Stefkovics, Ádám, and Jan Karem Höhne. 2024. "Examining the Link between the 'Middle Means Typical' Heuristic and Answer Behavior." *Survey Practice* 17 (July). https://doi.org/10.29115/SP-2024-0009.

ARTICLES

Examining the Link between the 'Middle Means Typical' Heuristic and Answer Behavior

Ádám Stefkovics[1,2], Jan Karem Höhne[3]

[1] IQSS, Harvard University, [2] HUN-REN Centre for Social Sciences, [3] DZHW, Leibniz University Hannover

Keywords: data quality, heuristics, probability-based online panel, questionnaire design, rating scale, web survey

https://doi.org/10.29115/SP-2024-0009

Survey Practice

Vol. 17, 2024

Question interpretation in web surveys may not only depend on the textual content but also on visual design aspects. Research has shown that respondents seem to make use of interpretative heuristics when answering questions potentially influencing their answer behavior. In this study, we investigate the implications of the 'middle means typical' (MMT) heuristic, which suggests that respondents perceive the middle option of a scale as the most typical one. For this purpose, we use data from a survey experiment embedded in the probability-based German Internet Panel (N = 4,679) varying the inclusion of a non-substantive "Don't know" option (with or without separation from the substantive options) and scale polarity (unipolar or bipolar). The four questions under investigation were adopted from the Big5 inventory dealing with agreeableness and openness. The results suggest that the MMT heuristic has a minor impact on answer behavior, as the separation of non-substantive options did not affect answer distributions and response times (as a measure of response effort). However, scale polarity influenced answer behavior and response times. Similar to what has been observed in previous studies, unipolar scales elicited more middle answers and bipolar scales elicited more positive answers. Bipolar scales also resulted in longer response times. Although design violations against the MMT heuristic do not seem to impact answer behavior, we still recommend exercising caution when designing scales with non-substantive options. We also highlight the necessity of testing scales differing with respect to polarity.

Introduction and research question

In self-administered surveys, such as web surveys, respondents usually interpret questions based on various aspects. While the literal meaning of questions and answer options is still important, a long line of research suggests that respondents' answers may be also driven by non-textual or visual elements of the survey in general and the questions in particular (see Höhne et al. 2021; Höhne and Yan 2020; Toepoel and Dillman 2011; Tourangeau, Couper, and Conrad 2004, 2007). For example, this includes graphical features, symbols, shapes, and spaces that can be easily and cost-efficiently included in web surveys (Couper, Tourangeau, and Kenyon 2004; Tourangeau, Couper, and Conrad 2004).

Seeking a better understanding of how visual elements impact respondents' answer process survey researchers draw upon psychological theories, such as from Gestalt psychology (Schwarz 2007; Toepoel and Dillman 2011). Building on these theories, Tourangeau, Couper, and Conrad (2004, 370–373) suggested five so-called interpretative heuristics that respondents follow when interpreting questions: 1) 'middle means typical' (i.e., the middle option of a scale is seen as the most typical one), 2) 'left and top means first' (i.e., the leftmost or top option is seen as the "first" one from a conceptual perspective), 3) 'near means related' (i.e., physically close options are seen as conceptually related), 4) 'up means good' (i.e., the topmost option in a vertically aligned scale is seen as the most desirable one), and 5) 'like means close' (i.e., visually similar looking options are seen as conceptually closer). Accordingly, each heuristic gives a specific meaning to visual and/or spatial cues in survey questions (Tourangeau, Couper, and Conrad 2004, 370).

In this article, we specifically focus on the middle means typical (MMT) heuristic and how design violations of it may impact respondents' answer behavior. Tourangeau, Rips, and Rasinski (2000) suggested that respondents see the middle option of a scale as the most typical one and potentially use it as a kind of anchor or reference point to gauge their own attitude or opinion. We consider design violations of the MMT heuristic when the rating scale is designed in a way that the conceptual midpoint and the visual midpoint do not overlap (or mismatch). In order to investigate the MMT heuristic, Tourangeau, Couper, and Conrad (2004, 373–376) manipulated the presentation of the conceptual and visual midpoints of the scale by varying the presentation of the non-substantive options (i.e., "Don't know" and "No opinion") from the substantive options (i.e., "Far too much," "Too much," "About the right amount," "Too little," and "Far too little"). Specifically, the authors either included the non-substantive options along with the substantive options (resulting in a mismatch between conceptual and visual midpoints) or separated them by a space or divider line (resulting in a match of the conceptual and visual midpoints). Interestingly, respondents' answers were shifted towards the visual midpoint (i.e., to the "Too little" option) when the non-substantive options were not separated. In addition, for one out of the two questions used in the experiment, the non-substantive options were selected more frequently by respondents when they were separated. Thus, the authors advocate following the design recommendations of the MMT heuristic and separating non-substantive options from the remaining ones.

A study by Höhne et al. (2021) replicating the Tourangeau, Couper, and Conrad (2004) experiment in Germany reported mixed results. Using eye-tracking technology the authors found that the visual midpoint received the same amount of attention, regardless of whether it matched or mismatched the conceptual midpoint. Respondents paid more attention to the visual than to the conceptual midpoint when both were mismatched. However, they

could not find any evidence that respondents paid more attention to the non-substantive options when they were separated from the remaining options. Building on these findings, Höhne et al. concluded that if at all only a small subset of respondents makes use of the MMT heuristic when answering survey questions.

It remains open whether and under what conditions respondents draw on the MMT heuristic considering the middle option as the most typical one. Considering the study by Tourangeau, Couper, and Conrad (2004) and its replication by Höhne et al. (2021) both studies employed bipolar instead of unipolar scales. Empirical studies on scale design show that scale polarity can impact respondents' communicative and cognitive answer processes (see Höhne, Krebs, and Kühnel 2021, 2022; Menold 2019; Schaeffer and Dykema 2020; Schaeffer and Presser 2003). In particular, the meaning of the conceptual midpoints of bipolar and unipolar scales, including their specific relation to the other substantive options, substantially differs between both polarities (Menold 2019; Wang and Krosnick 2020). While middle options in bipolar scales indicate a neutral position or the absence of a position, unipolar scales indicate a moderate position on the respective scale dimension. In addition, research shows that respondents frequently tend to select positive options in bipolar scales but middle options in unipolar scales (Höhne, Krebs, and Kühnel 2021, 2022; O'Muircheartaigh, Gaskell, and Wright 1995). Accordingly, it can be presumed that the MMT heuristic may be more common in unipolar than bipolar scales. In addition, there is a lack of studies investigating the response effort of unipolar and bipolar scales varying the inclusion of non-substantive options. We therefore address the following research question:

> How do design violations of the MMT heuristic affect answer behavior (i.e., answer distributions) and response effort (i.e., response times) in unipolar and bipolar scales?

In this study, we attempt to provide new evidence on how the MMT heuristic affects respondents' answer behavior by building on a survey experiment embedded in the probability-based German Internet Panel. Specifically, respondents were randomly assigned to one out of four experimental groups varying with respect to the inclusion of a non-substantive "Don't know" option and scale polarity. We additionally use response times in seconds to investigate response effort (Höhne and Yan 2020). In what follows, we describe the data collection, sample, experimental design, and questions used. We then report the statistical results and provide a methodology-driven discussion and conclusion, including perspectives for future research.

Method

Data

Data were collected from the German Internet Panel, which was part of the Collaborative Research Center 884 'Political Economy of Reforms' at the University of Mannheim. The German Internet Panel is based on an initial recruitment in 2012 and two refreshing recruitments in 2014 and 2018. While the recruitments in 2012 and 2014 were based on a three-stage stratified probability sample of the German population, the recruitment in 2018 was based on a two-stage stratified probability sample of the German population. For a detailed methodological description of the German Internet Panel, we refer interested readers to Blom, Gathmann, and Krieger (2015).

The German Internet Panel invites all panel members every two months to participate in a self-administered web survey that deals with a variety of economic, political, and social topics. In addition to respondents' survey answers, the German Internet Panel collects a variety of paradata, such as response times in seconds (s). Each web survey lasts about 20 minutes. For their participation in each wave, respondents receive a compensation of 4 Euros.

At the beginning of each wave, respondents are directed to a short welcome page announcing the approximate length of the web survey and informing them that the compensation for their participation will be credited to their study account after web survey completion. The survey questions used in this study were included close to the beginning of the web survey.

Sample characteristics

For this study, we use data from wave 42 in July 2019 (data is publicly available through the GESIS Data Archive; see Blom et al. 2020). In total, 4,714 respondents participated in wave 42. Of these respondents, 35 broke off before being asked any study-relevant questions. As a result, 4,679 respondents remain for statistical analysis. (The cumulative response rate was 13.9%.) On average, these respondents were born between 1965 and 1969, and 48% of them were female. In terms of education, 14% had graduated from a lower secondary school, 31% from an intermediate secondary school, and 51% from a college preparatory secondary school. Further, 1% still attended school or had finished without a diploma, and 3% reported another degree than previously mentioned .

Table 1. Description of the experimental groups

Experimental group	Visual separation	Scale polarity	Group size
1	No separation	Unipolar	1,170
2	No separation	Bipolar	1,165
3	Separation	Unipolar	1,171
4	Separation	Bipolar	1,173

Experimental groups

Respondents were randomly assigned to one out of four experimental groups (between-subject-design). The groups are defined by the inclusion of a non-substantive "Don't know" option (with or without visual separation from substantive options) and scale polarity (unipolar or bipolar). Table 1 describes the experimental groups.

To evaluate the effectiveness of random assignment and the sample composition between the four experimental groups, we conducted chi-square tests. The results showed no significant differences regarding age, gender, and education.

Survey questions

We used four questions (or statements) from the Big5 inventory that were adopted from Rammstedt et al. (2013). The Big5 personality traits are one of the most widely studied and used traits for understanding and assessing human personality. It describes human personality through five dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism. In this study, the statements used dealt with agreeableness (two questions) and openness (two questions) and were written in German. One statement per survey page was presented (single presentation), and each statement was preceded by an answering instruction. The rating scales consisted of five points, were vertically aligned, and included an additional non-substantive "Don't know" option. In the following, we provide English translations of the four statements used in this study (Appendix A includes screenshots of the questions and Appendix B includes the original German wordings):

1. I trust others easily, believe in the good in people. (agreeableness)
2. I have little artistic interest. (openness)
3. I tend to criticize others. (agreeableness)
4. I have an active imagination, am fanciful. (openness)

Bipolar scale: 1) Completely applicable, 2) Somewhat applicable, 3) Neither/nor, 4) Somewhat inapplicable, 5) Completely inapplicable, and 6) Don't know

Unipolar scale: 1) Applies completely, 2) Applies somewhat, 3) It depends, 4) Rather does not apply, 5) Does not apply at all, and 6) Don't know

Results

To investigate our research question, we first look at the answer distributions of all four questions across all four experimental groups. In line with our research question, we mainly compare unipolar scales with and without separation of the non-substantive "Don't know" option (groups 1 and 3) as well as bipolar scales with and without separation of the non-substantive "Don't know" option (groups 2 and 4). For this purpose, we conducted chi-square tests. Table 2 reports the answer distributions and test results.

Considering Table 2, irrespective of the way of including the non-substantive "Don't know" option, the answer distributions of unipolar and bipolar scales do not differ significantly. To put it differently, unipolar scales with and without separation as well as bipolar scales with and without separation result in almost identical answer distributions. This similarly applies to all four questions used in this study. In addition, the selection of the non-substantive "Don't know" option is very rare varying between less than 1% and less than 3%. The results of directed Z tests ($p_{separation} > p_{no\ separation}$) showed no statistically significant differences for unipolar and bipolar scales, respectively.

We additionally compared the answer distributions of unipolar and bipolar scales finding large differences between both scales. In unipolar scales, respondents tend to select the middle option ("It depends") more often than in bipolar scales. This is supported by the results of directed Z tests ($p_{unipolar} > p_{bipolar}$) and applies to scales with and without the separation of the non-substantive "Don't know" option. In bipolar scales, in contrast, respondents tend to select the positive options ("Completely applicable" and "Somewhat applicable") more often than in unipolar scales. This is supported by the results of directed Z tests ($p_{bipolar} > p_{unipolar}$), except for the second question.

In the next step, we investigated whether the four experimental groups differ with respect to response times in seconds (s). For this purpose, we initially accounted for response time outliers using the following definition procedure (Lenzner, Kaczmirek, and Lenzner 2010): the upper and lower one percentile of the response time distributions were defined as outliers and excluded from the analysis. We then collapsed response times across all four statements and conducted a one-way analysis of variance (ANOVA) using the Bonferroni α-inflation correction procedure to deal with the problem of multiple comparisons. Considering mean response times across experimental groups results in consistently lower response times for unipolar than bipolar scales (see Table 2). Bipolar scales produce significantly longer response times than their unipolar counterparts, irrespective of the way of including the non-

Table 2. Answer distributions (including chi-square test results) and response times in seconds (s)

	Unipolar scales				Bipolar scales			
	No separation (%)		Separation (%)		No separation (%)		Separation (%)	
Statement 1	$\chi^2(5) = 9.63, p = 0.087$				$\chi^2(5) = 2.42, p = 0.789$			
	Applies completely	7	Applies completely	8	Completely applicable	7	Completely applicable	7
	Applies somewhat	39	Applies somewhat	37	Somewhat applicable	53	Somewhat applicable	50
	It depends	30	It depends	33	Neither/nor	16	Neither/nor	16
	Rather does not apply	20	Rather does not apply	18	Somewhat inapplicable	20	Somewhat inapplicable	22
	Does not apply at all	4	Does not apply at all	2	Completely inapplicable	3	Completely inapplicable	4
	Don't know	1	Don't know	1	Don't know	1	Don't know	1
Statement 2	$\chi^2(5) = 2.28, p = 0.809$				$\chi^2(5) = 3.61, p = 0.607$			
	Applies completely	9	Applies completely	9	Completely applicable	9	Completely applicable	8
	Applies somewhat	29	Applies somewhat	29	Somewhat applicable	29	Somewhat applicable	31
	It depends	21	It depends	20	Neither/nor	13	Neither/nor	14
	Rather does not apply	23	Rather does not apply	24	Somewhat inapplicable	31	Somewhat inapplicable	30
	Does not apply at all	18	Does not apply at all	16	Completely inapplicable	17	Completely inapplicable	15
	Don't know	0	Don't know	1	Don't know	1	Don't know	1
Statement 3	$\chi^2(5) = 3.33, p = 0.649$				$\chi^2(5) = 9.48, p = 0.091$			
	Applies completely	3	Applies completely	3	Completely applicable	3	Completely applicable	3
	Applies somewhat	23	Applies somewhat	24	Somewhat applicable	34	Somewhat applicable	36
	It depends	41	It depends	41	Neither/nor	25	Neither/nor	27
	Rather does not apply	28	Rather does not apply	28	Somewhat inapplicable	30	Somewhat inapplicable	30
	Does not apply at all	5	Does not apply at all	3	Completely inapplicable	5	Completely inapplicable	3
	Don't know	1	Don't know	1	Don't know	3	Don't know	1
Statement 4	$\chi^2(5) = 7.60, p = 0.180$				$\chi^2(5) = 5.97, p = 0.309$			
	Applies completely	19	Applies completely	19	Completely applicable	19	Completely applicable	17
	Applies somewhat	45	Applies somewhat	46	Somewhat applicable	53	Somewhat applicable	55
	It depends	23	It depends	21	Neither/nor	14	Neither/nor	13
	Rather does not apply	11	Rather does not apply	12	Somewhat inapplicable	11	Somewhat inapplicable	13
	Does not apply at all	2	Does not apply at all	1	Completely inapplicable	2	Completely inapplicable	1
	Don't know	1	Don't know	1	Don't know	1	Don't know	2
Response times (s)	56.7		55.0		60.9		59.8	

Note: Due to rounding, the percentages may not add up to 100%. Some of the statements (1 and 4) were positively formulated and some others were negatively formulated (2 and 3). We did not recode the scales to illustrate respondents' general answer tendency. We report aggregated response times for all four statements.

substantive "Don't know" option. This is indicated by the model fit of the ANOVA: $F(3,4478) = 11.01$, $p < 0.001$. Appendix C additionally presents the response time distributions.

Discussion and conclusion

This study contributes to the understanding of the use of interpretative heuristics (Tourangeau, Rips, and Rasinski 2000) while answering survey questions. Specifically, we were interested in whether design violations of the MMT heuristic affect respondents' answer behavior. To this end, we investigated the inclusion of the non-substantive "Don't know" option (with or without visual separation) in unipolar and bipolar scales.

The results showed that for all four questions used in this study, the way of separating the non-substantive "Don't know" option resulted in almost no differences. To put it differently, we obtained almost identical answer distributions for unipolar and bipolar scales, respectively. In addition, we did not find any differences with respect to the selection of the "Don't know" option. This suggests that the MMT heuristic only plays a minor role for both scale polarities. This finding is in line with the findings of Höhne et al. (2021) but differs from the findings of Tourangeau, Couper, and Conrad (2004). However, in this study, we used questions with only one instead of two non-substantive options. Thus, our study included questions with six options (five substantive options and one non-substantive option) with no clear visual midpoint in the without separation groups. It is possible that respondents' answers would have shifted more strongly with two non-substantive options and a clear visual midpoint. We therefore recommend that future studies include a more tailored question and scale design.

Consistent with earlier findings (Höhne, Krebs, and Kühnel 2021, 2022; Menold 2019; Schaeffer and Dykema 2020; Schaeffer and Presser 2003), we found large differences when comparing unipolar and bipolar scales. Middle options were selected more frequently in unipolar scales, while positive options were selected more frequently in bipolar scales. These findings challenge the comparability of answers to scales differing in polarity and underscore the importance of standardized and well tested measurement instruments. This particularly applies to cross-cultural, cross-national comparisons.

Interestingly, we found that bipolar scales with and without separation of the non-substantive "Don't know" option produce significantly longer response times than their unipolar counterparts. The fact that both scales, including statements, only differ in three syllables (syllables influence reading or processing time; Baddeley 1992), supports the notion that scale polarity is associated with response effort. Unipolar scales consist of options that are organized along a continuum that, for example, runs from the uppermost to the lowermost point. Bipolar scales, in contrast, consist of options that

are organized along a continuum with two opposite (positive and negative) ends. The answer options commonly run from the uppermost positive point through a "transition point" (Schaeffer and Presser 2003) that is located in the middle of the scale to the uppermost negative point. This complex conceptual structure of bipolar scales may increase response effort resulting in longer response times. However, we found no evidence that the separation of non-substantive options affects response effort in terms of response times.

This study comes with some methodological limitations offering opportunities for future research. First, using "it depends" as the middle option for the unipolar scale may dilute its unipolar character and potentially limit the comparability of answers. We therefore recommend that future studies employ a more tailored middle option (e.g., "applies moderately"). Second, we only used a set of questions measuring personality traits (agreeableness and openness). It remains open whether and to what extent our findings also hold for other question topics. Third, in this study, we only looked at answer behavior but did not consider data quality. Thus, future research could extend this research by, for example, comparing criterion validity. This could be achieved by including criterion measures that are closely related to the target measures (Höhne and Yan 2020; Yeager and Krosnick 2012). Finally, we used data from a probability-based online panel. However, web surveys are frequently conducted with samples from nonprobability online panels, and thus it is important to test the application of the MMT heuristic in such panels as well.

Even though our findings indicate that violations against the MMT heuristic are only a minor threat to answer behavior, we recommend that survey researchers and practitioners exercise caution when designing scales including non-substantive options. The reason is that the exact mechanisms underlying the MMT heuristic are still not fully discovered and that they may vary with the question topic, the number of answer options, language, scale labelling, and other design aspects. Finally, this study added new evidence to the ongoing concern regarding the comparability of answers to questions with unipolar and bipolar scales. It also highlights the necessity of standardized and well tested measurement instruments.

Contact Information

Ádám Stefkovics, Ph.D.
HUN-REN Centre for Social Sciences
Budapest, Hungary, 1097, Tóth Kálmán street 4.
stefkovics.adam@tk.hu

Submitted: January 31, 2024 EST, Accepted: May 26, 2024 EST

REFERENCES

Baddeley, Alan. 1992. "Working Memory." *Science* 255 (5044): 556–59. https://doi.org/10.1126/science.1736359.

Blom, Annelies G., Marina Fikel, Sabine Friedel, Jan Karem Höhne, Ulrich Krieger, Tobias Rettig, and Alexander Wenz. 2020. "German Internet Panel, Wave 42 (July 2019)." GESIS Data Archive, Cologne. https://doi.org/10.4232/1.13465.

Blom, Annelies G., Christina Gathmann, and Ulrich Krieger. 2015. "Setting Up an Online Panel Representative of the General Population: The German Internet Panel." *Field Methods* 27 (4): 391–408. https://doi.org/10.1177/1525822X15574494.

Couper, Mick P., Roger Tourangeau, and Kristin Kenyon. 2004. "Picture This!: Exploring Visual Effects in Web Surveys." *Public Opinion Quarterly* 68 (2): 255–66. https://doi.org/10.1093/poq/nfh013.

Höhne, Jan Karem, Dagmar Krebs, and Steffen-M. Kühnel. 2021. "Measurement Properties of Completely and End Labeled Unipolar and Bipolar Scales in Likert-Type Questions on Income (in)Equality." *Social Science Research* 97 (July):102544. https://doi.org/10.1016/j.ssresearch.2021.102544.

———. 2022. "Measuring Income (In)Equality: Comparing Survey Questions With Unipolar and Bipolar Scales in a Probability-Based Online Panel." *Social Science Computer Review* 40 (1): 108–23. https://doi.org/10.1177/0894439320902461.

Höhne, Jan Karem, Timo Lenzner, Cornelia E. Neuert, and Ting Yan. 2021. "Re-Examining the Middle Means Typical and the Left and Top Means First Heuristics Using Eye-Tracking Methodology." *Journal of Survey Statistics and Methodology* 9 (1): 25–50. https://doi.org/10.1093/jssam/smz028.

Höhne, Jan Karem, and Ting Yan. 2020. "Investigating the Impact of Violations of the 'Left and Top Means First' Heuristic on Response Behavior and Data Quality." *International Journal of Social Research Methodology* 23 (3): 347–53. https://doi.org/10.1080/13645579.2019.1696087.

Lenzner, Timo, Lars Kaczmirek, and Alwine Lenzner. 2010. "Cognitive Burden of Survey Questions and Response Times: A Psycholinguistic Experiment." *Applied Cognitive Psychology* 24 (7): 1003–20. https://doi.org/10.1002/acp.1602.

Menold, Natalja. 2019. "Response Bias and Reliability in Verbal Agreement Rating Scales: Does Polarity and Verbalization of the Middle Category Matter?" *Social Science Computer Review* 39 (1): 130–47. https://doi.org/10.1177/0894439319847672.

O'Muircheartaigh, Colm, George Gaskell, and Daniel B. Wright. 1995. "Weighing Anchors: Verbal and Numeric Labels for Response Scales." *Journal of Official Statistics* 11:295–308.

Rammstedt, Beatrice, Christoph J. Kemper, Mira Céline Klein, Constanze Beierlein, and Anastassiya Kovaleva. 2013. "A Short Scale for Assessing the Big Five Dimensions of Personality: 10 Item Big Five Inventory (BFI-10)." *Methods, Data, Analyses* 7 (17). https://doi.org/10.12758/MDA.2013.013.

Schaeffer, Nora Cate, and Jennifer Dykema. 2020. "Advances in the Science of Asking Questions." *Annual Review of Sociology* 46 (1): 37–60. https://doi.org/10.1146/annurev-soc-121919-054544.

Schaeffer, Nora Cate, and Stanley Presser. 2003. "The Science of Asking Questions." *Annual Review of Sociology* 29 (1): 65–88. https://doi.org/10.1146/annurev.soc.29.110702.110112.

Schwarz, Norbert. 2007. "Cognitive Aspects of Survey Methodology." *Applied Cognitive Psychology* 21 (2): 277–87. https://doi.org/10.1002/acp.1340.

Toepoel, Vera, and Don A. Dillman. 2011. "Words, Numbers, and Visual Heuristics in Web Surveys: Is There a Hierarchy of Importance?" *Social Science Computer Review* 29 (2): 193–207. https://doi.org/10.1177/0894439310370070.

Tourangeau, Roger, Mick P. Couper, and Frederick Conrad. 2004. "Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly* 68 (3): 368–93. https://doi.org/10.1093/poq/nfh035.

———. 2007. "Color, Labels, and Interpretive Heuristics for Response Scales." *Public Opinion Quarterly* 71 (1): 91–112.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge University Press.

Wang, Rui, and Jon A. Krosnick. 2020. "Middle Alternatives and Measurement Validity: A Recommendation for Survey Researchers." *International Journal of Social Research Methodology* 23 (2): 169–84. https://doi.org/10.1080/13645579.2019.1645384.

Yeager, David Scott, and Jon A. Krosnick. 2012. "Does Mentioning 'Some People' and 'Other People' in an Opinion Question Improve Measurement Quality?" *Public Opinion Quarterly* 76 (1): 131–41. https://doi.org/10.1093/poq/nfr066.

Appendices

Appendix A

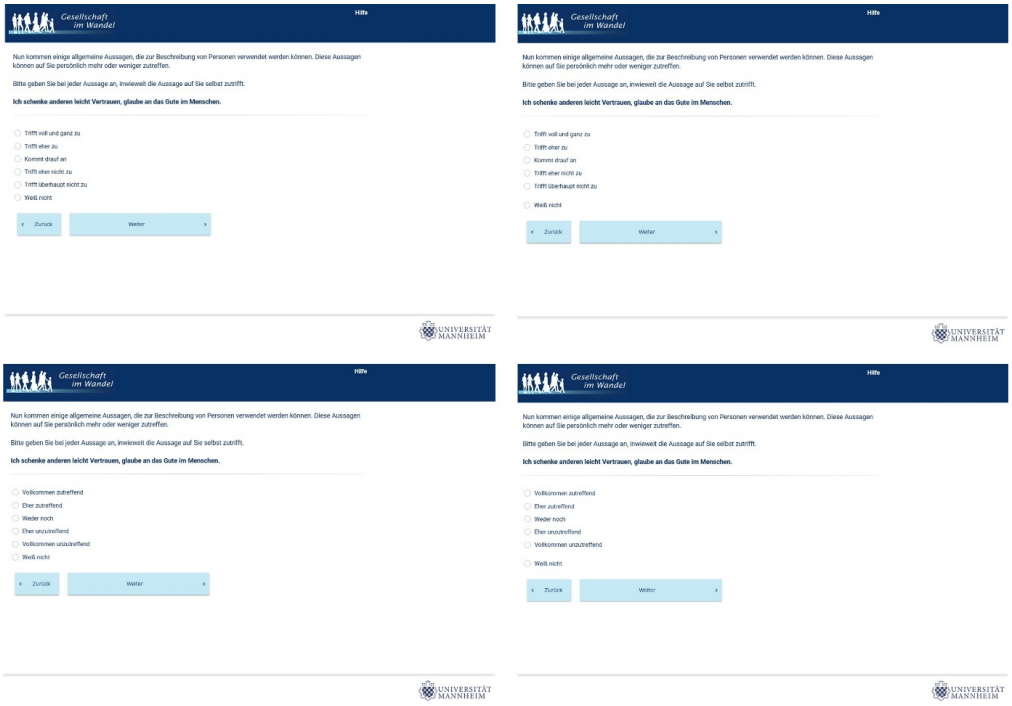

Figure A1. Example screenshots of the first question on agreeableness.

Note: The first group is shown in the upper left corner (no separation, unipolar), the second group is shown in the lower left corner (no separation, bipolar), the third group is shown in the upper right corner (separation, unipolar), and the fourth group is shown in the lower right corner (separation, bipolar).

Appendix B

Original German wordings of the questions and answer options.

1. Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen (agreeableness)
2. Ich habe nur wenig künstlerisches Interesse (openness)
3. Ich neige dazu, andere zu kritisieren (agreeableness)
4. Ich habe eine aktive Vorstellungskraft, bin fantasievoll (openness)

Bipolar scale: 1) Vollkommen zutreffend, 2) Eher zutreffend, 3) Weder noch, 4) Eher unzutreffend, 5) Vollkommen unzutreffend, and 6) Weiß nicht

Unipolar scale: 1) Trifft voll und ganz zu, 2) Trifft eher zu, 3) Kommt drauf an, 4) Trifft eher nicht zu, 5) Trifft überhaupt nicht zu, and 6) Weiß nicht

Note: Each statement was preceded by an answering instruction (see

screenshots in Appendix A). The rating scales consisted of five points, were vertically aligned, and included an additional non-substantive "Don't know" option.

Appendix C

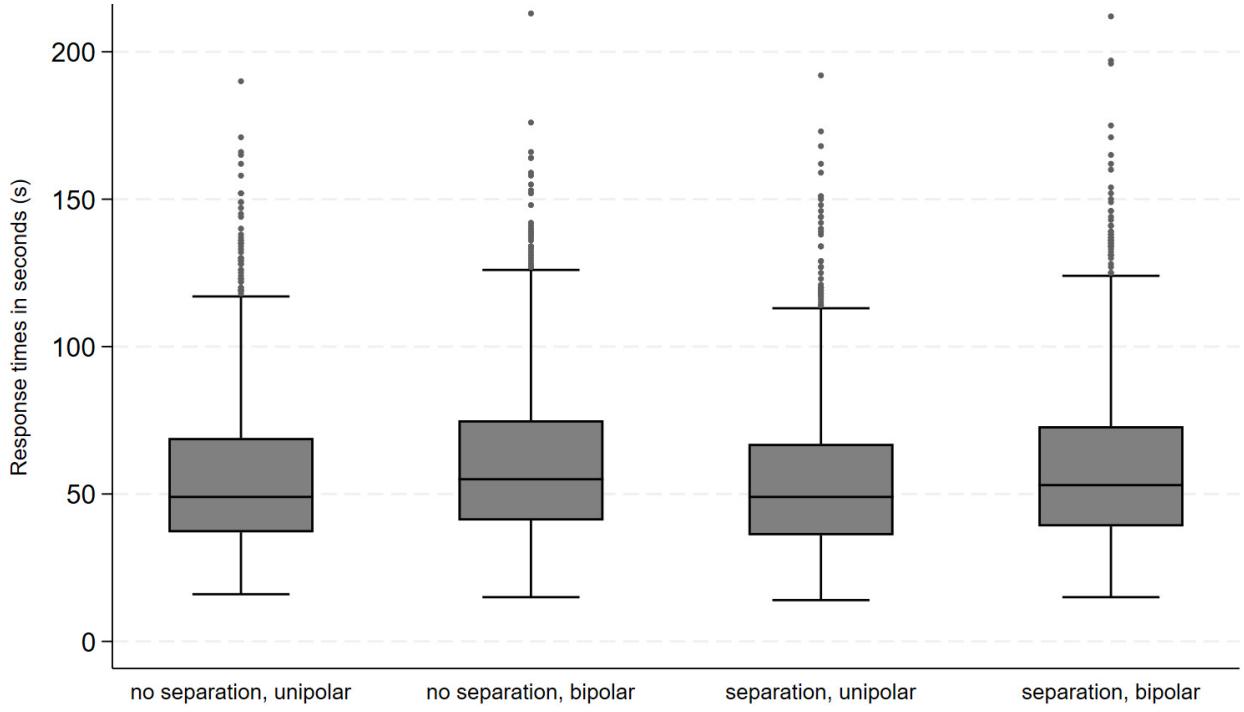

Figure C1. Box plots displaying the response time distributions across the four experimental groups.

Note: Experimental groups: 1) "no separation, unipolar," 2) "no separation, bipolar," 3) "separation, unipolar," and 4) "separation, bipolar."