Do you agree? Do you strongly agree? The effect of the number of response categories on response processes and verification of substantive hypotheses

Artur Pokropek, Tomasz Żółtak, Marek Muszyński

Institute of Philosophy and Sociology, Polish Academy of Sciences
Department of Computational Social Sciences, Institute of Philosophy and Sociology of the Polish Academy of Sciences, 00-330 Warszawa, ul. Nowy Świat 72, Phone: 0048501764306;
Emails: artur.pokropek@ifispan.edu, tomasz.zoltak@ifispan.edu.pl;
marek.muszynski@ifispan.edu.pl

Author Note

Artur Pokropek is the head of the Department of Computational Social Sciences at the Institute of Philosophy and Sociology, Polish Academy of Sciences.

Tomasz Żółtak is an Assistant Professor in the Department of Computational Social Sciences at the Institute of Philosophy and Sociology, Polish Academy of Sciences.

Marek Muszyński is an Assistant Professor in the Department of Computational Social Sciences at the Institute of Philosophy and Sociology, Polish Academy of Sciences.

Funding

This research is financed by the National Science Centre (NCN) research grant (2019/33/B/HS6/00937) Understanding response styles in self-report data: consequences, remedies and sources.

Final version of this paper was published as

Artur Pokropek, Tomasz Żółtak, Marek Muszyński, Do You Agree? Do You Strongly Agree? The Effect of the Number of Response Categories on Response Processes and Verification of Substantive Hypotheses, *International Journal of Public Opinion Research*, Volume 38, Issue 1, Spring 2026, edaf070, https://doi.org/10.1093/ijpor/edaf070

Abstract

This study investigates how the number and labeling of response categories in survey scales affect respondent behavior, psychometric properties, and substantive conclusions. Using data from a web-based survey experiment with over 2,800 participants, we randomly assigned respondents to scales that varied in the number of response options and labeling formats. Engagement, response style tendencies, reliability and convergent validity were assessed through a combination of process data (e.g., response times, cursor movements), self-reports, and psychometric modeling. Our findings suggest that scales with a greater number of response categories tend to exhibit higher reliability, although the improvements are modest and depend on the specific scale. While scales with more response options required longer completion times and prompted more complex response behavior, the number of response categories did not influence substantive conclusions in simple regression models.

Keywords: rating scales, number of response categories, Likert-type, survey design, response styles, computer-based paradata.

Do you agree? Do you strongly agree? The effect of the number of response categories on response processes and verification of substantive hypotheses

Despite growing research on the optimal number of response options and labeling formats, survey designers often rely more on personal experience than on empirical evidence when making such decisions (Simms et al., 2019). This study contributes systematic evidence to inform ongoing debates in the field. Beyond traditional analyses of reliability and validity, we also assess respondents' attitudes toward different response scales, examine tendencies for response styles, and analyze response processes using computer-based paradata, including response times and cursor movements.

Prior Research on the Number of Response Categories

Reliability. Longer response scales are argued to improve measurement precision by increasing variance and allowing for greater response granularity (Finn et al., 2015; Flamer, 1983; Hilbert et al., 2016; Weng, 2004). However, critics argue that this added variance may reflect noise rather than signal, and that more response options can confuse respondents, add more cognitive burden, and lead to satisficing, particularly among cognitively burdened individuals (Krosnick, 1991; Krosnick et al., 1996). More complex response scales can lead to processing errors on all stages, starting from more difficult comprehension, due to more information to be processed, up to more challenging response selection (Tourangeau et al., 2000). Human limitations in distinguishing subtle differences in complex constructs also challenge the usefulness of longer scales (Cox, 1980; Symonds, 1924; Tourangeau et al., 2000). Simulations by Lozano et al. (2008) suggest that reliability and validity improve up to seven

categories, beyond which benefits plateau, a finding supported by several studies (Menold & Tausch, 2015; Preston & Colman, 2000; Rakhshani et al., 2024; Simms et al., 2019; Weng, 2004). Still, shorter scales (2–3 options) can sometimes yield comparable reliability (Jones & Loe, 2013), and others suggest that 4–7 options minimize measurement error (Abulela & Khalaf, 2024). Maydeu-Olivares et al. (2009) found that while more categories increase reliability and item information in IRT models, they worsen model fit—indicating that some of the added variance may remain psychometrically unaccounted for.

Validity. Empirical findings are mixed: some studies report no differences across scale lengths (Maydeu-Olivares et al., 2009; Simms et al., 2019), while others argue that scales with up to 11 options may be too long and less valid (Revilla et al., 2014). Excessive options can reduce discriminant validity, as respondents may struggle to differentiate between similar choices (Preston & Colman, 2000; Lozano et al., 2008), especially during the selection phase of answering, where initial judgments are mapped onto available options (Tourangeau & Rasinski, 1989). This increases cognitive load and may trigger method effects such as careless responding or response styles (Revilla et al., 2014). In contrast, scales with five to seven points are often found to optimize psychometric properties by balancing precision and cognitive demands (Krosnick & Presser, 2010; Weijters et al., 2010). Consequently, while longer response scales may capture finer nuances, their advantages in validity may be offset by potential respondent fatigue and increased response scale cognitive complexity.

Response processes. Paradata such as response times and cursor movements offer novel insights into respondents' cognitive engagement beyond traditional survey measures. Studies show that longer scales and labeling strategies can increase cognitive load, observable through paradata (Horwitz et al., 2017; Stieger & Reips, 2010). For instance, Horwitz et al. (2017) found

that complex questions led to longer hovering over items, indicating cognitive difficulty. Stieger and Reips (2010) noted differences in response times and cursor paths across question formats, though quick "click-through" behavior occurred regardless of format.

Recent work highlights the potential of cursor tracking to detect careless or inattentive responding. Pokropek et al. (2024) showed that analyzing cursor speed, hesitations, and interaction with approximate areas of interest can identify disengagement. Similarly, Fernández-Fontelo et al. (2023) demonstrated that mouse movement data can predict difficulties during web surveys, aiding questionnaire improvement. Still, systematic studies on how specific scale formats affect paradata are scarce (Couper & Peterson, 2017; Pokropek et al., 2023).

Careless or insufficient effort responding (C/IER)—inattentive answering not based on content—may also be linked to scale format. It results from low attention, motivation, or ability and leads to poor data quality. Indicators of C/IER include straightlining and Mahalanobis distance (Meade & Craig, 2012). While research is limited, some formats—such as numeric-only labels—are associated with higher rates of non-differentiation (Gummer & Kunz, 2021).

Self-reported Engagement. Self-reported indicators of engagement and attentiveness, such as task involvement and subjective interest, have received limited empirical attention in research on response scale design. While some evidence suggests that respondents prefer scales with 7 to 10 points (formats also associated with higher reliability and validity; Preston & Colman, 2000), very few studies systematically examined how subjective engagement is linked to data quality, e.g. lower propensity of C/IER. Nonetheless, self-reported measures can provide valuable insights into how respondents experience and interpret different formats, offering a complementary perspective to traditional psychometric evaluations (Meade & Craig, 2012).

Response Styles. The number of response categories in rating scales has been shown to influence not only psychometric properties but also susceptibility to specific response styles, such as extreme response style (ERS), midpoint response style (MRS) or acquiescence response style (ARS). Response styles are systematic tendencies in how respondents use response scales and can affect the interpretation of results, especially when results are not RS-adjusted (He & Van de Vijver, 2015; Ulitzsch et al., 2024). Weijters et al. (2010) demonstrated that scales with more response options tend to reduce extreme response styles but may simultaneously increase the use of midpoint categories, especially when verbal labels are sparse or ambiguous. This suggests that expanding scale length can shift rather than eliminate response style biases. Kieruj and Moors (2010) found that scales with more response categories, e.g. 9-, 10- or 11-point scales, lead to increased levels of MRS, but that ERS is visible in all response scales, from 5- to 11-point. A similar observation pointing to no relation between number of response categories and ERS was made by Kutscher and Eid (2020), while Corrado and Joxhe (2020) suggested that 11-point scales evoke more ERS than 7-point.

Other studies found that no particular response scale format is related to ARS, which is rather an effect of fatigue or careless responding, not function of response style (Moors et al., 2014). The relation between response styles and response scale format needs further research, especially that the multitude of models used to estimate response styles effects leads to low comparability of results between different studies (Schoenmakers et al., 2024).

Labelled categories. Verbal labels (e.g., "strongly agree") help respondents interpret response options, especially for complex constructs, while numeric-only labels (e.g., 1–5) may be easier to process on small screens. However, numeric-only formats tend to confuse respondents, reduce psychometric quality, and decrease interpretability (Gummer & Kunz, 2021;

Menold, 2020; Weng, 2004). Full labelling of all categories is also preferred over end-only labelling, which is associated with more extreme responses and lower reliability and validity (Menold et al., 2014; Menold & Tausch, 2015).

Yet, there is a limit to how many labels participants can meaningfully differentiate—on long scales like 11-point, some may ignore intermediate options (Kutscher & Eid, 2020; Revilla et al., 2014; Menold et al., 2014; Weng, 2004). Labelling format also affects response styles, with more ERS seen in numeric-only or end-labelled scales (Gummer & Kunz, 2021; Moors et al., 2014), likely due to increased visual focus on endpoints (Menold et al., 2014).

Odd or Even. Odd-numbered response scales (e.g., 5- or 7-point) include a neutral midpoint, allowing respondents to express neutrality when constructs have both positive and negative aspects. This can reduce bias by avoiding forced choices. In contrast, even-numbered scales (e.g., 4- or 6-point) exclude a midpoint, potentially encouraging more decisive responses. However, recent findings show that the midpoint may signal neutrality or serve as a form of non-response, introducing potential measurement error if not accounted for (Tijmstra & Bolsinova, 2025).

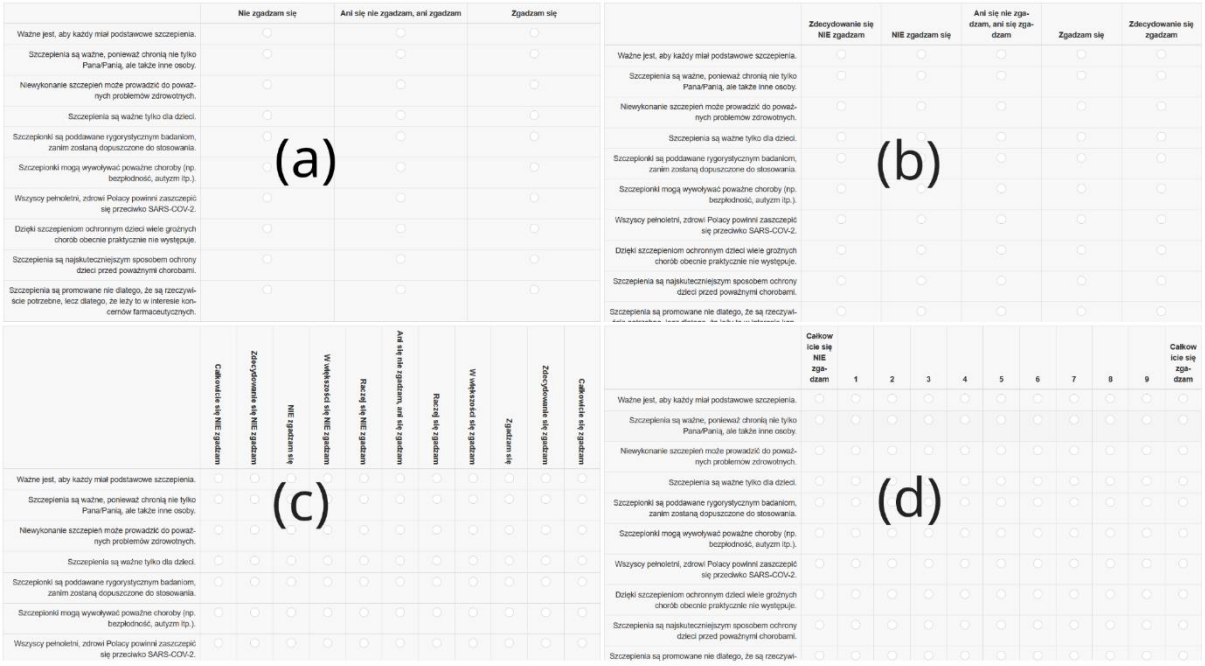
Research aims

In this work, we present the results of a self-administered online survey experiment in which we randomly assigned respondents to scales of different numbers of response categories: 3-, 4-, 5-, 6-, 7-, 10- and 11-point, and differently labelled response options. Figure 1 presents

four examples of scales with different numbers of categories and labelling options.

Figure 1

Example of Survey Screens with a Different Number of Response Options and Labelling



Note. The panels show: a) three categories (all labelled); (b) five categories (all labelled); (c) eleven categories (all labelled); (d) eleven categories (end labelled)

We examine four groups of effects that could differ between scales with varying numbers of response categories and labelling: (1) respondent behaviours related to engagement and attentiveness, (2) subjective experiences, and (3) tendencies toward response styles, as well as (4) scale's psychometric characteristics: reliability and validity, here operationalised as correlations with other variables.

First, we analyse whether scales with varying numbers of response options induce differences in engagement and attentiveness during the response process. This aspect is examined using process data indicators, such as response time and cursor movements, but also

popular inattentive responding indices, such as straightlining and Mahalanobis distance (Horwitz et al., 2017; Pokropek et al., 2023), which provide insights, although indirect, into the cognitive effort invested by respondents. Additionally, we include self-reported attentiveness measures (Meade & Craig, 2012) to capture subjective experiences of focus, effort, interest, and perceived burden during the task. Scales with more response options may be perceived as more complex or time-consuming, potentially affecting the respondents' willingness to engage, thus affecting the overall quality of responses (Revilla et al., 2014; Tourangeau et al., 2000).

Third, we evaluate whether the number of response categories influences the extent to which respondents exhibit response styles (RS), such as extreme or midpoint responding (Wetzel et al., 2016). This aspect is critical for identifying whether the design of response scales introduces bias that might undermine the validity of scale-based inferences. To this end, we used Multidimensional Nominal Response IRT models (MNRMs), including both middle and extreme response styles (Falk & Cai, 2016; Wetzel & Carstensen, 2017). MNRM is a popular model used for RS adjustment with evidence-based properties and a large versatility of use and estimation (Schoenmakers et al., 2024).

Then, we examine whether scales' length differences are reflected in their psychometric properties. Specifically, we measure the reliability of scales using the omega coefficient (McDonald, 1999).

Finally, we compared the validity of different response scale formats, assessed by examining correlations between the scales and several variables that, according to established social theories, should positively correlate with the constructs measured by the scales. In the subsequent analysis, we employed Ordinary Least Squares (OLS) regression to examine whether

scales with different numbers of response categories produced statistically distinct conclusions regarding their relationships with other variables.

In the light of literature review we formulate the following research hypothesis:

Hypothesis 1: Participants will spend more time on scales with more response categories, especially when all categories are labelled.

Hypothesis 2: Respondents will report more burden when responding to scales with more response categories, due to their larger cognitive complexity.

Hypothesis 3: Scales with more response categories, especially with only end labels, will lead to higher careless responding, as evidenced by straightlining, Mahalanobis distance, and vertical speed/acceleration cursor moves indices.

Hypothesis 4a: The variance of stylistic latent traits will increase with the number of response categories, due to elevated cognitive burden, which may increase the likelihood of some respondents shifting into more automatic and less reflective response processes.

Hypothesis 4b: The variance of stylistic latent traits will be lower in all-labelled response scales in comparison to end-labelled scales, reflecting that labels help participants to process response categories.

Hypothesis 5: Scales with more response categories will yield higher reliability up to approximately 7 categories, beyond which improvements will plateau or be minimal, as estimated by McDonald's omega coefficient.

Hypothesis 6: The relationship between scale length and validity will follow a curvilinear pattern, with improvements up to 5-7 categories, after which validity may stabilize or even slightly decrease for scales with 10-11 points.

While these hypotheses reflect the patterns most consistently documented in prior

research, our analytical approach allows for the detection of both linear and non-linear relationships. The inclusion of scales ranging from 3 to 11 points enables us to identify potential plateau effects and optimal scale lengths, particularly for psychometric properties where the literature suggests diminishing returns beyond seven categories (Lozano et al., 2008; Preston & Colman, 2000; Simms et al., 2019).

Method

Participants

Data was collected between 23rd November and 6th December 2021 in web-based mode, using a Polish opt-in Internet panel Ariadna. Respondents participated in the study using their own computers (either desktop or laptop); responding on a smartphone was not allowed, because of the fundamental differences between survey layout and ways of interaction with survey interface between computers and mobile devices. At the current state, the between-devices comparability of our log-data collection method cannot be guaranteed.

The survey lasted for about 15 minutes and was preceded by a Pathfinder cognitive abilities test that had a median time of completion of around 15 minutes – information on the length of each part of the study is provided in table S1 in the supplementary materials. The cognitive test had norms only for people up to the age of 50, hence our sample was truncated at this age.

Respondents were remunerated for the completed interview with points that could be exchanged into small gifts. The estimated value of this study is around 1€. The participation rate reached 30% (out of the total number of respondents invited to the survey, 30% clicked the link and initiated the cognitive test).

Out of the total of 4,124 participants who started the cognitive test, only 3,703 made it through to the survey part. At this point they were randomly assigned to the response scale and format condition. 369 respondents were assigned to the condition with a slider response scale, which was excluded from our analysis. Out of the remaining 3,334 participants, 3,176 successfully finished the survey. Thus, the break-off rate on the level of the cognitive test achieved 10% and on the survey level it amounted to only 5%. Moreover, the participants were excluded from the analysis due to the following reasons: (1) excessively long completion of the questionnaire (more than 2 hours; 2 exclusions), (2) excessively fast completion of the questionnaire (five minutes or less; 324 exclusions). The exclusions served to retain only participants that finished the study and responded in at least partially attentive mode in order to compare the regular response processes and data between the response scales of different formats. This resulted in a final size of the analysed sample of 2,850 respondents.

The collected sample used gender, age group, and self-reported education level to control for sample comparability to the general Polish population. In order to model its structure the following quotas were used: a) gender: 50% female, 50% male; b) age group: 18-29: 30%; 30-39: 35%; 40-50: 35%; c) education level: 35% higher education; 65% primary and secondary education level. Maximum diversion of the final marginal distributions was set to ±10%.

Dropout and exclusions only marginally affected the sample characteristics: we found no significant differences in gender or education level composition neither between respondents who dropped out during the cognitive test preceding the questionnaire and those who proceeded to the questionnaire nor between respondents who dropped out while completing the questionnaire or were excluded because of excessively long or fast survey completion and those retained in the final sample. Although there were statistically significant differences in age

distribution, indicating that younger respondents were slightly more likely to drop out or be excluded, these differences were small in absolute terms and did not meaningfully distort the overall sample composition (more details in Tables S2 and S3).

Additionally, respondents were excluded, but only from the analysis including cursor moves indices because of the technical problems preventing us from capturing paradata correctly on a given survey screen (see Supplementary Online Materials). Characteristics of the final sample are presented in Table 1.

Table 1

Socio-demographic Characteristics of Respondents

Characteristic	Value	N	Percentage
Gender	Male	1319	46.3%
	Female	1531	53.7%
Age Group	18-29	683	24.0%
	30-39	921	32.3%
	40-50	1246	43.7%
Education Level	Secondary	1535	53.9%
	Tertiary	1315	46.1%
Total		2850	100.0%

Materials

Respondents were randomly assigned to one of 9 experimental conditions, varying the number and format of response categories (3, 4, 5, 6, 7, 10, 11 all-labelled categories, 10 and 11 end-labelled categories; see Table 2 and Supplementary Online Materials Table S4 and S5).

Additionally, they were randomly assigned to one of three instruction sets (standard, warning, or appeal) delivered via a short video. The instructions aimed to influence participants' motivation to respond attentively, with standard instructions providing a basic overview of the

survey, warning instructions cautioning against careless responses as that careless respondents could be identified, and appeal instructions emphasising the social importance of the survey (see Pokropek et al., 2023 for more details and Table S9 for instructions used in the procedure). This manipulation was part of a separate study not directly related to the present analyses; therefore, we do not elaborate on it further here (see Table S6 for related analyses).

Approval to conduct this study was sought and obtained in a written form from an institutional ethical research committee.

The study utilised four survey scales: the Vaccination Scale, the Institutional Trust Scale, the Reading Attitudes Scale, and the Joy of Reading Scale. All scales were delivered in Polish.

The Vaccination Scale assesses attitudes toward vaccinations using ten items covering importance, societal benefits, and safety concerns, including three reverse-coded statements. Based on the Eurobarometer (European Commission, 2019), it was expanded to capture both supportive and sceptical views.

The Institutional Trust Scale includes 25 items measuring trust in 20 real and 5 fictitious institutions (e.g., national government, police, WHO), the latter used to detect inattentiveness or overclaiming. A “I don’t know this institution” option was included for all items. The scale consists of two dimensions: national and international trust, which were also analyzed separately (see Table S7).

The Reading Attitudes Scale measures reading self-efficacy through six items, including three reverse-coded statements. It was adapted from PISA (OECD, 2010).

The Joy of Reading and Reading Habits Scale includes 11 items, five reverse-coded. The version used was from PISA 2009, which included the most items (OECD, 2010: 112).The inclusion of diverse topics was intentional – it allowed us to test the effects of response scale

formats across a range of constructs varying in emotional intensity, content, and social desirability.

The general cognitive ability test Pathfinder, administered prior to the survey in all conditions, is a brief, 15-minute gamified assessment that measures both verbal and nonverbal abilities through five subtests, including vocabulary knowledge, verbal analogies, and matrix reasoning (Malanchini et al., 2021). Pathfinder was used as a measure of cognitive abilities to validate several survey scales, such as reading attitudes and reading habits, which are theoretically expected to correlate (e.g., Bråten et al., 2025). The distribution of Pathfinder scores also served other research objectives not addressed in this paper.

Table 2

Number of Participants in Specific Experimental Conditions

	Standard instruction		**Warning instruction**		**Appeal instruction**		**Overall**	
Response scale	**N**	**Pct**	**N**	**Pct**	**N**	**Pct**	**N**	**Pct**
3 cat (all labelled)	109	3.8%	100	3.5%	99	3.5%	308	10.8%
4 cat. (all labelled)	95	3.3%	111	3.9%	107	3.8%	313	11.0%
5 cat. (all labelled)	99	3.5%	111	3.9%	120	4.2%	330	11.6%
6 cat. (all labelled)	118	4.1%	100	3.5%	113	4.0%	331	11.6%
7 cat. (all labelled)	103	3.6%	120	4.2%	88	3.1%	311	10.9%
10 cat. all labelled	115	4.0%	115	4.0%	116	4.1%	346	12.1%
10 cat. end labelled	103	3.6%	103	3.6%	103	3.6%	309	10.8%
11 cat. all labelled	103	3.6%	92	3.2%	106	3.7%	301	10.6%
11 cat. end labelled	116	4.1%	92	3.2%	93	3.3%	301	10.6%
Overall	961	33.7%	944	33.1%	945	33.2%	2850	100.0%

Note. Percentages of the whole sample used in the analysis are reported. Pearson's chi-squared test of independence between conditions: $\chi^2(16, N = 2850) = 15.017$, $p = 0.523$.

Procedure

The study included several parts, starting with instructions delivered in the form of short videos (see Table S8), followed by a Polish adaptation (Muszyński et al., 2025) of the cognitive abilities test Pathfinder (Malanchini et al., 2021), tracked by several survey scales, manipulation check (questions about instruction comprehension) and a set of additional self-reported measures: responding diligence, interest in the study, and survey difficulty scales. The median time of completing the survey was about 30 minutes. The order of measurement and manipulations used in the study is displayed in Table S1. All participants responded to the measures in the same order.

Design

The main model for comparing groups is the Analysis of variance (ANOVA) conducted in the 9 x 3 between-participants design. The first factor had nine levels corresponding to the number and labelling of response categories (3, 4, 5, 6, 7, 10, and 11 all-labelled, plus 10 and 11 end-labelled). The second factor comprised three levels corresponding to three types of experimental instruction (standard, warning, and appeal).

We started by testing whether there were interaction effects between two factors. Across 34 models we found only a single one in which the interaction was statistically significant at the 0.05 level (vertical acceleration – joint analysis of all scales), but was no longer significant after applying Holm's correction for multiple comparisons. We therefore concluded that the instruction condition did not meaningfully moderate the response scale length and format effects and decided to report in the paper results from the models including only the main effects (results

of the omnibus *F* tests from the models including interaction terms, as well as tests regarding the instruction main effects, are available in the Supplementary Online Materials Table S6).

Power analysis indicated that the achieved power was as high as 0.90, even for small interaction effect sizes of *f=0.1*. This was calculated using a conventional 0.05 significance level and the denominator degrees of freedom ranging from 2445 to 2823 (depending on the number of observations available for a given measure (Faul et al., 2007).

Careless responding indices

We used three C/IER indices: (1) straightlining, (3) Mahalanobis distance, (4) survey screen time. This set covers response pattern, inconsistency, and response time analysis, main subtypes of C/IER analysis (Meade & Craig, 2012). Mahalanobis distance indicates participant distance from the data centre in a multivariate normal distribution and helps to identify outliers. The straightlining index was operationalised as the longest streak of identical survey responses. This index was chosen due to its simplicity of calculation and interpretation. To reduce the risk of confusion between valid and invalid straightlining (Reuning & Plutzer, 2020; Schonlau & Toepoel, 2015), most of our scales included reverse-coded items. Time spent on each of the analysed survey screens was logged using the natural logarithm to account for skewness. A larger number of failed attention checks, more straightlining, larger Mahalanobis distance and very short response times were assumed as indicating C/IER.

Cursor Movement Indicators

As the cursor's position was close to the eye gaze on the screen (Kirsh & Joy, 2020) the cursor moves can serve as an ersatz of eye-tracking and provide additional information on survey behaviour. In this study, we analysed mouse movement indicators based on metrics proposed by

Pokropek et al. (2023), inspired by previous studies (Horwitz et al., 2017; Stieger & Reips, 2010). Specifically, we focused on vertical movement indicators, including vertical speed, vertical acceleration, and the number of vertical direction changes (flips):

Mean vertical speed (vY): Measured in pixels per second, with higher values interpreted as indicative of rapid responding, potentially signalling C/IER.

Mean absolute vertical acceleration (aY): Measured in pixels per squared second, with higher values indicative of C/IER.

Number of vertical direction changes (flipsY): A low number of flips was associated with straightlining behaviour, indicative of C/IER, while a high number of flips is related to tendencies to reconsider previous items, suggesting more careful consideration of answers or even hesitancy/indecisiveness.

To address non-normal distributions in the data, we transformed the indices: the number of flips was square root-transformed, while vertical speed and acceleration were log-transformed.

Self-reported Task Engagement, Interest and Subjective Difficulty

We used three self-reported indices to assess task engagement, interest in the survey task, and task difficulty to measure participants' subjective opinions on the survey response process.

Task Involvement. This scale measured participants' self-reported engagement and attentiveness by items like "I carefully read each question in this survey." or "I completed this survey in a hurry." The scale comprised nine items (including five reversed positions) and employed a 5-point Likert-type response scale.

Interest. This scale measured participants' self-reported interest with items like "I

enjoyed participating in this survey." or "I was bored while participating in this survey." The scale comprised eight items (including two reversed items) and employed a 5-point Likert-type response scale.

Subjective task difficulty. This scale measured participants' self-reported interest with items like "Filling out this survey was difficult." or "Filling out this survey brought me satisfaction." The scale comprised six items (including one reversed) and employed a 5-point Likert-type response scale. This instrument was modelled after a Task Loading Index (TLX; Hart & Staveland, 1988).

Multidimensional Nominal Response Models

To assess the effect of response scale length and format on response scales (RS), we used a generalisation of the Multidimensional Nominal Response Models (MNRM) with a priori specified response style patterns (Falk & Cai, 2016; Wetzel & Carstensen, 2017). A separate model was estimated for each response scale length and format (see Supplementary Online Materials for detailed specification).

Results

Respondent Behaviours: engagement and attentiveness indices

Table 3 presents the results of the ANOVA analyses for C/IER indices across four scales: Vaccinations (va), Institutional trust (it), Reading attitudes (ra), and Reading habits (rh), as well as for all scales analysed jointly (all).

Table 3

ANOVA results for C/IER indices across response scale formats

Index	Scale	N	F	p	Adjusted p	partial η^2
log(screen time)	va	2850	10.32	0.000	0.000	0.027
	it	2850	16.72	0.000	0.000	0.044
	ra	2850	9.09	0.000	0.000	0.028
	rh	2850	10.24	0.000	0.000	0.025
	all	2850	14.43	0.000	0.000	0.039
straightlining	va	2850	5.07	0.000	0.000	0.015
	it	2850	7.38	0.000	0.000	0.017
	ra	2850	13.22	0.000	0.000	0.031
	rh	2850	6.12	0.000	0.000	0.018
	all	2850	3.92	0.000	0.003	0.012
Mahalanobis	va	2850	0.18	0.994	1.000	0.000
	it	2472	0.28	0.973	1.000	0.001
	ra	2850	0.36	0.941	1.000	0.001
	rh	2850	0.14	0.997	1.000	0.000
	all	2472	0.12	0.998	1.000	0.000

Note. va -Vaccinations; it - Institutional trust; ra - Reading attitudes; rh - Reading habits; all - all four scales analysed together. Adjusted *p* - Holm-corrected (correction was applied considering all the tests reported in tables 3-5).

Significant differences were observed for log-transformed screen time across all individual scales and the combined scales (all). The strongest effect was observed for Institutional trust ($F(1, 2849) = 16.72$, $p < .001$, adj. $p < .001$, partial $\eta^2 = .044$), followed by the combined scales ($F(1, 2849) = 14.43$, $p < .001$, adj. $p < .001$, partial $\eta^2 = .039$). For other scales, the effect was slightly weaker, with partial η^2 ranging from 0.025 to 0.028, but still statistically significant at the 0.001 level, also after applying Holm's correction. Longstring analysis also showed significant effects, particularly for Reading attitudes ($F(1, 2849) = 13.22$, $p < .001$, adj. p

< .001, partial η^2 = .031) and Institutional trust ($F(1, 2849) = 7.38$, $p < .001$, adj. $p < .001$, partial η^2 = .017). The weakest effect was found for the combined scales ($F(1, 2849) = 3.92$, $p = .003$, adj. $p < .012$, partial η^2 = .012). Mahalanobis distance did not yield significant results for any scale or the combined scales (all $p > .05$, adjusted $p > .999$, partial $\eta^2 < .001$).

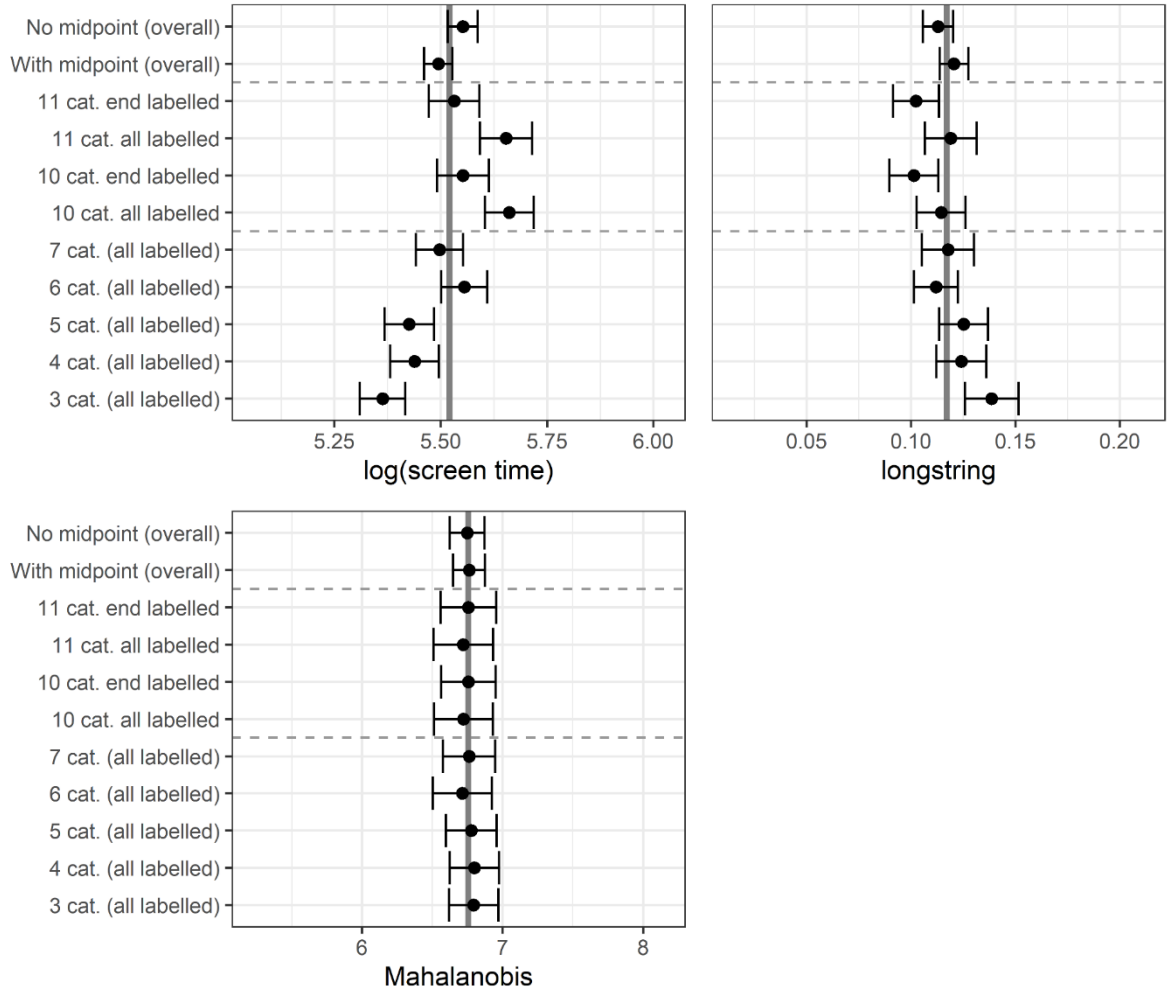
The Partial Prediction Plots in Figure 2 illustrate the relationships between C/IER indices and the response scale format for the combined scales. The results indicate that respondents required less time to complete scales with fewer response options. Scales with more response categories took the most time to complete, which aligns with the expectation that these scales involve more processing. In case of 10- and 11-point scales full labelling caused longer response times than end labelling - probably an indicator of time that participants needed to process the labels.

For the straightlining index, the highest values were observed for shorter response scales, while the lowest values occurred for scales with larger numbers of categories. This pattern suggests that respondents may have been less attentive and more likely to repeat their answers when responding to shorter scales. However, this interpretation should be cautiously approached, as it is inherently easier to produce repetitive responses with only three categories than with eleven.

Additionally, in Figure 2, average results for scales with a midpoint (3-, 5-, 7-, 11-point) were provided compared to those without a midpoint (4-, 6-, 10-point). There were no significant differences in C/IER indices between these two scales, suggesting that the presence or absence of a midpoint does not substantially affect respondent inattentive behaviour (at least as measured by this set of indices).

Figure 2

Partial prediction plots of C/IER indices for the combined scales across response scale formats



Note. The grey vertical line denotes the mean across conditions.

Table 4, analogous to Table 3, presents the ANOVA results for cursor movement indices corrected for screen size.

Table 4

ANOVA results for cursor movement indices across response scale formats

Index	Scale	N	F	p	Adjusted p	partial η^2
vertical speed	va	1827	1.69	0.095	1.000	0.007
	it	1495	1.46	0.167	1.000	0.007
	ra	2477	1.10	0.356	1.000	0.004
	rh	2096	1.67	0.102	1.000	0.007
	all	1137	1.84	0.065	1.000	0.011
vertical acceleration	va	1827	1.01	0.428	1.000	0.005
	it	1495	2.24	0.022	0.470	0.013
	ra	2477	2.30	0.019	0.408	0.007
	rh	2096	1.48	0.161	1.000	0.005
	all	1137	2.02	0.041	0.737	0.015
number of vertical flips	va	1827	1.18	0.308	1.000	0.005
	it	1495	1.20	0.294	1.000	0.006
	ra	2477	9.47	0.000	0.000	0.029
	rh	2096	2.19	0.025	0.484	0.008
	all	1137	1.12	0.349	1.000	0.007

Note. va -Vaccinations; it - Institutional trust; ra - Reading attitudes; rh Reading habits; all - All 4 scales together. Adjusted *p* - Holm-corrected (correction was applied considering all the tests reported in tables 3-5).

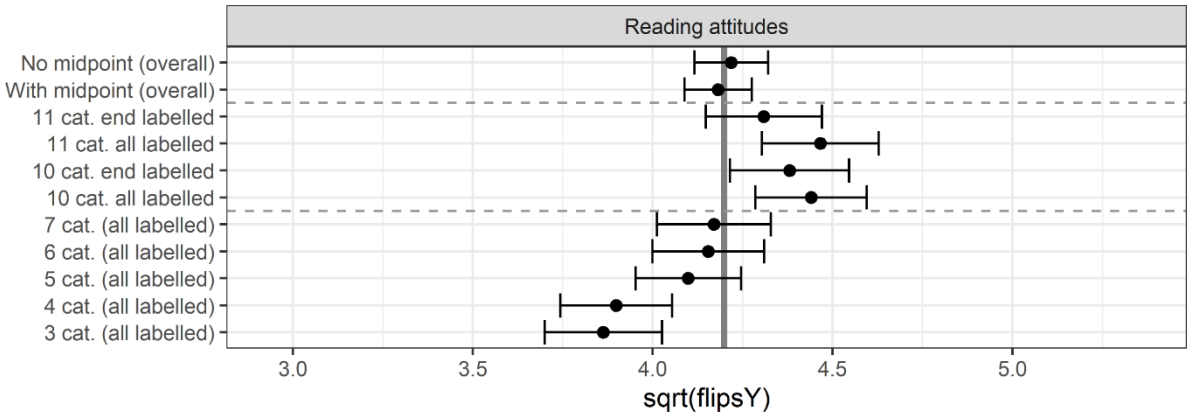
In most cursor movement indices, no significant differences were found after adjusting p-values for multiple comparisons using Holm's method. The only significant result (according to adjusted p-values) occurred for the number of vertical flips in the Reading attitudes scale ($F(1, 2476) = 9.47$, $p < .001$, adj. $p < .001$, partial $\eta^2 = .029$). For all other indices and scales, adjusted p-values were non-significant, suggesting that cursor movement patterns were largely consistent across scales of different formats.

The significant result for the number of vertical flips in the Reading attitudes scale is illustrated in Figure 3. The plot suggests that respondents made fewer direction changes on

shorter scales, indicating less deliberation and fewer revisions. This pattern is consistent with the idea that assessing one's opinions on a simpler, shorter response scale is easier, while longer scales require more effort to position oneself accurately. Additionally, longer scales may lead to more frequent mistakes and, consequently, a higher likelihood of making corrections.

Figure 3

Number of vertical flips in reading attitudes scale across response scale formats



Note. The grey vertical line denotes the mean across conditions.

Subjective experiences

Table 5 summarises the ANOVA results for self-reported engagement measures, including engagement, attentiveness, interest, and burden, analysed across all conditions.

Table 5

ANOVA Results for self-reported engagement measures across response scale formats

Self-reported	N	F	p	Adjusted p	partial η^2
Engagement	2850	2.54	0.009	0.216	0.007
Attentiveness	2850	1.57	0.128	1.000	0.004
Interest	2850	1.24	0.270	1.000	0.003
Burden	2850	2.23	0.022	0.470	0.006

Note. Adjusted *p* - Holm-corrected (correction was applied considering all the tests reported in tables 3-5).

These results suggest that self-reported measures of engagement, attentiveness, interest, and burden were largely consistent across conditions, with no meaningful differences observed.

Response Styles

Table 6 presents the fit indices comparing models with and without response styles (RS) across various response scale designs. The analysis utilised three fit indices: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and sample-adjusted BIC. The additional variance explained by response style traits (ERS and MRS) was also calculated as an index of RS intensity (Scharl & Gnambs, 2024). For scales with 3 and 4 response categories, only ERS is specified due to the limited number of categories, which restricts the feasibility of more complex modelling. The results indicate that models incorporating response styles consistently provided a better fit across all scales, as shown by AIC, BIC, and sample-adjusted BIC values.

Table 6

Fit of models with and without RS

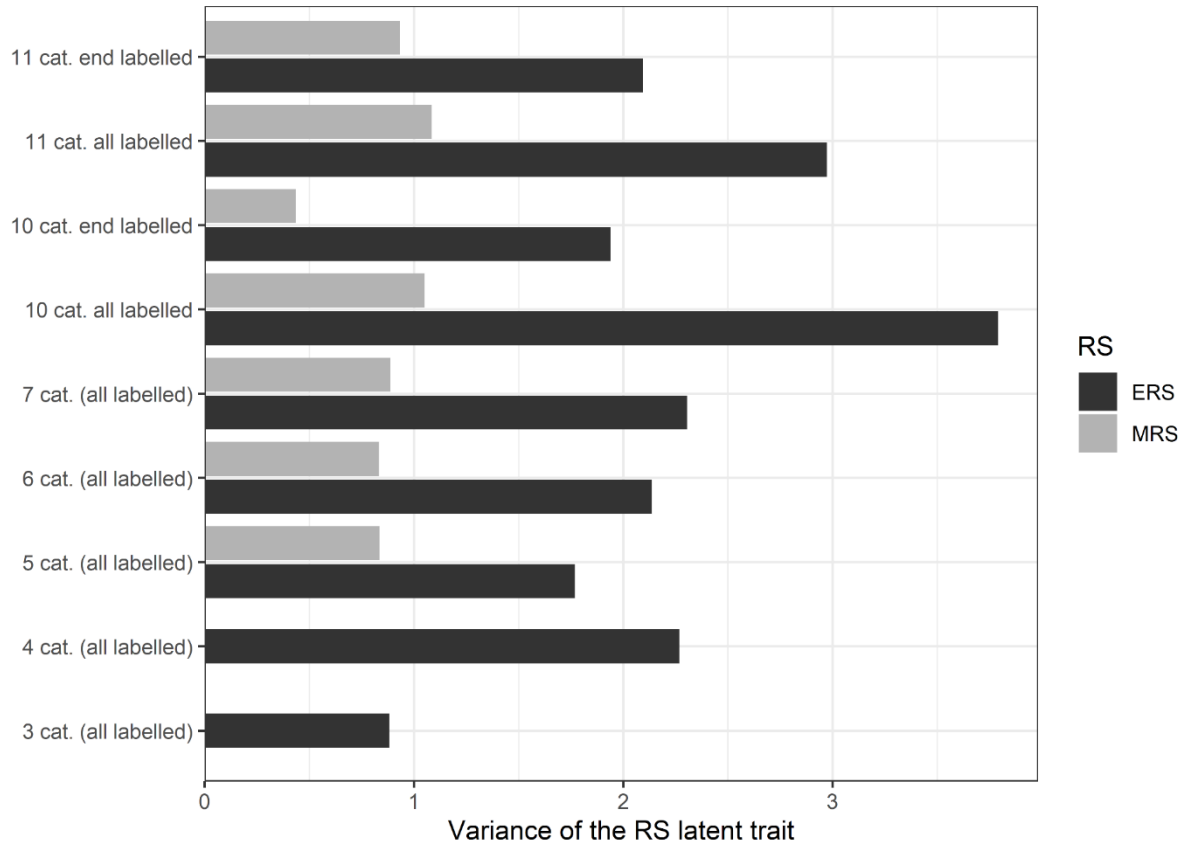
Scale	N	Model	AIC	BIC	SABIC	Var(ERS)	Var(MRS)
3 cat. (all labelled)	308	Without RS	18,821.4	19,190.6	18,876.6	-	-
		With RS	**18,020.1**	**18,258.8**	**18,055.8**	0.88	-
4 cat. (all labelled)	313	Without RS	24,120.4	24,648.6	24,201.4	-	-
		With RS	**22,773.4**	**23,170.5**	**22,834.3**	2.27	-
5 cat. (all labelled)	330	Without RS	31,204.0	31,899.2	31,318.7	-	-
		With RS	**29,255.7**	**29,688.8**	**29,327.2**	1.77	0.84
6 cat. (all labelled)	331	Without RS	36,115.8	36,971.3	36,257.6	-	-
		With RS	**33,805.2**	**34,398.3**	**33,903.5**	2.14	0.83
7 cat. (all labelled)	311	Without RS	37,813.3	38,808.0	37,964.4	-	-
		With RS	**35,867.5**	**36,604.2**	**35,979.4**	2.30	0.89
10 cat. all labelled	346	Without RS	51,757.0	53,268.7	52,022.0	-	-
		With RS	**49,084.8**	**50,331.0**	**49,303.2**	3.79	1.05
10 cat. end labelled	309	Without RS	46,617.1	48,084.3	46,837.9	-	-
		With RS	**44,624.5**	**45,834.1**	**44,806.5**	1.94	0.44
11 cat. all labelled	301	Without RS	47,190.4	48,799.3	47,422.9	-	-
		With RS	**44,672.5**	**46,025.6**	**44,868.0**	2.97	1.08
11 cat. end labelled	301	Without RS	46,413.3	48,018.5	46,645.3	-	-
		With RS	**44,189.4**	**45,538.8**	**44,384.4**	2.09	0.93

Note. Values for better fitting models are bolded. For response scales of 3 and 4 categories, column Var(ERS) reports the variance of the latent trait expressing both ERS and MRS.

Figure 4 illustrates the intensity of response styles (RS), measured as the variance of RS latent traits (ERS and MRS), across response scales of varying formats.

Figure 4

Intensity of RS, measured as a variance of RS latent traits, across response scale formats



Note. For scales with 3 and 4, only ERS is specified because the limited number of categories does not allow for modelling more RS.

The results indicate that extreme response styles (ERS) consistently exhibit higher variance than midpoint response styles (MRS) across all response scale formats. For fully labelled response categories, longer response scales - particularly those with 10 and 11 categories - tend to show higher ERS variances. In contrast, shorter response scales demonstrate lower variances. However, this pattern is disrupted by the results for the end-labelled 10- and 11-category scales, where ERS variance is comparable to scales with fewer response categories.

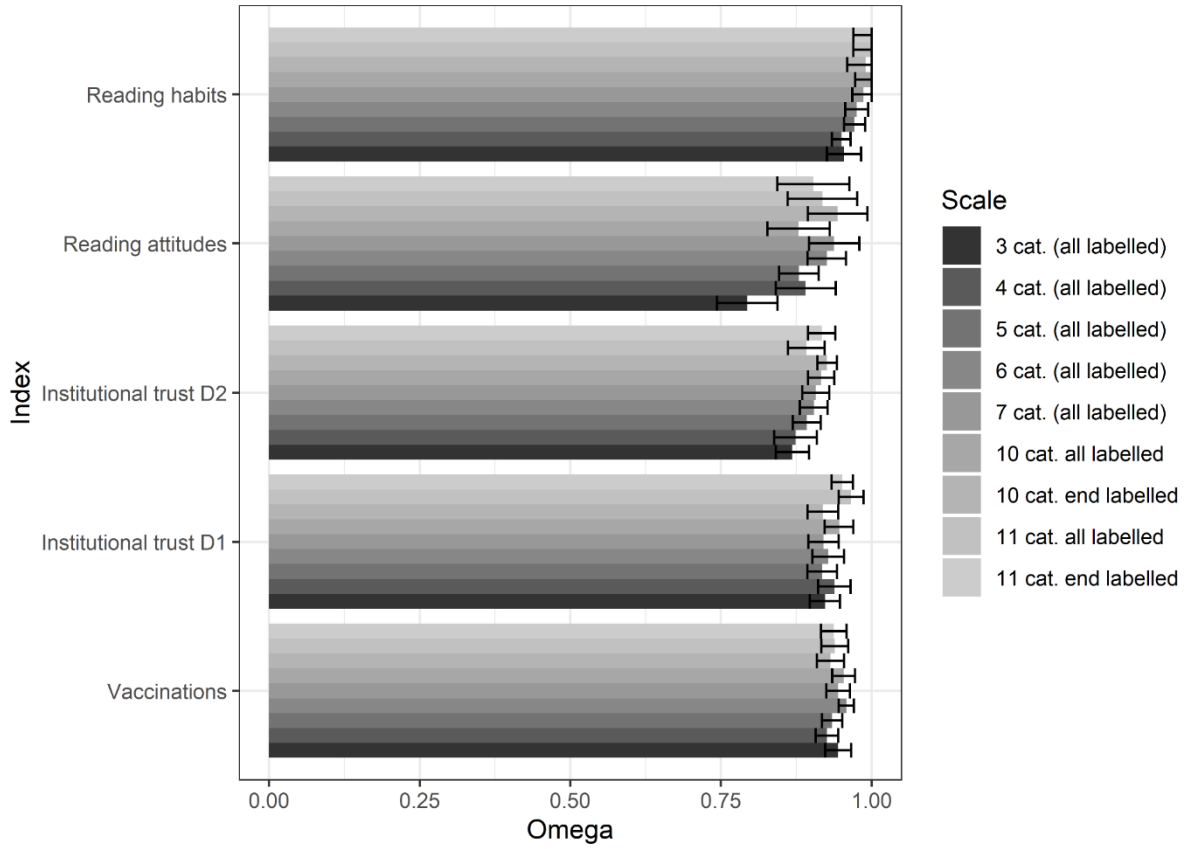
MRS variances are generally smaller and relatively consistent across most scale formats, except for the end-labelled 10-category scale, which exhibits a notably lower variance than other analysed response scales. Despite some observable tendencies, the overall picture is mixed, and it seems that similar response styles and to a similar degree, are present across all response scale formats.

Reliability of scales

Figure 5 presents the reliability estimates of the analysed scales across different response scale formats. Omega reliability coefficients and their confidence intervals (CIs) were estimated using a parametric bootstrap (method 'bsi' in the function ci.reliabilities() from the MBESS R package, Kelley et al., 2018). For the institutional trust scale, listwise deletion was applied whenever respondents endorsed an additional response category indicating that they do not know a given institution.

Figure 5

Reliability Omega Coefficients with Confidence Intervals for Scales Across Different Response Scale Formats



The differences in reliability, as measured by Omega coefficients (Zinbarg et al., 2005), are small and scale-dependent. For some scales (reading attitudes, institutional trust on the second dimension), increasing the number of response categories improved reliability. The highest differences were visible on the Reading Attitudes scale, which had the fewest items. Reliability was lower with three categories (Omega = 0.794), but increased significantly as the number of response categories grew, reaching 0.944 with ten end-labelled categories. Still, for other measurement tools, differences are not statistically significant. The differences in reliability

between scales with all response categories labelled and those with only the endpoints labelled were minimal. For most scales, the Omega values were very similar across the two labelling methods, suggesting that the choice of labelling has a negligible impact on reliability (see Supplementary Online Materials Table S6 for details).

Validity

In order to assess any differences in construct validity, we estimated a set of linear regression models in which we investigated pairwise relationships between sum-score indices computed for each of the five scales discussed in the previous section and between these indices and the variables used to assess validity, which were also part of the procedure:

- The sum-score index of generalised trust computed given responses to three items, initially included in the European Social Survey (ESS). The scale and format of the response scale differed according to the experimental conditions. Still, the item format differed slightly: each question was shown on a separate survey screen, and the response scale was anchored by two specific statements at each end.
- The sum-score index of attitudes towards immigrants scale (originally included in the ESS round 1 and 7 questionnaire). Response scale format again varied according to the experimental condition.
- Sum-score of the Pathfinder general cognitive ability test.
- Respondent's age.
- Respondent's education level, coded binary as tertiary versus secondary and primary.

Overall, there were 45 such models. In each of them, we controlled for the differences in response scale length and format and type of instruction given to the respondents at the

beginning of the survey using a set of dummy variables. Moreover, we included the interaction of the dependent variable with response scale length and format and the interaction between the dependent variables and the type of instruction. To assess whether response scale length and format affect the results of substantial analyses, we examined the statistical significance of this first interaction.

Out of 45 comparisons, we found seven (i.e. ~15%) statistically significant interactions at the 0.05 level. However, after applying Holm's correction for multiple comparisons, none of the interactions remained statistically significant at the 0.05 level. The largest observed difference in slope estimates across all comparisons was 0.266, measured in the metric of a standardised regression coefficient. This suggests that while minor differences in results may occur, the number of response categories does not significantly influence the substantive relationships examined in this analysis (see Supplementary Online Materials Table S7 for details).

Discussion

Implications for Engagement and Response Styles

First, differences in engagement and response behavior across response scale formats were small. However, longer scales took more time and showed more vertical flips—indicating greater cognitive effort. This supports Hypothesis 1: participants spent more time on scales with more categories. Labelled 10- and 11-point scales also led to longer response times than unlabelled ones, further confirming Hypothesis 1.

Nonetheless, in contrast to what we expected, these differences did not translate into major variations in self-reported engagement or subjective task difficulty. Hypothesis 2 was not

confirmed then, as participants reported the same level of burden, attentiveness, interest, and engagement regardless of the response scale they have used. It seems that the longer response time in case of scales with more response categories is just related to processing labels and not additional cognitive problems related to the number of response categories (Menold et al., 2014). Previous research suggested that too few labels to rely on, rather than too many categories to choose from, is a major problem for survey respondents (Gummer & Kunz, 2021).

In contrast to Hypothesis 3, we did not identify systematic relations between employed C/IER indices and type of response scale. The scale with the fewest response categories (3-point) was associated with shorter response times and higher tendencies for longstring behavior. The presence of a neutral midpoint in scales with odd number of response categories did not appear to significantly influence engagement indices, suggesting that respondents adapt equally well to both even- and odd-numbered response scales.

The investigation of response styles (RS) showed no consistent evidence that scale length significantly affects tendencies toward extreme or midpoint responding, corroborating some previous evidence (Kieruj & Moors, 2010; Kutscher & Eid, 2020). The ERS and MRS seem to be present in all response scales analysed, as indicated by model fit indices (Table 6). The variance of ERS and MRS seems related to the number of response categories, with scales with more categories characterised by more RS variance - both ERS and MRS variance rises from 5-point to 11-point response scales (and from 3- to 4-point). Adding more response options seems to evoke more idiosyncratic response tendencies and more response variability, thus confirming Hypothesis 4a.

The previously obtained result that end-labelled scales evoke more ERS does not seem to be confirmed here (Moors et al., 2014). On the contrary, all-labelled 10- and 11-point response

scales yielded higher variances of ERS and MRS, rejecting Hypothesis 4b. It seems that label processing may also be subjected to a sizable inter-individual variability that covaries with stylistic tendencies.

These results go in-line with claims that RS are more individual-driven idiosyncratic response tendencies that differ much more across respondents than across response scale formats (Ulitzsch et al., 2023).

Implications for Psychometric Properties

Reliability analysis showed that the number of response categories modestly influenced internal consistency, with small, scale-dependent effects. Contrary to expectations, reliability increased only slightly with more categories, with meaningful gains observed only for the shortest (6-item) Reading Attitudes scale. Differences between all- and end-labelled scales were negligible, indicating limited impact of labelling on reliability. These findings partially support Hypothesis 5, which predicted that reliability improvements would plateau after approximately 7 categories. Indeed, we observed that reliability increased only slightly with more categories, with the most substantial gains occurring up to 7 points, after which improvements were marginal. This confirms the expected non-linear pattern, suggesting that longer scales offer only minimal reliability benefits beyond the 7-point threshold, while shorter scales remain sufficiently reliable.

Substantive Conclusions

We showed that the number of response categories does not meaningfully alter the substantive conclusions drawn from regression analyses. While Hypothesis 6 predicted a

curvilinear relationship with validity potentially decreasing for 10-11 point scales, we found no significant differences in validity across scale lengths. This suggests even greater robustness of substantive findings across different response scale formats than anticipated, aligning with some earlier studies (Maydeu-Olivares et al., 2009). Even the largest observed difference in slope estimates (0.266 in standardised terms) was relatively minor, further supporting the robustness of substantive findings across different response scale formats.

Significance of the Study

This study is among the first to systematically examine the interplay between response scale design, respondent behaviour, and substantive results in a large, diverse sample. Our findings suggest that the overall impact of these design choices on substantive research conclusions is minimal. Scales with more response categories may offer higher reliability, but this comes at the price of longer response times.

Limitations and Future Directions

Several limitations should be noted. First, the study was conducted in Poland, limiting generalizability to other cultural contexts. Response styles vary across cultures and languages (He & Van de Vijver, 2015), so caution is needed when applying these findings elsewhere. Second, while we tested multiple scales and conditions, results may not generalize to other constructs or formats. Instruments with fewer items or weaker latent structures may be more sensitive to scale format changes (Maydeu-Olivares et al., 2009). Third, only desktop and laptop users were included, excluding smartphones. This limits external validity, though necessary to capture mouse movement data. Still, recent findings suggest comparable data

quality across devices (Maslovskaya et al., 2024), indicating device effects may be minimal. The web-based format also introduces potential mode effects. Online surveys differ from face-to-face ones in anonymity, social desirability, and respondent attentiveness. These factors may explain mixed findings in the literature, with no scale validity differences found in self-administered modes (Maydeu-Olivares et al., 2009), but some identified in face-to-face contexts (Revilla et al., 2014). Additionally, participants were part of an internet opt-in panel, likely more experienced and familiar with surveys than the general population. This may reduce the generalizability of engagement-related results. Future studies should explore scale use across more diverse populations, considering variables like age and cognitive ability. The questionnaire also covered varied topics (e.g., vaccine attitudes, trust, reading), which could limit applicability to more focused surveys. However, this diversity allowed testing across constructs with different emotional and social desirability profiles, reducing topic-related bias. Lastly, the study did not address acquiescence bias—respondents' tendency to agree regardless of content—which can distort findings, especially in agree–disagree formats (Krosnick & Presser, 2010). Although not analyzed here, its potential impact on data quality warrants attention in future work.

Conclusion & Recommendations

This study provides empirical evidence on Likert-type scale design in social science research. While longer scales (7–11 points) show slightly higher reliability and more complex response patterns, these differences are small and do not meaningfully affect substantive conclusions. Even the largest observed differences in standardized regression coefficients across formats were minor, suggesting limited influence on construct validity. Only dichotomous scales

consistently showed lower measurement quality (Lundmark et al., 2016).

Our results confirm this non-linear pattern, particularly for reliability, where improvements beyond 7 categories were minimal, as predicted in our revised Hypothesis 5 (Lundmark et al., 2016; Preston & Colman, 2000; Rakhshani et al., 2025; Revilla et al., 2014; Weijters et al., 2010). Fewer than five options may reduce variability and precision, while more than seven increase cognitive load without added benefit (Revilla et al., 2014). Longer scales may also increase fatigue, especially in phone surveys or among less motivated respondents (Krosnick, 1991; Dawes, 2008), and very short scales (2–3 points) often lead to skewed distributions and ceiling/floor effects (Rakhshani et al., 2024). These may require more complex modelling.

Although more response options increased response time—likely due to label processing—this did not affect self-reported engagement or perceived difficulty. We also found no consistent link between response scale length and careless responding, suggesting that factors like motivation and cognitive ability play a bigger role (Krosnick, 1991).

We recommend 5- to 7-point fully labelled scales for most contexts, balancing quality, usability, and modelling ease. Scale choice should reflect the construct, survey mode, and target population.

Beyond psychometric considerations, practical implementation factors strongly favor shorter scales. With the increasing prevalence of mobile devices in survey research (Gummer et al., 2023), longer scales often require scrolling on small screens, potentially increasing respondent burden and breakoff rates (Couper & Peterson, 2017). This is particularly problematic for 10- and 11-point scales, which may not display fully even on tablets. Furthermore, mixed-mode surveys face additional challenges: scales that display horizontally on

paper questionnaires may need vertical formatting on screens, potentially affecting response comparability across modes (Singh, 2023). Displaying response category labels on small screens becomes problematic when scales are long. However, reducing the number of labels to accommodate screen size can negatively affect measurement quality and is therefore not recommended (Gummer & Kunz, 2021).

Longer response scales are also problematic in telephone surveys, as respondents struggle to retain more than 5-7 options in working memory without visual aids (Bowling, 2005; Krosnick & Presser, 2010). These practical considerations, combined with our finding that 5-7 point scales achieve nearly optimal psychometric properties, provide compelling reasons to avoid longer scales unless specific research needs justify the additional complexity. Survey designers should also consider their target population's device usage patterns and potential mode effects when selecting scale lengths, as these implementation factors may have greater impact on data quality than small differences in theoretical reliability.

In sum, while response scale format can affect data quality marginally, it does not decisively shape substantive conclusions. Researchers should prioritize practical factors—such as respondent comfort, mode, and construct characteristics—over the pursuit of a single “optimal” number of categories.

References

Abulela, M. A., & Khalaf, M. A. (2024). Does the number of response categories impact validity evidence in self-report measures? A scoping review. *SAGE Open*, *14*(1).

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, *27*(3), 281–291. https://doi.org/10.1093/pubmed/fdi031

Bråten, I., Skovdahl, O., Anmarkrud, Ø., & Strømsø, H. I. (2025). Does reading still make you smarter? It depends. *Reading and Writing*, 1-22. https://doi.org/10.1007/s11145-025-10668-2

Couper, M. P., & Peterson, G. J. (2017). Why do web surveys take longer on smartphones? *Social Science Computer Review*, *35*(3), 357–377. https://doi.org/10.1177/0894439316629932

Cox III, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 17(4), 407–422.

European Commission. (2019). *Europeans' attitudes towards vaccination: Special Eurobarometer 488 – Data annex (Wave EB91.2)*.

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, 21(3), 328–347. https://doi.org/10.1037/met0000059

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191. https://doi.org/10.3758/BF03193146

Fernández-Fontelo, A., Kieslich, P. J., Henninger, F., Kreuter, F., & Greven, S. (2023). Predicting Question Difficulty in Web Surveys: A Machine Learning Approach

Based on Mouse Movement Features. *Social Science Computer Review*, *41*(1), 141-162. https://doi.org/10.1177/08944393211032950

Finn, J. A., Ben-Porath, Y. S., & Tellegen, A. (2015). Dichotomous versus polytomous response options in psychopathology assessment: Method or meaningful variance? *Psychological Assessment*, 27(1), 184–193.

Flamer, S. (1983). Assessment of the multitrait-multimethod matrix validity of Likert scales via confirmatory factor analysis. *Multivariate Behavioral Research*, 18(3), 275–306.

Gummer, T., Höhne, J. K., Rettig, T., Roßmann, J., & Kummerow, M. (2023). Is there a growing use of mobile devices in web surveys? Evidence from 128 web surveys in Germany. *Quality & Quantity*, *57*(6), 5333-5353.

Gummer, T., & Kunz, T. (2021). Using only numeric labels instead of verbal labels: Stripping rating scales to their bare minimum in web surveys. *Social Science Computer Review*, 39(5), 1003–1029.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research.

He, J., & Van de Vijver, F. J. R. (2015). Effects of response styles on cross-cultural comparisons: The role of context, sample characteristics, and question characteristics. *International Journal of Psychology*, 50(6), 456–468. https://doi.org/10.1002/ijop.12167

Hilbert, S., Küchenhoff, H., Sarubin, N., Nakagawa, T. R., & Bühner, M. (2016). The influence of the response format in a personality questionnaire: An analysis of a dichotomous, a Likert-type, and a visual analogue scale. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, 23(1).

Horwitz, R., Kreiner, S., & Christensen, K. B. (2017). Analyzing process data using item

response theory. *Psychometrika*, 82(4), 1082–1103. https://doi.org/10.1007/s11336-017-9555-5

Jones, W. P., & Loe, S. A. (2013). Optimal number of questionnaire response categories: More may not be better. *SAGE Open*, 3(2), 2158244013489691.

Kelley, K., Kelley, M. K., & Imports, M. A. S. S. (2018). The MBESS R package. [Computer software]. MBESS. Retrieved from https://CRAN.R-project.org/package=MBESS

Kieruj, N. D., & Moors, G. (2010). Variations in response style behavior by response scale format in attitude research. *International Journal of Public Opinion Research*, 22(3), 320–342.

Kirsh, I., & Joy, M. (2020, July). Exploring pointer assisted reading (PAR): Using mouse movements to analyze web users' reading behaviors and patterns. In *International Conference on Human-Computer Interaction* (pp. 156–173). Springer International Publishing.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3), 213–236.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, 1996(70), 29–44. https://doi.org/10.1002/ev.1033

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), Handbook of survey research (2nd ed., pp. 263–314). Emerald Group Publishing.

Kutscher, T., & Eid, M. (2020). The effect of rating scale length on the occurrence of inappropriate category use for the assessment of job satisfaction: An experimental online study. *Journal of Well-Being Assessment*, 4, 1–35.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, 4(2), 73–79.

Lundmark, S., Gilljam, M., & Dahlberg, S. (2016). Measuring generalized trust: An examination of question wording and the number of scale points. *Public Opinion Quarterly*, *80*(1), 26-43.

Malanchini, M., Rimfeld, K., Gidziela, A., Cheesman, R., Allegrini, A. G., Shakeshaft, N., ... & Plomin, R. (2021). Pathfinder: A gamified measure to integrate general cognitive ability into the biological, medical, and behavioural sciences. *Molecular Psychiatry*, 26(12), 7823–7837.

Maslovskaya, O., Smith, P. W., & Durrant, G. (2024). Do respondents using smartphones produce lower quality data? Evidence from the first large-scale UK mixed-device survey–Understanding Society Wave 8. *International Journal of Social Research Methodology*, 1-14.

Mathews, C. O. (1929). The relationship between number of response categories and reliability in attitude measurement. *Journal of Educational Psychology*, 20(6), 436–440.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–455. https://doi.org/10.1037/a0028085

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

Menold, N., & Tausch, A. (2016). Measurement of latent variables with different rating scales: Testing reliability and measurement equivalence by varying the verbalization and

number of categories. *Sociological Methods & Research*, 45(4), 678–699.

Moors, G., Kieruj, N. D., & Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology*, 44(1), 369–399.

Muszyński, M., Banasik-Jemielniak, N., Żółtak, T., Rimfeld, K., Shakeshaft, N. G., Schofield, K. L., ... & Pokropek, A. (2025). Moving intelligence measurement online: adaptation and validation of the Polish version of the Pathfinder general cognitive ability test. *Quality & Quantity*, 1-26. https://doi.org/10.1007/s11135-025-02254-z

OECD. (2010). *PISA 2009 Results: Learning to Learn – Student Engagement, Strategies and Practices (Volume III)*. http://dx.doi.org/10.1787/9789264083943-en

Pokropek, A., Żółtak, T., & Muszyński, M. (2023). Mouse chase: Detecting careless and unmotivated responders using cursor movements in web-based surveys. *European Journal of Psychological Assessment*, 39(4), 299–306.

Pokropek, A., Żółtak, T., & Muszyński, M. (2024). Identifying careless responding in web-based surveys: Exploiting sequence data from cursor trajectories and approximate areas of interest. *Zeitschrift für Psychologie, 232*(2), 95–105. https://doi.org/10.1027/2151-2604/a000518

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104(1), 1–15.

Rakhshani, A., Donnellan, M. B., Roberts, B. W., & Lucas, R. E. (2024). Brief Report: Does the Number of Response Options Matter for the BFI-2? Conceptual Replication and Extension. *Assessment, 31*(4), 855-862.

Reuning, K., & Plutzer, E. (2020). Valid vs. invalid straightlining: The complex relationship between straightlining and data quality. *Survey Research Methods, 14*(5), 439–459. https://doi.org/10.18148/srm/2020.v14i5.7766

Revilla, M., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research*, 43(1), 73–97.

Scharl, A., & Gnambs, T. (2024). The impact of different methods to correct for response styles on the external validity of self-reports. *European Journal of Psychological Assessment*, 40(1), 13–21. https://doi.org/10.1027/1015-5759/a000731

Schoenmakers, M., Tijmstra, J., Vermunt, J., & Bolsinova, M. (2024). Correcting for extreme response style: Model choice matters. *Educational and Psychological Measurement*, 84(1), 145–170.

Schonlau, M., & Toepoel, V. (2015). Straightlining in web survey panels over time. *Survey Research Methods, 9*(2), 125–137. https://doi.org/10.18148/srm/2015.v9i2.6128

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, 31(4), 557.

Singh, R. K. (2023). *ESS mode change: Response scale report*. GESIS–Leibniz Institute for the Social Sciences. Retrieved from https://europeansocialsurvey.org/sites/default/files/2024-09/ESS_mode_change_response_scale_report.pdf

Stieger, S., & Reips, U. D. (2010). What are participants doing while filling in an online questionnaire? A paradata collection tool and an empirical study. *Computers in Human Behavior*, 26(6), 1488–1495.

Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology*, 7(6), 456–461.

Tijmstra, J., & Bolsinova, M. (2025). Modeling within-and between-person differences in the use of the middle category in Likert scales. *Applied Psychological Measurement,* 01466216251322285.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*(3), 299-314.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Ulitzsch, E., Henninger, M., & Meiser, T. (2024). Differences in response-scale usage are ubiquitous in cross-country comparisons and a potential driver of elusive relationships. *Scientific Reports,* 14(1), 10890.

Ulitzsch, E., Lüdtke, O., & Robitzsch, A. (2023). The role of response style adjustments in cross-country comparisons: A case study using data from the PISA 2015 questionnaire. *Educational Measurement: Issues and Practice*, 42(3), 65–79.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. International Journal of Research in Marketing, 27(3), 236–247

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test–retest reliability. *Educational and Psychological Measurement*, 64(6), 956–972.

Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*, 33(5), 352–364.

https://doi.org/10.1027/1015-5759/a000291

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response styles and acquiescence over time and their effects on personality trait change. *Journal of Research in Personality*, 67, 87–100. https://doi.org/10.1016/j.jrp.2016.04.005

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7

List of Tables and Figures

Supplementary materials to: Do you agree? Do you strongly agree? The effect of the number of response categories on response processes and verification of substantive hypotheses

Experimental procedure

Table S1 Schematic outlook of experimental procedure

Ordinal number	Measurement	Screens	Response scale length/format manipulation	Median Time
1.	Movies with instruction manipulation	1	No	2 minutes
2.	Cognitive abilities test Pathfinder	Many	No	15 minutes
3.	Survey: vaccinations	2	Yes	15 minutes
4.	Survey: general trust (source: ESS)	3	Yes	
5.	Survey: institutional trust (adapted from ESS)	2	Yes	
6.	Survey: reading attitudes (source: PISA)	1	Yes	
7.	Survey: reading materials (source: PISA)	2	Yes	
8.	Manipulation check	1	No	
9.	Self-reported survey diligence and comments	4	No	
10.	Debriefing	1	No	

Note: ESS: European Social Survey; PISA: Programme for International Student Assessment

Sample dropout

Table S2 Comparison of characteristics of respondents who dropped out of the sample during the Pathfinder test and respondents who proceeded to the questionnaire

Characteristic	Value	N		Pct.		Test
		Dropped	Passed	Dropped	Passed	
Gender	Male	218	1,735	51.8%	46.9%	$\chi^2(1)=3.487$, p=0.062
Gender	Female	203	1,968	48.2%	53.1%	
Age	18-29	129	911	30.6%	24.6%	$\chi^2(2)=7.445$, p=0.024
Age	30-39	132	1,230	31.4%	33.2%	
Age	40-50	160	1,562	38.0%	42.2%	
Education	Secondary	229	2,008	54.4%	54.2%	$\chi^2(1)=0.000$, p=0.989
Education	Tertiary	192	1,695	45.6%	45.8%	

Table S3 Comparison of characteristics of respondents who dropped out of the sample during the questionnaire or were excluded because of excessively short/long survey completion time, with respondents included in the analyses

Characteristic	Value	N		Pct.		Test
		Dropped or excluded	Passed	Dropped or excluded	Passed	
Gender	Male	242	1,319	50.0%	46.3%	$\chi^2(1)=2.152$, p=0.142
Gender	Female	242	1,531	50.0%	53.7%	
Age	18-29	152	683	31.4%	24.0%	$\chi^2(2)=24.189$, p=0.000
Age	30-39	176	921	36.4%	32.3%	
Age	40-50	156	1,246	32.2%	43.7%	
Education	Secondary	278	1,535	57.4%	53.9%	$\chi^2(1)=1.994$, p=0.158
Education	Tertiary	206	1,315	42.6%	46.1%	

Sample exclusions in analyses including cursor moves indices

Analyses including cursor moves indices could be performed only for the respondents for whom paradata collection was successful on a given survey screen(s). The reasons for which paradata collection could fail were:

1. Respondent spent more than 2 minutes answering the survey screen, i.e. span that exceeded the maximum time for log-data collection.
   The amount of cursor moves data grows very fast with time and at some point it may become so large that its sheer size (in MBs) causes difficulties to process it and to send it over the Internet. These technical difficulties may cause lags in the survey interface and slow down respondent's web browser and computer. To diminish the probability of encountering such problems, on the basis of our previous experience, we decided to limit the period of process data collection to 2 minutes. Please also mind that exceeding this limit for a screen with only 11 survey items means that probably respondents are engaged in off-screen multitasking or have made a pause in the survey anyway.
2. Respondent resized the survey browser window while answering the scale.
3. Respondent left the survey browser window while answering the scale.
4. There were broken records in paradata collected on a given survey screen (i.e. some errors occurred during data transfer between respondent's computer and the server).

Frequencies of these problems are summarised in Table S4 below.

Table S4 Sample exclusions in analyses including cursor moves indices

Problem	**Vaccinations**		**Institutional trust**		**Reading attitudes**		**Reading habits**		**All 4 scales together**	
	N	**Pct.**	**N**	**Pct.**	**N**	**Pct.**	**N**	**Pct.**	**N**	**Pct.**
Over 2 minutes spent on a single survey screen	321	11.3%	255	8.9%	70	2.5%	128	4.5%	570	20.0%
Resized a browser window	67	2.4%	46	1.6%	36	1.3%	40	1.4%	88	3.1%
Left browser window	59	2.1%	71	2.5%	32	1.1%	34	1.2%	139	4.9%
Broken records (transmission errors)	754	26.5%	1193	41.9%	269	9.4%	623	21.9%	1462	51.3%
At least one problem	1023	35.9%	1355	47.5%	373	13.1%	754	26.5%	1713	60.1%

No problems	1827	64.1%	1495	52.5%	2477	86.9%	2096	73.5%	1137	39.9%

Wording of response categories

Table S5 Wording of response scale categories for different response scale lengths and formats used in the experiment (English translation in parenthesis)

4 cat. (all labelled)	**5 cat. (all labelled)**	**6 cat. (all labelled)**	**7 cat. (all labelled)**
Zdecydowanie się NIE zgadzam (*Strongly disagree)*	Zdecydowanie się NIE zgadzam (*Strongly disagree)*	Zdecydowanie się NIE zgadzam (*Strongly disagree)*	Zdecydowanie się NIE zgadzam (*Strongly disagree)*
NIE zgadzam się (*Disagree*)	NIE zgadzam się (*Disagree*)	NIE zgadzam się (*Disagree*)	NIE zgadzam się (*Disagree*)
Zgadzam się (*Agree*)	Ani się nie zgadzam, ani się zgadzam (*Neither disagree nor agree*)	Raczej się NIE zgadzam (*Rather disagree*)	Raczej się NIE zgadzam (*Rather disagree*)
Zdecydowanie się zgadzam (*Strongly agree*)	Zgadzam się (*Agree*)	Raczej się zgadzam (*Rather agree*)	Ani się nie zgadzam, ani się zgadzam (*Neither disagree nor agree*)
	Zdecydowanie się zgadzam (*Strongly agree*)	Zgadzam się (*Agree*)	Raczej się zgadzam (*Rather agree*)
		Zdecydowanie się zgadzam (*Strongly agree*)	Zgadzam się (*Agree*)
			Zdecydowanie się zgadzam (*Strongly agree*)
10 cat. all labelled	**10 cat. end labelled**	**11 cat. all labelled**	**11 cat. end labelled**
Całkowicie się NIE zgadzam (*Entirely disagree*)	Całkowicie się NIE zgadzam (*Entirely disagree*)	Całkowicie się NIE zgadzam (*Entirely disagree*)	Całkowicie się NIE zgadzam (*Entirely disagree*)
Zdecydowanie się NIE zgadzam (*Strongly disagree)*	1	Zdecydowanie się NIE zgadzam (*Strongly disagree)*	1
NIE zgadzam się (*Disagree*)	2	NIE zgadzam się (*Disagree*)	2
W większości się NIE zgadzam (*Mostly disagree*)	3	W większości się NIE zgadzam (*Mostly disagree*)	3

Raczej się NIE zgadzam (*Rather disagree*)	4	Raczej się NIE zgadzam (*Rather disagree*)	4
Raczej się zgadzam (*Rather agree*)	5	Ani się nie zgadzam, ani się zgadzam (*Neither disagree nor agree*)	5
W większości się zgadzam (*Mostly agree*)	6	Raczej się zgadzam (*Rather agree*)	6
Zgadzam się (*Agree*)	7	W większości się zgadzam (*Mostly agree*)	7
Zdecydowanie się zgadzam (*Strongly agree*)	8	Zgadzam się (*Agree*)	8
Całkowicie się zgadzam (*Entirely agree*)	Całkowicie się zgadzam (*Entirely agree*)	Zdecydowanie się zgadzam (*Strongly agree*)	9
		Całkowicie się zgadzam (*Entirely agree*)	Całkowicie się zgadzam (*Entirely agree*)

Multidimensional Nominal Response Models

To assess the effect of response scale length and format on response scales (RS), we used a generalisation of the Multidimensional Nominal Response Models (MNRM) with a priori specified response style patterns (Falk & Cai, 2016; Wetzel & Carstensen, 2017). A separate model was estimated for each response scale length and format. Under the MNRM, the probability of a response in category *k* (among *m* response categories) of item *i* is modelled as:

$$P(Y_i = k|W, \Theta, b_{ik}) = \frac{exp(b_{ik}+w_{ikVac}\theta_{Vac}+w_{ikITr1}\theta_{ITr1}+w_{ikItr2}\theta_{ITr2}+w_{ikRd1}\theta_{Rd1}+w_{ikRd2}\theta_{Rd2}+w_{kERS}\theta_{ERS}+w_{kMRS}\theta_{MRS})}{\sum_{h=1}^{m} exp(b_{ih}+w_{ikVac}\theta_{Vac}+w_{ikITr1}\theta_{ITr1}+w_{ikItr2}\theta_{ITr2}+w_{ikRd1}\theta_{Rd1}+w_{ikRd2}\theta_{Rd2}+w_{hERS}\theta_{ERS}+w_{hMRS}\theta_{MRS})} \quad (1)$$

where b_{ik} is item category intercept parameter, θ_{Vac}, θ_{ITr2}, θ_{ITr2}, θ_{Rd1}, θ_{Rd2} are latent variables describing, respectively views on vaccines, trust in domestic institutions, trust in international institutions, reading attitudes, enjoyment of reading activities and θ_{ERS}, θ_{MRS} are latent traits describing extreme and middle RS. Latent traits are assumed to be multivariate and normally distributed, with variance and covariance parameters being freely estimated and means set to 0. Finally, $w_{ik\circ}$ and $w_{k\circ}$ are weights defined in the so-called scoring matrix. These weights were defined in the following way:

- w_{ikVac}, w_{ikITr1}, w_{ikITr2}, w_{ikRd1}, and w_{ikRd2} were set to $k - 1$ if a given item belonged to a given scale in the questionnaire and were set to 0 otherwise.
- w_{kERS} were set to 1 for the most extreme category at each end of the response scale and set to 0 for all the other categories.
- w_{kMRS} were set to 1 for a single middle category on odd-length response scales and two middle categories on even-length response scales and set to 0 for all the other categories.

For response scales of length 3 and 4 , the model specified according to Equation 1 is not identified because choosing response category/categories that do not indicate ERS inevitably means choosing the one(s) that indicate MRS. Consequently, in these cases, we estimated the model of a simpler form:

$$P(Y_i = k|W, \Theta, b_{ik}) = \frac{exp(b_{ik}+w_{ikVac}\theta_{Vac}+w_{ikITr1}\theta_{ITr1}+w_{ikItr2}\theta_{ITr2}+w_{ikRd1}\theta_{Rd1}+w_{ikRd2}\theta_{Rd2}+w_{kERS}\theta_{ERS})}{\sum_{h=1}^{m} exp(b_{ih}+w_{ikVac}\theta_{Vac}+w_{ikITr1}\theta_{ITr1}+w_{ikItr2}\theta_{ITr2}+w_{ikRd1}\theta_{Rd1}+w_{ikRd2}\theta_{Rd2}+w_{hERS}\theta_{ERS})} \quad (2)$$

It is important to remember that "ERS" in this model includes both ERS and MRS, with "negative ERS" (i.e. reluctance to choose extreme categories of the response scale) being indistinguishable from "positive MRS" (i.e. the tendency to choose the middle categories) and vice versa.

In order to verify whether RS are present in our data, we compared the fit of models including RS with models that do not include RS latent traits and are equivalent to the multidimensional Partial Credit Model:

$$P(Y_i = k|W, \Theta, b_{ik}) = \frac{exp(b_{ik}+w_{ikVac}\theta_{Vac}+w_{ikITr1}\theta_{ITr1}+w_{ikItr2}\theta_{ITr2}+w_{ikRd1}\theta_{Rd1}+w_{ikRd2}\theta_{Rd2})}{\sum_{h=1}^{m} exp(b_{ih}+w_{ikVac}\theta_{Vac}+w_{ikITr1}\theta_{ITr1}+w_{ikItr2}\theta_{ITr2}+w_{ikRd1}\theta_{Rd1}+w_{ikRd2})} \quad (3)$$

Model fit was compared using information criteria indices (AIC, BIC, SABIC). In this analysis, we also assessed how the variance of RS latent traits varied between different response scale formats.

Supplementary results

Respondent Behaviours

Table S6 ANOVA results regarding effects of type of instruction and response scale length and format on cursor movement indices and careless responding indices

Measure	**Scale**	**N**	**Type of instruction**				**Interaction: response scale length and format x type of instruction**			
			F	**p**	**Adj. p**	**partial η^2**	**F**	**p**	**Adj. p**	**partial η^2**
log(vY)	va	1827	2.05	0.129	1.000	0.002	1.01	0.445	1.000	0.008
	it	1495	0.26	0.771	1.000	0.000	0.89	0.578	1.000	0.010
	ra	2477	2.44	0.088	1.000	0.002	0.67	0.825	1.000	0.004
	rh	2096	0.99	0.372	1.000	0.001	1.43	0.120	1.000	0.012
	all	1137	0.16	0.850	1.000	0.000	1.60	0.062	1.000	0.023
log(aY)	va	1827	3.46	0.032	0.920	0.004	1.55	0.074	1.000	0.013
	it	1495	3.17	0.042	1.000	0.004	1.31	0.179	1.000	0.017
	ra	2477	4.81	0.008	0.254	0.004	1.14	0.310	1.000	0.007
	rh	2096	2.52	0.081	1.000	0.002	1.08	0.371	1.000	0.010
	all	1137	1.51	0.220	1.000	0.003	1.68	0.046	1.000	0.027
sqrt(flipsY)	va	1827	0.44	0.644	1.000	0.001	0.83	0.649	1.000	0.007
	it	1495	1.10	0.334	1.000	0.001	1.30	0.188	1.000	0.013
	ra	2477	0.78	0.460	1.000	0.001	1.04	0.412	1.000	0.007
	rh	2096	1.11	0.330	1.000	0.001	0.78	0.711	1.000	0.007
	all	1137	2.21	0.110	1.000	0.004	1.19	0.267	1.000	0.018
log(screen time)	va	2850	0.96	0.384	1.000	0.001	0.55	0.920	1.000	0.003
	it	2850	1.04	0.353	1.000	0.001	0.61	0.881	1.000	0.003
	ra	2850	0.33	0.720	1.000	0.000	0.87	0.608	1.000	0.005
	rh	2850	0.78	0.458	1.000	0.001	1.12	0.325	1.000	0.006

Measure	Scale	N	Type of instruction				Interaction: response scale length and format x type of instruction			
			F	p	Adj. p	partial η^2	F	p	Adj. p	partial η^2
	all	2850	1.67	0.189	1.000	0.001	0.55	0.922	1.000	0.003
longstring	va	2850	1.78	0.170	1.000	0.001	1.29	0.190	1.000	0.007
	it	2850	0.84	0.434	1.000	0.001	0.79	0.701	1.000	0.005
	ra	2850	0.57	0.564	1.000	0.000	1.31	0.183	1.000	0.006
	rh	2850	2.36	0.094	1.000	0.002	0.74	0.754	1.000	0.004
	all	2850	1.23	0.291	1.000	0.001	0.61	0.882	1.000	0.004
Mahalanobis	va	2850	0.33	0.720	1.000	0.000	0.73	0.761	1.000	0.004
	it	2472	0.47	0.625	1.000	0.000	1.54	0.077	1.000	0.009
	ra	2850	1.08	0.339	1.000	0.001	1.25	0.224	1.000	0.008
	rh	2850	4.79	0.008	0.254	0.003	1.23	0.234	1.000	0.007
	all	2472	2.02	0.134	1.000	0.002	1.03	0.424	1.000	0.006
Self-reported engagement		2850	19.42	0.000	0.000	0.013	1.02	0.431	1.000	0.005
Self-reported attentiveness		2850	14.06	0.000	0.000	0.010	1.01	0.439	1.000	0.006
Self-reported interest		2850	10.10	0.000	0.001	0.008	1.21	0.254	1.000	0.007
Self-reported burden		2850	1.89	0.151	1.000	0.001	0.69	0.808	1.000	0.004

Notes: va –Vaccinations; it – Institutional trust; ra – Reading attitudes; rh – Reading habits; all – all four scales analysed together. Adj. p – Holm-corrected (correction was applied separately to type of instruction effects across all the models and to interaction effects across all the models).

Reliability of scales

Table S7 Reliability of scales with 95% confidence intervals

Scale	N	Statistic	Vaccinations	Institutional trust D1	Institutional trust D2	Reading attitudes	Reading habits
3 cat. (all labelled)	308	Omega	0.944	0.923	0.868	0.794	0.954
		CI	(0.923, 0.966)	(0.898, 0.948)	(0.841, 0.896)	(0.743, 0.844)	(0.926, 0.982)
4 cat. (all labelled)	313	Omega	0.926	0.938	0.874	0.891	0.950
		CI	(0.907, 0.945)	(0.911, 0.965)	(0.838, 0.909)	(0.841, 0.940)	(0.934, 0.965)
5 cat. (all labelled)	330	Omega	0.934	0.918	0.892	0.879	0.972
		CI	(0.917, 0.952)	(0.893, 0.943)	(0.869, 0.915)	(0.846, 0.912)	(0.954, 0.989)
6 cat. (all labelled)	331	Omega	0.958	0.928	0.904	0.926	0.975
		CI	(0.946, 0.971)	(0.902, 0.954)	(0.881, 0.927)	(0.894, 0.958)	(0.956, 0.994)
7 cat. (all labelled)	311	Omega	0.945	0.920	0.907	0.938	0.987
		CI	(0.925, 0.964)	(0.895, 0.945)	(0.885, 0.929)	(0.896, 0.979)	(0.968, 1.000)
10 cat. all labelled	346	Omega	0.954	0.946	0.916	0.879	1.000
		CI	(0.935, 0.973)	(0.922, 0.970)	(0.894, 0.938)	(0.827, 0.930)	(0.973, 1.000)
10 cat. end labelled	309	Omega	0.932	0.919	0.926	0.944	0.990
		CI	(0.909, 0.954)	(0.894, 0.944)	(0.910, 0.942)	(0.894, 0.993)	(0.960, 1.000)
11 cat. all labelled	301	Omega	0.939	0.966	0.892	0.918	0.998
		CI	(0.916, 0.961)	(0.946, 0.987)	(0.861, 0.922)	(0.861, 0.976)	(0.970, 1.000)
11 cat. end labelled	301	Omega	0.937	0.951	0.917	0.903	0.998
		CI	(0.916, 0.958)	(0.934, 0.969)	(0.895, 0.940)	(0.844, 0.963)	(0.970, 1.000)

Notes: Omegas and their CIs estimated using parametric bootstrap (method 'bsi' in function `ci.reliabilities()` from the MBESS R package). In case of institutional trust listwise deletion of missing data was used. Institutional trust D1, D2 - first and second scale dimension, respectively.

Validity

Table S8 Validity analysis

Model number	Dependent variable	Predictor	N	Interaction effect: response scale format x predictor			Max std. diff.
				F	p	Adj. p	
1	it1	gt	2845	2.78	0.005	0.205	0.244
2	it1	va	2845	2.48	0.011	0.488	0.236
3	va	it1	2845	2.35	0.016	0.688	0.213
4	va	it2	2843	2.27	0.020	0.854	0.265
5	ra	rh	2850	2.01	0.041	1.000	0.248
6	it1	it2	2843	2.00	0.042	1.000	0.260
7	it2	it1	2843	1.98	0.046	1.000	0.264
8	it2	ca	2843	1.89	0.058	1.000	0.266
9	rh	ra	2850	1.83	0.068	1.000	0.151
10	rh	ca	2850	1.76	0.081	1.000	0.245
11	va	gt	2850	1.74	0.084	1.000	0.208
12	it2	age	2843	1.49	0.093	1.000	
13	it2	gt	2843	1.56	0.132	1.000	0.198
14	it1	age	2845	1.32	0.175	1.000	
15	it1	ra	2845	1.43	0.177	1.000	0.236
16	rh	age	2850	1.31	0.181	1.000	
17	it2	va	2843	1.35	0.215	1.000	0.171
18	it2	ati	2843	1.34	0.219	1.000	0.188
19	va	ca	2850	1.32	0.230	1.000	0.212
20	rh	it2	2843	1.26	0.261	1.000	0.188
21	ra	it1	2845	1.24	0.269	1.000	0.206
22	rh	gt	2850	1.16	0.322	1.000	0.186
23	ra	education	2850	1.13	0.337	1.000	
24	ra	ca	2850	1.12	0.346	1.000	0.195
25	ra	age	2850	1.09	0.354	1.000	
26	it1	education	2845	1.07	0.378	1.000	
27	rh	va	2850	1.07	0.381	1.000	0.207
28	va	rh	2850	1.07	0.384	1.000	0.165

Model number	Dependent variable	Predictor	N	Interaction effect: response scale format x predictor			Max std. diff.
				F	p	Adj. p	
29	ra	gt	2850	1.02	0.420	1.000	0.153
30	it2	rh	2843	1.02	0.421	1.000	0.182
31	va	ati	2850	0.97	0.456	1.000	0.150
32	it1	ca	2845	0.96	0.463	1.000	0.150
33	rh	it1	2845	0.96	0.467	1.000	0.154
34	va	age	2850	0.88	0.589	1.000	
35	it1	rh	2845	0.81	0.591	1.000	0.147
36	it2	education	2843	0.80	0.603	1.000	
37	ra	it2	2843	0.74	0.655	1.000	0.154
38	va	ra	2850	0.73	0.666	1.000	0.135
39	ra	va	2850	0.73	0.668	1.000	0.150
40	it2	ra	2843	0.72	0.674	1.000	0.154
41	va	education	2850	0.71	0.685	1.000	
42	rh	education	2850	0.37	0.936	1.000	
43	ra	ati	2850	0.35	0.947	1.000	0.105
44	rh	ati	2850	0.24	0.984	1.000	0.097
45	it1	ati	2845	0.12	0.999	1.000	0.062

Notes: va – Vaccinations; it1 – Institutional trust, dimension 1; it2 – Institutional trust, dimension 2; ra – Reading attitudes; rh – Reading habits; gt – General trust, ati – Attitudes towards immigrants, ca – Cognitive ability;
In models there were also included: main effect of the predictor, main effect of the response scale format, main effect of the typo of instruction, interaction of the type instruction with the predictor variable.
Holm's method was used to adjust p-values for multiple comparisons.
Max std. diff. – for continuous predictor variables only: maximum pairwise difference of the standardized slope parameter for the predictor variable between different response scale formats (standardization was performed using dependent and predictor variables' grand means and standard deviations).

Instructions used in the procedure

Table S9 Instructions used in the procedure

Instruction type	Original version (Polish)	Translated version (English)
Regular	Szanowni Państwo, zapraszamy Państwa do udziału w około 30-minutowym badaniu, które będzie się składało z dwóch części: testu umiejętności poznawczych oraz ankiety. W pierwszej części prosimy o rozwiązanie szeregu zagadek logicznych, a w drugiej zwracamy się o odpowiedzenie na szereg pytań dotyczących ważnych kwestii społecznych, między innymi zaufania do instytucji publicznych. Dziękujemy za zainteresowanie naszym badaniem i jeszcze raz zapraszamy od udziału w nim!	Dear Sir/Madam, We invite you to participate in an approximately 30-minute study consisting of two parts: a cognitive skills test and a survey. In the first part, we ask you to solve a series of logical puzzles, and in the second, to answer several questions on important social issues, including trust in public institutions. Thank you for your interest in our study, and we look forward to your participation!
Appeal	Szanowni Państwo, zapraszamy Państwa do udziału w około 30-minutowym badaniu, które będzie się składało z dwóch części: testu umiejętności poznawczych oraz ankiety. W pierwszej części prosimy o rozwiązanie szeregu zagadek logicznych, a w drugiej zwracamy się o odpowiedzenie na szereg pytań dotyczących ważnych kwestii społecznych, między innymi zaufania do instytucji publicznych. Zwracamy się o odpowiedzenie na szereg pytań dotyczących ważnych kwestii społecznych, m.in. zaufania do instytucji publicznych. Celem naszego badania jest poznanie zdania osób mieszkających w Polsce na temat kluczowych spraw społecznych, m. in. poziomu zaufania, jakim cieszą się poszczególne instytucje publiczne, stosunku do emigracji, szczepień na COVID, jak również poziomu czytelnictwa w kraju. To ważne tematy, znajdujące się w centrum aktualnych zainteresowań polskich i światowych naukowców. Zebranie dokładnych danych na ten temat pozwoli na	Dear Sir/Madam, We invite you to participate in an approximately 30-minute study consisting of two parts: a cognitive skills test and a survey. In the first part, we ask you to solve a series of logical puzzles, and in the second, to answer a series of questions about important social issues, including trust in public institutions. The goal of our study is to gather the opinions of people living in Poland on key social topics, such as the level of trust in public institutions, attitudes toward migration, COVID vaccinations, and the state of reading habits in the country. These are important issues that are currently the focus of interest among Polish and international researchers. Collecting accurate data on these topics will help uncover the real opinions of Poles on these matters.

	poznanie rzeczywistych opinii Polek i Polaków w tych sprawach. Dlatego też apelujemy do Państwa o zaangażowanie w wypełnianie ankiety. Prosimy o jak najdokładniejsze i jak najbardziej prawdziwe odpowiedzi, w przeciwnym razie nie będziemy w stanie wypełnić celów postawionych przed naszym badaniem. Przygotowaliśmy ją z najwyższą dbałością o metodologiczną poprawność i wygodę użytkowników, dlatego prosimy Państwa o wzajemność i udzielenie szczerych i starannych odpowiedzi. Nauki społeczne mogą się rozwijać tylko w oparciu o kooperację osób badanych. Dziękujemy za zainteresowanie naszym badaniem i jeszcze raz zapraszamy od udziału w nim!	We therefore urge you to engage thoughtfully when completing the survey. Please provide the most accurate and truthful answers possible. Otherwise, we will not be able to fulfill the objectives of our study. The survey has been designed with the utmost care for methodological accuracy and user convenience, and we kindly ask for reciprocity in providing honest and thoughtful responses. Social sciences can only advance through the cooperation of study participants. Thank you for your interest in our study, and we invite you once again to take part!
Warning	Szanowni Państwo, zapraszamy Państwa do udziału w około 30-minutowym badaniu, które będzie się składało z dwóch części: testu umiejętności poznawczych oraz ankiety. W pierwszej części prosimy o rozwiązanie szeregu zagadek logicznych, a w drugiej zwracamy się o odpowiedzenie na szereg pytań dotyczących ważnych kwestii społecznych, między innymi zaufania do instytucji publicznych. Chcemy podkreślić, że nasze badanie zawiera pytania skonstruowane specjalnie, by identyfikować osoby badane, które odpowiadają w sposób nieuważny lub udzielają nieprawdziwych informacji. Prosimy o udzielanie tylko odpowiedzi szczerych i godnych ze stanem faktycznym. Pragniemy zwrócić uwagę, że osobym które będą wypełniać ankietę w sposób nierzetelny mogą nawet utracić prawo do wynagrodzenia za ankietę. Dziękujemy za zainteresowanie naszym badaniem i jeszcze raz zapraszamy od udziału w nim!	Dear Sir/Madam, We invite you to participate in an approximately 30-minute study consisting of two parts: a cognitive skills test and a survey. In the first part, we ask you to solve a series of logical puzzles, and in the second, to answer several questions on important social issues, including trust in public institutions. Please note that our study includes questions specifically designed to identify respondents who provide careless or dishonest answers. We kindly request that you respond sincerely and accurately. Be aware that participants who complete the survey unreliably may risk losing their compensation for participation. Thank you for your interest in our study, and we look forward to your participation!

Note. Survey instructions were read on video by one of the researchers from our Institute. The videos included subtitles that were shown to all participants.