scale. One possible explanation is that the negative values triggered stronger negative evaluations of the effects of climate change. While positivity bias implies that respondents tend to avoid overly negative evaluations, in the context of climate change, which is a highly debated topic in the media, at the centre of the political agenda, and in many ways controversial, respondents may have wanted to stressits strong negative effects. Indeed, Schwarz and Clore (1983) showed that individuals' current emotions can impact their responses to survey questions, suggesting that emotionally charged topics like climate change may elicit more extreme responses.

The discrepancy between our findings and previous studies (Schwarz et al. 1991; O'Muircheartaigh et al. 1995; Tourangeau et al. 2007), suggests that interpretations of the 0/10 and −5/+5 scales could be item- or topic-specific. Unfortunately, we do not have other items to test this assumption. As such, additional research is needed to assess the effect of numerical values on a diverse set of items, combining those related to divisive subjects, more neutral topics and topics generally perceived as positive.

Concerning the middle response category, hypothesis 2 predicted that items measured with the −5/+5 scale would exhibit a higher proportion of responses at the midpoint than those measured with a 0 to 10 scale. This hypothesis is also rejected. In fact, the proportion of midpoint responses in the 0/10 scale was consistently higher than that of the −5/+5 scale. In this context, our results replicated those of O'Muircheartaigh et al. (1995). This is due to the way numeric labels shape the perceived meaning of the midpoint. On a 0/10 scale, "5" is typically interpreted as a moderate or balanced response. However, "0" on a -5/+5 scale can be perceived as the absence of any impact rather than a moderate position. O'Muircheartaigh, et al. (1995) found that midpoint selection depends on both scale format and topic salience. On the salient topic of climate change, respondents may have avoided "0" to signal engagement, even when their views were not strongly polarised. Indeed. when the midpoint is denoted as "0", respondents might perceive it as absence of evaluation. To avoid the appearance of "sitting on the fence" – especially on a topic like climate change –

respondents were less likely to select "0", and more inclined to select "5", viewing it as a moderate position that is in the middle of a continuum between 0 and 10 (O'Muircheartaigh et al. 1995). In sum, the lower selection of "0" reflects the interaction between numeric interpretation and the emotional and normative weight of the climate change topic.

This study offers important contributions to both public opinion research and survey methodology. The findings underscore how the numerical structure of response scales can shape answers on politically and socially salient topics such as climate change. The consistent trend toward lower mean scores on the −5/+5 scale suggests that numeric labels influence how respondents perceive negative and positive evaluations, even when verbal anchors are held constant. This challenges assumptions of positivity bias and highlights the role of context-specific interpretations, emphasising the need to examine how the choice of numeric labels on rating scales interacts with the perceived importance and emotional charge of a topic, ultimately influencing response patterns in public opinion research.

Our findings also complicate prior conclusions – particularly those suggesting that respondents avoid negative values – and instead show that negative values can prompt stronger negative responses on topics perceived as important and emotionally relevant. Moreover, the differential interpretations of midpoints across formats (i.e., "0" on a -5/+5 scale vs. "5" on a 0–10 scale) add to the literature on response behaviour. Together, these findings deepen our understanding of how numeric formatting in rating scales can affect both data quality and substantive conclusions. From a practical perspective, the findings also caution against combining data from surveys that use different scales, even when the scale length and verbal labels are identical, if the numerical labels differ, due to comparability issues.

6 Conclusions

Our research confirms that respondents use the two scales differently, and their values are not equivalent. Despite the clear polarity of the items in our study, unlike in Schwarz et al. (1991), respondents remain uncertain about the range of the scale and the distance between the two verbally labelled endpoints (Tourangeau et al. 2007).

The tasks associated with optimally completing questionnaires involve considerable cognitive effort from respondents. They rely on and pay close attention to every detail of the questions to complete their task (Tourangeau et al. 2007). In this process, they often leverage elements that might seem secondary, such as the numbers attached to scale points.

Contrary to earlier research suggesting that respondents avoid negative values on bipolar scales (Schwarz et al. 1991; O'Muircheartaigh et al. 1995; Tourangeau et al. 2007), we consistently find lower means on the −5/+5 scale compared to the 0/10 scale. Because climate change is often associated with severe consequences, the presence of negative numbers may have made the lower end of the −5/+5 scale more salient and cognitively accessible, encouraging respondents to register more criticism. In contrast, the 0/10 scale may have felt more

abstract or softened by the lack of explicit negative numbers, resulting in higher average scores. This finding suggests that, on emotionally charged topics, scale polarity can amplify emotional interpretation rather than suppress it as predicted by positivity bias theory. Rather than avoiding negative numbers, respondents may have used them to express the perceived seriousness of the issue.

Given this divergence from previous research, we suggest that more research is conducted to better understand the impact of numerical labels. Indeed, while previous research indicates a general tendency to avoid negative values, our study suggests that this bias weakens in the context of highly salient and morally charged topics like climate change. In conclusion, the effect of numeric labels might not be consistent, but topic-dependent. For this reason, additional research is needed to assess a diverse set of items, combining those related to divisive subjects, neutral topics, and those generally perceived as positive.

Despite its contributions, our study has certain limitations. Although it covers three countries incorporating a cross-national component while most previous studies were conducted in the US, additional research including a broader and more diverse set of countries is warranted. In addition, only in Hungary did we employ random sampling procedures, while in Great Britain and Portugal quota samples were drawn. Finally, while we analysed five different items, future research should encompass a wider range of topics and items.

7 Appendix

See Fig. 4 and Tables 5, 6.

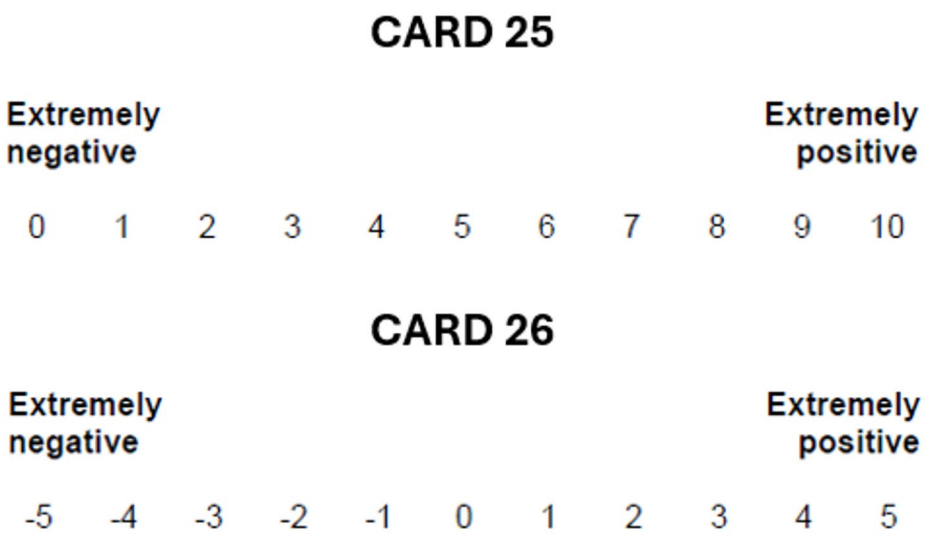

Fig. 4 Showcards

Table 5 Descriptive statistic for the five items (harmonised values) by experimental condition

	0–10						– 5+5						Co-hen's d
	Mean	Median	Mode	SD	Skewness	Kurtosis	Mean	Median	Mode	SD	Skewness	Kurtosis	
Impact of climate change on people in the world?	3.19	3	2	2.469	0.690	-0.012	2.56	2	0	2.322	1.115	1.180	0.262
How negative or positive impact of climate change on: you personally?	3.52	4	5	2.120	0.354	0.268	3.09	3	5	2.122	0.520	0.431	0.202
How negative or positive impact of climate change on: people in Britain?	3.40	3	5	2.122	0.402	0.135	2.89	3	2	2.146	0.871	0.991	0.238
How negative or positive impact of climate change on: people in developing countries?	2.91	2	0	2.401	0.863	0.395	2.22	2	0	2.270	1.324	1.682	0.295
How negative or positive impact of climate change on: nature and the environment?	2.65	2	0	2.599	1.095	0.622	1.93	1	0	2.327	1.643	2.618	0.291

Table 6 T-test and cohen's d by country

		1–10		– 5 + 5					
		M	*SD*	*M*	*SD*	*t*	p-value	Mean difference	effect size (Cohen's d)
GB	Impact of climate change on people in the world?	3.52	2.643	3.13	2.873	2.046	0.041	0.389	0.141
	How negative or positive impact of climate change on: you personally?	4.12	2.100	4.06	2.253	0.410	0.682	0.061	0.027
	How negative or positive impact of climate change on: people in Britain?	4.12	2.142	3.78	2.423	2.159	0.031	0.342	0.148
	How negative or positive impact of climate change on: people in developing countries?	3.23	2.727	2.67	2.716	2.947	0.003	0.558	0.205
	How negative or positive impact of climate change on: nature and the environment?	3.21	2.928	2.56	2.941	3.232	0.001	0.652	0.221
HU	Impact of climate change on people in the world?	2.96	2.055	2.21	1.812	6.543	0.000	0.752	0.387
	How negative or positive impact of climate change on: you personally?	3.40	1.887	2.79	1.768	5.560	0.000	0.607	0.333
	How negative or positive impact of climate change on: people in Britain?	3.25	1.846	2.59	1.701	6.156	0.000	0.654	0.372
	How negative or positive impact of climate change on: people in developing countries?	2.73	1.959	1.98	1.742	6.642	0.000	0.749	0.404
	How negative or positive impact of climate change on: nature and the environment?	2.36	2.140	1.56	1.651	7.073	0.000	0.803	0.419
PT	Impact of climate change on people in the world?	3.17	2.665	2.47	2.230	4.853	0.000	0.701	0.283
	How negative or positive impact of climate change on: you personally?	3.22	2.249	2.66	2.111	4.429	0.000	0.563	0.256
	How negative or positive impact of climate change on: people in Britain?	3.05	2.227	2.52	2.132	4.143	0.000	0.529	0.242
	How negative or positive impact of climate change on: people in developing countries?	2.86	2.509	2.13	2.308	5.115	0.000	0.732	0.302
	How negative or positive impact of climate change on: nature and the environment?	2.54	2.686	1.82	2.289	4.903	0.000	0.715	0.287

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript. Authors declare they have no financial interests related to the paper. All authors contributed to the study conception and design. The first draft of the manuscript was written by Gianmaria Bottoni and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

Aybar, C., Pérez, V., Pavía, J.M.: Scale matters: unravelling the impact of likert scales on political self-placement. Qual. Quant. **58**, 3725–3746 (2024). https://doi.org/10.1007/s11135-023-01825-2

Couper, M.P., Tourangeau, R., Kenyon, K.: Picture this! exploring visual effects in web surveys. Pub. Opin. Q. **68**(2), 255–266 (2004)

DeCastellarnau, A.: A classification of response scale characteristics that affect data quality: a literature review. Qual. Quant. **52**(4), 1523–1559 (2018)

Diedenhofen, B.: Cocron: statistical Comparisons of Two or more Alpha Coefficients (Version 1.0–1). (2016)

Evans, J.: On a scale from 1 to 10, how much do the numbers used in survey scales really matter? Pew Research Center. (2019). https://www.pewresearch.org/decoded/2019/03/06/on-a-scale-from-1-to-10-how-much-do-the-numbers-used-in-survey-scales-really-matter/ (2019). Accessed 13 June 2025

Gummer, T., Kunz, T.: Using only numeric labels instead of verbal labels: stripping rating scales to their bare minimum in web surveys. Social Sci. Comput. Rev. **39**(5), 1003–1029 (2021). https://doi.org/10.1177/0894439320951765

Höhne, J.K., Krebs, D., Kühnel, S.M.: Measuring income (in) equality: comparing survey questions with unipolar and bipolar scales in a probability-based online panel. Soc. Sci. Comput. Rev. **40**(1), 108–123 (2022)

Kennedy, C.: Bipolar scale. In: Lavrakas, P.J. (ed.) Encyclopedia of Survey Research Methods, pp. 63–64. Sage (2008)

Krebs, D.: The impact of response format on attitude measurement. In: Salzborn, S., Davidov, E., Reinecke, J. (eds.) Methods, theories, and Empirical Applications in the Social Sciences, pp. 105–113. Springer, Berlin (2012)

Krosnick, J.A.: Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl. Cogn. Psychol. **5**(3), 213–236 (1991)

Krosnick, J.A., Fabrigar, L.R.: Designing rating scales for effective measurement in surveys. In: Lyberg, L. E., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., Trewin, D. (eds) Survey Measurement and Process Quality. Wiley, Hoboken, pp. 141–64 (1997).

Menold, N.: Rating-scale labeling in online surveys: an experimental comparison of verbal and numeric rating scales with respect to measurement quality and respondents' cognitive processes. Sociol. Methods Res. **49**(1), 79–107 (2020)

Menold, N.: Response bias and reliability in verbal agreement rating scales: does polarity and verbalization of the middle category matter? Soc. Sci. Comput. Rev. **39**(1), 130–147 (2021)

O'Muircheartaigh, C., Gaskell, G., Wright, D.B.: Weighing anchors: verbal and numeric labels for response scales. J. Off. Stat. **11**, 295–308 (1995)

Saris, W.E., Gallhofer, I.: Estimation of the effects of measurement characteristics on the quality of survey questions. In: Survey Research methods (Vol. 1, No. 1, pp. 29–43). (2007)

Schaeffer, N.C., Presser, S.: The science of asking questions. Ann. Rev. Sociol. **29**(1), 65–88 (2003)

Scherpenzeel, A.C., Saris, W.E.: The validity and reliability of survey questions: a meta-analysis of MTMM studies. Sociol. Methods Res. **25**(3), 341–383 (1997)

Schwarz, N., Clore, G.L.: Mood, misattribution, and judgments of well-being: Informative and directive functions of affective States. J. Pers. Soc. Psychol. **45**(3), 513 (1983)

Schwarz, N., Hippler, H.J.: The numeric values of rating scales: a comparison of their impact in mail surveys and telephone interviews. Int. J. Public. Opin. Res. **7**(1), 72–74 (1995)
Schwarz, N., Knäuper, B., Hippler, H.J., Noelle-Neumann, E., Clark, L.: Rating scales numeric values May change the meaning of scale labels. Pub. Opin. Q. **55**(4), 570–582 (1991)
Schwarz, N., Grayson, C.E., Knauper, B.: Formal features of rating scales and the interpretation of question meaning. Int. J. Public. Opin. Res. **10**(2), 177–184 (1998)
Stefkovics, Á., Höhne, J.K.: Examining the link between the 'Middle means typical' heuristic and answer behavior. Surv. Pract. **17** (2024). https://doi.org/10.29115/SP-2024-0009
Thurstone, L.L.: Attitudes can be measured. Am. J. Sociol. **33**(4), 529–554 (1928)
Toepoel, V., Dillman, D.A.: Words, numbers, and visual heuristics in web surveys: is there a hierarchy of importance? Social Sci. Comput. Rev. **29**(2), 193–207 (2011)
Tourangeau, R., Rips, L.J., Rasinski, K.: The Psychology of Survey Response. Cambridge University Press, Cambridge (2000)
Tourangeau, R., Couper, M.P., Conrad, F.: Spacing, position, and order: Interpretive heuristics for visual features of survey questions. Pub. Opin. Q. **68**(3), 368–393 (2004)
Tourangeau, R., Couper, M.P., Conrad, F.: Color, labels, and interpretive heuristics for response scales. Pub. Opin. Q. **71**(1), 91–112 (2007)
Weng, L.J.: Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. Educ. Psychol. Meas. **64**(6), 956–972 (2004)
Wildt, A.R., Mazis, M.B.: Determinants of scale response: label versus position. J. Mark. Res. **15**(2), 261–267 (1978)