Re-Examining the Middle Means Typical and the Left and Top Means First Heuristics using Eye-Tracking Methodology

Jan Karem Höhne
University of Mannheim (Germany)
Universitat Pompeu Fabra (Spain)

Timo Lenzner
GESIS – Leibniz Institute for the Social Sciences (Germany)

Cornelia E. Neuert
GESIS – Leibniz Institute for the Social Sciences (Germany)

Ting Yan
Westat (USA)

Abstract

Web surveys are a common self-administered mode of data collection using written language to convey information. This language is usually accompanied by visual design elements, such as numbers, symbols, and graphics. As shown by previous research, such elements of survey questions can affect response behavior because respondents sometimes use interpretive heuristics, such as the *middle means typical* and the *left and top means first* heuristics, when answering survey questions. In this study, we adopted the designs and survey questions of two experiments reported in Tourangeau, Couper, and Conrad (2004). One experiment varied the position of non-substantive response options in relation to other substantive response options and the second experiment varied the order of the response options. We implemented both experiments in an eye-tracking study. By recording respondents' eye movements, we are able to observe how they read question stems and response options and to draw conclusions about the survey response process the questions initiate. This enables us to investigate the mechanisms underlying the two interpretive heuristics and to test the assumptions of Tourangeau et al. (2004) about the ways in which interpretive heuristics influence survey responding. The eye-tracking data reveal mixed results for the two interpretive heuristics. For the *middle means typical* heuristic it remains somewhat unclear whether respondents seize on the conceptual or visual midpoint of a response scale when answering survey questions. For the *left and top means first* heuristic we found that violations of the heuristic increase response effort in terms of eye fixations. These results are disussed in the context of the findings of the original studies.

Keywords: eye tracking, interpretive heuristics, lab experiment, question processing, response behavior, web survey research

This document is a preprint and thus it may differ from the final version: Höhne, J.K., Lenzner, T., Neuert, C.E., & Yan, T. (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, 25–50. DOI: 10.1093/jssam/smz028.

Introduction

Survey respondents are "cooperative communicators", who use any information provided by the survey instrument to understand and answer a question when completing a self-administered survey (Schwarz 1996). According to the principles of Gestalt psychology, information is communicated through four different visual elements (see Dillman, Smyth, and Christian 2014): words (conveying meaning that helps to understand what is being asked), numbers (conveying additional meaning that helps to understand sequence or order), symbols (e.g., figures convey additional meaning based on what they represent), and graphics (e.g., shapes and visual images convey additional meaning based on what they represent).[1] Following this notion, respondents not only draw on words, but also other visual elements in the form of numbers, symbols, and graphics (Christian and Dillman 2004; Couper, Tourangeau, and Kenyon 2004; Dillman, Smyth, and Christian 2009; Schwarz, Grayson, and Knäuper 1998; Schwarz et al. 1991; Smith 1995; Toepoel and Dillman 2011a, 2011b; Tourangeau, Rips, and Rasinski 2000; Tourangeau et al. 2004, 2007).

Web surveys make it easy to employ different visual elements, such as symbols and graphics (Couper et al. 2004; Dillman, Smyth, and Christian 2014; Toepoel and Dillman 2011a). These visual elements help respondents to correctly fill out a questionnaire and make the survey experience more enjoyable. However, they can also influence respondents' response behavior in unintended ways and thus their answers to the survey questions. One explanation for the impact of visual elements on response behavior is that respondents sometimes make use of so-called interpretive heuristics when answering survey questions. Expanding beyond the principles of Gestalt psychology, Tourangeau et al. (2004), for instance, proposed five such heuristics that assign meaning to spatial and/or visual elements: (1) *middle means typical* (i.e., respondents see the middle response option as the most typical one), (2) *left and top means first* (i.e., respondents see the leftmost or top response option as the first one in a conceptual sense), (3) *near means related* (i.e., respondents see response options or questions that are physically close to each other to be related), (4) *up means good* (i.e., respondents see the top response option as the most desirable one)[2], and (5) *like means close* (i.e., respondents see visually similar response options as conceptually closer).

One limitation of the Tourangeau et al. (2004) study is that only so-called "indirect data" (see Galesic et al. 2008; Galesic and Yan 2011), such as survey answers and response times, were examined. There is no direct evidence on how respondents process the survey questions. To overcome this limitation and to extend the research on interpretive heuristics, in this study, we adopted the same designs and survey questions of two experiments reported by Tourangeau et al. (2004). The first experiment varied the position of non-substantive response options in relation to other substantive response options and the second experiment varied the order of the response options. We address the following main research question: How do violations of the

[1] The Gestalt psychology covers several principles, such as the „principle of proximity" stating that objects that are closer together will be seen as a group. We refer interested readers to Dillman, Smyth, and Christian (2014) and Toepoel and Dillman (2011a) for a comprehensive discussion of the principles of Gestalt psychology and their implications for questionnaire design.

[2] This heuristic can be seen as a special form of the *left and top means first* heuristic (Tourangeau et al. 2004).

middle means typical and the *left and top means first* heuristics influence respondents' processing of survey questions?

In the following, we discuss and outline the theoretical considerations with respect to these two interpretive heuristics. For a detailed discussion of the other three heuristics, werefer interested readers to Toepoel and Dillman (2011a, 2011b) and Tourangeau et al. (2004, 2007).

Background

Respondents are found to use response scales to infer the distribution of an attitude, opinion, or behavior in the general population (Schwarz and Hippler 1987). For instance, they conclude that the middle option of a response scale expresses the most typical or average value in the population and thus it serves as a standard of comparison (Schwarz et al. 1985). According to the *middle means typical* heuristic, the middle option of a response scale serves as an anchor or reference point (see Tourangeau et al. 2000) because it is seen as the typical one. Tourangeau et al. (2004) investigated the use of this heuristic in web surveys by manipulating the presentation of the conceptual midpoint of a response scale; they specifically examined response scales in which the visual midpoint did not coincide with the conceptual midpoint (i.e., the visual midpoint of the response scale fell to one side of the conceptual midpoint). For this purpose, the authors manipulated how non-substantive options, such as "don't know" and "no opinion", were included in response scales. The non-substantive options were either presented as additional radio buttons (resulting in a visual midpoint that was different from the conceptual midpoint) or were separated from the substantive options by a space or a divider line (resulting in a visual midpoint that coincided with the conceptual one). As predicted by the heuristic, for most questions respondents' answers shifted toward the visual midpoint in the condition in which non-substantive options were presented as additional radio buttons causing the visual midpoint falling to one side of the conceptual midpoint. The authors interpreted these results as supporting evidence of respondents' application of the *middle means typical* heuristic and concluded "that the meaning of each response option is partly based on its relative position within the array of response options" (Tourangeau et al. 2004, p. 376). In addition, more respondents selected the non-substantive responses in the condition in which non-substantive responses were separated from other scale points. This indicates that the visual separation by a space or divider line drew respondents' attention to the non-substantive options shown as independent of the substantive options (see Toepoel and Dillman 2011a).

The *left and top means first* heuristic suggests that the first option of a response scale – either the leftmost one in horizontally arranged scales or the top one in vertically arranged scales – is conceptually seen as the first one. In general, this interpretive heuristic corresponds to the reading direction in most Western languages, such as English, French, Spanish, and German (see Rayner 1998; Rayner and Pollastek 2006; Tourangeau et al. 2004). It assumes that respondents expect the first response option in a series of ordered options to be the starting point (e.g., "very good"). In addition, they expect the successive options to follow in a logical progression (e.g., "good", "neither good nor bad" etc.) and that the lowermost option represents the opposite endpoint (e.g., "very bad"). Tourangeau et al. (2004) examined this heuristic by varying the order of the response options. In the first condition, the response options were presented in a way consistent with the *left and top means first* heuristic (i.e., the top option was

one of the endpoints and the successive options followed in a logical order). In the second condition, the presentation of the response options was mildly inconsistent with the heuristic (i.e., the conceptual midpoint appeared at the bottom of the scale). In the third condition, the order of the response options was strongly inconsistent with the heuristic (i.e., "it depends", "strongly agree", "strongly disagree", "agree", and "disagree"). In line with the heuristic, response times increased with the degree of inconsistency, suggesting that the order discrepancies slowed respondents down (see also Holbrook et al. 2000). Furthermore, when the response option "it depends" was presented in the middle of the scale, more respondents selected this option than when it was presented at the top or at the bottom of the scale (Tourangeau et al. 2004). Again, the authors interpreted these results as evidence for the application of the *left and top means first* heuristic.

Eye-Tracking Methodology

In the present study, we rerun the experiments on *separating scale points from non-substantive responses* and *order of the response options* by Tourangeau et al. (2004) and re-examine the authors' assumptions about the ways in which these two interpretive heuristics influence survey responding. We extend the research of Tourangeau et al. (2004) by using eye-tracking methodology, which allows us to explore the underlying mechanisms affecting question processing and respondent behavior (see Galesic and Yan 2011). In eye-tracking studies, participants' eye movements are captured by infrared cameras while reading questionnaire instructions, question stems, and response options. These cameras record respondents' exact eye location and the number, duration, and order of their fixations.

The connection between eye movements and cognitive processing is based upon two assumptions (Just and Carpenter 1980, p. 330). The *immediacy assumption* states that objects that are fixated by the eyes are processed immediately so that their interpretation is not deferred. The *eye-mind assumption* postulates that the eyes remain fixated on an object as long as it is being processed. Taken together, these two assumptions suggest that the time spent fixating on an object is approximately equal to the time it is being processed. As a result, eye movements provide direct information about what respondents process and how intensely they process it (Neuert and Lenzner 2017).

More fixations and longer fixations signal a longer response process (see Galesic et al. 2008; Galesic and Yan 2011; Höhne 2019; Höhne and Lenzner 2015, 2018; Kamoen et al. 2011, 2017; Lenzner, Galesic and Kaczmirek 2011; Neuert and Lenzner 2017). A long response process could result from careful deliberation and thorough recall. It could also reflect respondents having difficulty with comprehension due to the use of a difficult or ambiguous word in a survey question, trouble recalling a specific event, and struggle with arriving at an answer or choosing between response options. For instance, Lenzner et al. (2011) showed that ambiguous noun phrases in survey questions produced higher fixation count and time than unambiguous noun phrases, suggesting that ambiguity increases response effort. In addition, the same ambiguous noun phrases were found to increase the selection of non-substantive responses and to decrease response consistency, resulting in poor data quality (see Lenzner 2012). This finding suggests that eye-tracking parameters are good indicators of response effort and data quality.

In previous research, fixation count and time were used to determine whether or not respondents actually read all parts of a question and whether some parts receive more attention than others (Graesser et al. 2006; Höhne and Lenzner 2015, 2018; Lenzner, Kaczmirek, and Galesic 2014; Neuert 2017). Both Galesic et al. (2008) and Höhne and Lenzner (2015) investigated the occurrence and causes of primacy effects and found that respondents fixated more frequently and longer on the beginning of the scales. Höhne (2019) and Höhne and Lenzner (2018) examined the response effort involved in answering agree/disagree and item-specific questions and observed that respondents fixated more frequently and longer on the item-specifc response options than on the agree/disagree response options. Kamoen et al. (2011, 2017) examined the response effort of answering contrastive – positively or negatively formulated – survey questions and found that negatively worded questions yielded more and longer fixations than positively worded questions. Finally, Neuert (2017) analyzed the processing of forced-choice and check-all-that-apply question formats and reported higher fixation counts and times for forced-choice than check-all-that-apply questions indicating a more deliberate response process initiated by the former format.

We use eye tracking to directly investigate the implications of the *middle means typical* and the *left and top means first* heuristics proposed by Tourangeau et al. (2004). In the next section, we outline the research designs and hypotheses for the two experiments, respectively. Subsequently, we describe the sample, the eye-tracking equipment, the procedure of the study, and the analytical strategies. We then present the results of each experiment. Finally, we discuss the theoretical and practical implications of the findings and suggest perspectives for future research.

Experimental Designs and Hypotheses

Experiment 1: Separating Scale Points from Non-Substantive Responses

We used the same two questions as Tourangeau et al. (2004) that dealt with the performance of the government (see Appendix A and B). Respondents were randomly assigned to one of three experimental conditions. For the first group (n = 44), non-substantive response options were presented as additional radio buttons (*non-separation condition; condition 1*), causing a mismatch between the conceptual and visual midpoint. The second group (n = 46) received the questions separating the non-substantive response options by a space (*space condition; condition 2*). The third group (n = 41) received the questions separating the non-substantive response options by a divider line (*line condition; condition 3*). In conditions 2 and 3, the visual midpoint of the response scale coincides with the conceptual midpoint whereas the visual midpoint in condition 1 falls to the lower side of the conceptual midpoint.

The *middle means typical* heuristic indicates that the visual midpoint of a scale should receive comparatively more attention and that respondents' focus should shift towards the visual midpoint when it is placed to one side of the conceptual midpoint. Hence, we postulate the following hypotheses.

H1.1: Respondents fixate equally long and equally often on the visual midpoint of a response scale, irrespective of whether it coincides with the conceptual midpoint – i.e., respondents pay the same amount of attention to the visual midpoint in conditions 2 and 3 (i.e., "about the right amount") as to the visual midpoint in condition 1 (i.e., "too little").

H1.2: Respondents fixate longer and more often on the visual midpoint than on the conceptual midpoint of a scale when the visual midpoint does not match the conceptual one – i.e., respondents pay more attention to the "too little" option than to the "about the right amount" option in condition 1.

H1.3: When the visual midpoint of a response scale is placed to one side of the conceptual midpoint (condition 1), respondents' attention shifts towards this side of the scale and they fixate longer and more often on the substantive options of this side – i.e., respondents pay more attention to the "too little" and "far too little" options in condition 1 than in conditions 2 and 3.

In addition to these hypotheses that were directly derived from the *middle means typical* heuristic, Tourangeau et al. (2004) found that respondents were more likely to select a non-substantive response option in the conditions in which these were separated from the other substantive options by a space or divider line. The authors speculated that dividing the non-substantive options may draw attention to these options and thus may increase the likelihood of respondents selecting them. Hence, we postulate the following additional hypothesis.

H1.4: Respondents fixate longer and more often on the non-substantive response options if these are separated from the other options by a space or divider line – i.e., respondents pay more attention to the non-substantive response options in the conditions 2 and 3 than in condition 1.

Experiment 2: Order of the Response Options

The two questions for the experiment on *order of the response options* were taken from Tourangeau et al. (2004) and dealt with physician-patient relations (see Appendix A and B). Again, respondents were randomly assigned to one of three experimental conditions. The first group (n = 43) received the response options presented in a consistent order (*consistent condition; condition 1*). The second group (n = 46) were shown response options in a mildly inconsistent order (*mildly inconsistent condition; condition 2*). The third group (n = 42) received two questions presenting the response options in a strongly inconsistent order (*strongly inconsistent condition; condition 3*).

The *left and top means first* heuristic indicates that respondents will be confused when response options do not follow a logical order and thus will need more time to process the scale in comparison to response scales following a conventional order. Hence, we postulate the following hypotheses.

H2.1: Respondents fixate longer and more often on the response options the more inconsistent they are with the *left and top means first* heuristic – i.e., respondents fixate longest and most often on the response options when the order is strongly inconsistent with the heuristic (condition 3), followed by the mildly inconsistent order (condition 2) and the consistent order (condition 1).

H2.2: Respondents read more response options and show more re-fixations between the response options, that is, they re-fixate an option they have read previously after reading at least one other option, the more inconsistent they are with the *left and top means first* heuristic. Specifically, respondents read most options and show most re-fixations in the strongly inconsistent condition (condition 3) followed by the mildly inconsistent condition (condition 2) and the consistent condition (condition 1).

In addition, Tourangeau et al. (2004) found that respondents were more likely to select the conceptual middle option "it depends" if it was presented in the middle of the response scale than if it was presented at the top or the bottom. The authors argued that the middle option's meaning is unambiguous when its position suggests that it represents the midpoint of the scale. If this is the case, then respondents should need less time to infer the meaning of the response option "it depends" if it is placed in the center of the scale. Hence, we postulate the following additional hypothesis.

H2.3: Respondents fixate shorter and less often on the conceptual middle option ("it depends") when it is presented in the middle of the scale than when it is presented at the top or the bottom of the scale.

Method

Participants

We recruited respondents from the respondent pool maintained by the institute as well as by word of mouth. In total, 131 respondents participated in both experiments. 38% were between 18 and 24 years old, 38% were between 25 and 44 years old, 19% were between 45 and 64 years old, and 5% were 65 years or older. 50% of the respondents were female. 7% had graduated from a lower secondary school, 14% from an intermediate secondary school and 79% from a college preparatory secondary school or university. 37% of the participants had participated in at least one web survey during the last three months.

In order to evaluate the effectiveness of random assignment and the sample composition between groups in both experiments, we conducted several chi-square tests for the reported sociodemographic characteristics mentioned above. No statistically significant differences could be observed.

Due to technical difficulties, the recordings of the eye movements for some respondents were not satisfactory because there was a systematic shift to the line below or above the one that was fixated. These respondents were excluded from the data, leaving 111 to 114 respondents for statistical analyses.

Eye-Tracking Equipment

The Senso Motoric Instruments (SMI) RED250mobile eye-tracking system was used to record participants' eye movements. To identify saccades and fixations, we used the SMI BeGaze version 2.3 built-in event detector for high speed eye-tracking data (i.e., for data recorded at a sampling rate of 200 Hz or higher) in the default setting: min. saccade duration: auto; peak velocity threshold: 40°/s; min. fixation duration: 50ms; peak velocity: a) start: 20% of saccade length and b) end: 80% of saccade length. The RED250mobile Eye Tracker is a mobile device that can be mounted on the bottom frame of a desktop monitor or laptop display. The system is typically accurate within 0.4° and has a resolution of 0.3°. It permits head movements within a range of 32×21 cm at 60 cm distance. In the set-up of this experiment, the eye tracker was mounted on the bottom frame of a 22" TFT monitor (resolution 1680×1050). Eye movements were recorded at a sampling rate of 250 Hz. The online questionnaire was programmed with a font size of 16 pixels and double-spaced text with a line height of 40 and 32 pixels for the question text and response options, respectively.

Procedure

The study was conducted at GESIS – Leibniz Institute for the Social Sciences in Mannheim (Germany) in April and May of 2017. Participants were seated in front of the SMI RED250mobile eye-tracking system so that their eyes were approximately 60 cm from it and they were instructed to sit down in a comfortable but stable position. Before the web survey started, the eye tracker was calibrated to ensure each participant's eyes could be accurately tracked.

The web survey contained several unrelated experiments, which were independently randomized to avoid systematic carryover effects. The entire eye-tracking study was supervised by an experimenter who stayed in an adjacent room to observe participants' eye movements on a separate computer screen and to assist in case of problems. The average completion time was about 30 minutes. For their participation in the whole study, respondents received a compensation of €20.

Analytical Strategies

To explore the mechanisms underlying interpretative heuristics and to examine how respondents processed the different question versions, we looked at the following eye-tracking parameters: fixation count, fixation time, number of response options read, and re-fixations (the latter two only for the experiment on *order of the response options*). These four eye-tracking parameters can be defined as follows: (1) fixation count is the total number of fixations on a specific area of interest (e.g., the response options) including re-readings, (2) fixation time is the total duration of fixations on a specific area of interest (e.g., the response options) including re-readings, (3) number of response options read refers to the total number of response options that respondents fixate (including re-fixations), and (4) re-fixations on a specific area of interest (e.g., the response options) are the total number of areas that respondents re-fixate (e.g., fixating a response option again after reading at least one other option).

Due to technical limitations the number of response options read as well as the number of re-fixations could not be automatically detected by the RED250mobile eye-tracking system. The questions were then coded by 2 coders, each of which coded the eye movements of one half of the respondents (n = 65 and n = 66, respectively). In addition, the eye movements of a randomly selected subset of 10% of the respondents (n = 13) were coded by both coders for the purpose of estimating reliability. Interrater agreement was excellent (Fleiss et al. 2003), with an Intraclass Correlation Coefficient (ICC) of 0.98. Discrepancies between the two ratings were examined and discussed with the second author until consensus was reached.

Results

In both experiments we conducted one-way Analyses of Variance (ANOVAs) using the Bonferroni α-inflation correction procedure for equal variances to deal with the problem of multiple comparisons to statistically analyze fixation counts, fixation times, number of response options read, and re-fixations. Furthermore, we calculated Cohen's d (see Cohen 1969) to determine the effect sizes. Since there are no distinct differences between specific questions, we decided to conduct all statistical analyses at the aggregated level in an attempt to reduce the number of statistical tests and to efficiently summarize the results.

In addition, we report the response distributions (see Appendix C) and response times (see Appendix D) for the two questions of the experiments on s*eparating scale points from non-substantive responses* and *order of the response options*.

Experiment 1: Separating Scale Points from Non-Substantive Responses

Fixation Count and Time on the Visual Midpoint

Supporting hypothesis 1.1, Table 1 shows no statistically significant differences between the three experimental conditions with respect to the fixation count and time on the visual midpoint of the scale. This result is additionally supported by Cohen's d, which indicates relatively small effect sizes ($d < 0.20$), except for fixation time between conditions 1 and 3 as well as conditions 2 and 3. All in all, the visual midpoints received the same amount of attention, irrespective of whether they fell on the conceptual midpoint of the scale (conditions 2 and 3) or whether it fell on the lower side of the conceptual midpoint of the scale (condition 1).[3]

Table 1. Mean differences and effect sizes (in parentheses) of fixation count and time between the visual midpoint ("too little") in condition 1 and the visual midpoint ("about the right amount") in conditions 2 and 3

Fixation count					
	Condition 1	Condition 2	F value ($df_1 = 2$)	df_2	p value
Condition 2	–0.37 (0.12)		0.38	111	0.685
Condition 3	–0.61 (0.19)	–0.25 (0.08)			
Fixation time (sec)					
Condition 2	–0.37 (0.19)		1.99	111	0.141
Condition 3	–0.77 (0.49)	–0.40 (0.26)			

Note. Mean differences were calculated by subtracting column means from row means. Cohen's d (in parentheses) indicates the effect size. Condition 1: non-separation; condition 2: space; condition 3: line.

We additionally investigated whether the visual midpoint was also the response option that received most attention in comparison to the other options. To do so, we looked at the mean fixation count and time on each option across all conditions. Figure 1 shows that respondents fixated longest and most often on the fourth option ("too little") in all conditions. Hence, the visual midpoint did not necessarily receive most attention. However, it is important to keep in mind that the fixation count and time on each of the options is not only determined by how intensively respondents process them, but also by the option they select. Options that end up being selected receive some attention simply by the fact that for selecting the option, respondents have to fixate them (see Galesic et al. 2008). An inspection of the response distributions (see Appendix C) reveals that the "too little" option was also selected most frequently, irrespective of the condition. This finding corresponds to the distributions reported by Tourangeau et al. (2004).

[3] We also controlled for the responses given in the analyses, but the main results did not change.

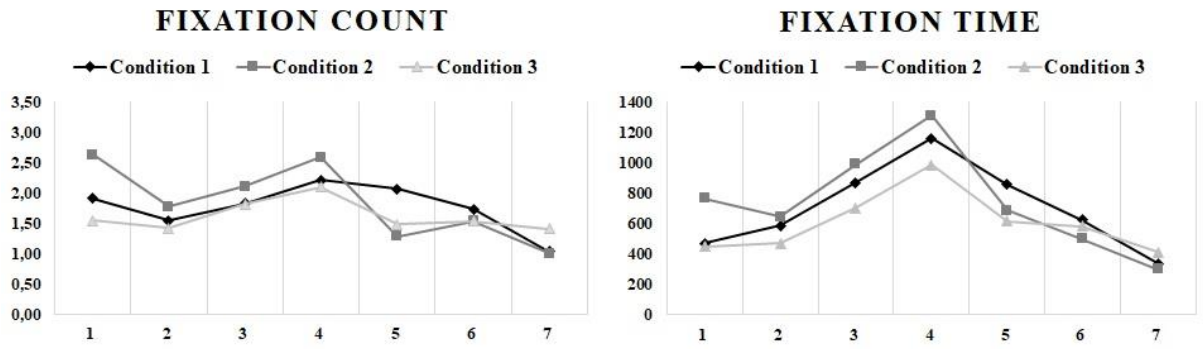

Figure 1. Mean fixation count and time for the seven response options across the three experimental conditions

Note. Condition 1: non-separation; condition 2: space; condition 3: line. Response options: 1) far too much, 2) too much, 3) about the right amount, 4) too little, 5) far too little, 6) don't know, and 7) no opinion. Fixation time is stated in milliseconds (ms).

Fixation Count and Time on the Conceptual and Visual Midpoint in the First Condition

In line with hypothesis 1.2, Table 2 shows that respondents pay more attention to the visual midpoint ("too little") than to the conceptual midpoint ("about the right amount") of the response scale in condition 1 (see also Figure 2). This was marginally significant for fixation count and significant for fixation time. Cohen's d also indicates a small to medium effect size ($d > 0.20$).

Table 2. Mean differences and effect sizes (in parentheses) of fixation count and time between the conceptual midpoint ("about the right amount") and the visual midpoint ("too little") in condition 1

	Eye-tracking parameters	
Condition 1	Fixation count	Fixation time (sec)
Conceptual – visual midpoint	-0.71^{+} $t(40) = -1.62$ (0.23)	-0.54^{*} $t(40) = -1.74$ (0.31)

Note. $^{+}p < 0.10$; $^{*}p < 0.05$. We conducted paired t-tests to determine the differences in the allocation of attention between the conceptual and visual midpoint in condition 1. Cohen's d (in parentheses) indicates the effect size. Condition 1: non-separation.

Fixation Count and Time on the Fourth and Fifth Response Options

According to hypothesis 1.3, we expected that placing the visual midpoint below the conceptual midpoint (condition 1) increases respondents fixation count and time on the two substantive options ("too little" and "far too little") of this side of the response scale. As expected, Table 3 shows that respondents consistently fixated longer and more often on these two substantive options in condition 1 than in the conditions 2 and 3. This finding is additionally supported by Figure 1. However, the differences were not statistically significant. Accordingly, Cohen's d indicates comparatively small effect sizes ($d < 0.30$).

Table 3. Mean differences and effect sizes (in parentheses) of fixation count and time on the fourth and fifth response option ("too little" and "far too little") between the three experimental groups

Fixation count					
	Condition 1	Condition 2	F value ($df_1 = 2$)	df_2	p value
Condition 2	–1.05 (0.18)		0.58	111	0.562
Condition 3	–1.12 (0.22)	–0.06 (0.01)			
Fixation time (sec)					
Condition 2	–0.17 (0.06)		0.61	111	0.530
Condition 3	–0.68 (0.28)	–0.51 (0.20)			

Note. Mean differences were calculated by subtracting column means from row means. Cohen's d (in parentheses) indicates the effect size. Condition 1: non-separation; condition 2: space; condition 3: line.

Fixation Count and Time on the Non-Substantive Response Options

According to hypothesis 1.4, we examined whether the non-substantive response options ("don't know" and "no opinion") received more attention if they were separated from the other options by a space (condition 2) or a divider line (condition 3) than if they were presented as additional radio buttons (condition 1). Contrary to our expectation, Table 4 shows no significant differences in fixation count and time on the non-substantive options between the three conditions. This finding also corresponds to average fixation count and time shown in Figure 1. Cohen's d indicates comparatively small effect sizes, except for fixation count and time between conditions 2 and 3. Hence, our findings are not in line with Tourangeau et al.'s (2004) speculation that non-substantive options receive more attention if they are separted from the other options.

Table 4. Mean differences and effect sizes (in parentheses) of fixation count and time on the non-substantive response options ("don't know" and "no opinion") between the three experimental groups

Fixation count					
	Condition 1	Condition 2	F value ($df_1 = 2$)	df_2	p value
Condition 2	–0.68 (0.15)		0.53	111	0.589
Condition 3	0.46 (0.09)	1.14 (0.27)			
Fixation time (sec)					
Condition 2	–0.36 (0.19)		0.69	111	0.503
Condition 3	0.12 (0.06)	0.49 (0.27)			

Note. Mean differences were calculated by subtracting column means from row means. Cohen's d (in parentheses) indicates the effect size. Condition 1: non-separation; condition 2: space; condition 3: line.

Experiment 2: Order of the Response Options

Fixation Count and Time on the Full Scale

In line with hypothesis 2.1, Table 5 shows that respondents fixated more often and longer on the response options the more inconsistent they were with the *left and top means first* heuristic. Significant differences in fixation count were found between conditions 1 and 3 and in fixation time between both conditions 1 and 2 as well as conditions 1 and 3. Cohen's d additionally supports these results, indicating strong effect sizes ($d > 0.60$). All in all, it seems that the order inconsistencies indeed required respondents to engage in a more effortful processing. These findings are in line with Holbrook et al. (2000) and Tourangeau et al. (2004), who were able to show that order discrepancies slowed respondents down.

Table 5. Mean differences and effect sizes (in parentheses) of fixation count and time between the three experimental groups

Fixation count					
	Condition 1	Condition 2	F value ($df_1 = 2$)	df_2	p value
Condition 2	3.00 (0.31)		3.81	108	0.025
Condition 3	7.06* (0.61)	4.05 (0.35)			
Fixation time (sec)					
Condition 2	1.79* (0.62)		4.24	108	0.017
Condition 3	1.87* (0.62)	0.08 (0.02)			

Note. *$p < 0.05$. Mean differences were calculated by subtracting column means from row means. Cohen's d (in parentheses) indicates the effect size. Condition 1: consistent order; condition 2: mildly inconsistent order; condition 3: strongly inconsistent order.

Number of Response Options Read and Re-fixated

Consistent with hypothesis 2.2, Table 6 shows that the number of response options read increases with the order discrepancies. Respondents read marginally significantly more options in condition 2 than in condition 1 and significantly more options in condition 3 than in condition 1. For these two comparisons, Cohen's d shows medium strong effect sizes ($d > 0.50$).

Table 6 also shows that the number of re-fixations increases with the order inconsistencies. This effect was significant between condition 1 (consistent order) and condition 2 (mildly inconsistent order) and marginally significant between condition 1 and condition 3 (strongly inconsistent order). Cohen's d reveals medium strong effect sizes ($d > 0.50$) for these comparisons.

Table 6. Mean differences and effect sizes (in parentheses) of number of response options read and re-fixated between the three experimental groups

No. of response options read					
	Condition 1	Condition 2	F value ($df_1 = 2$)	df_2	p value
Condition 2	1.11^{+} (0.55)		4.50	108	0.013
Condition 3	1.44** (0.52)	0.34 (0.06)			
No. of response options re-fixated					
Condition 2	1.95* (0.56)		3.55	108	0.032
Condition 3	1.72^{+} (0.52)	–0.23 (0.07)			

Note. $^{+}p < 0.10$; *p < 0.05; **p < 0.01. Mean differences were calculated by subtracting column means from row means. Cohen's d (in parentheses) indicates the effect size. Condition 1: consistent order; condition 2: mildly inconsistent order; condition 3: strongly inconsistent order.

Fixation Count and Time on the Conceptual Midpoint

With respect to hypothesis 2.3, we compared the fixation count and time on the conceptual middle option ("it depends"). Contrary to our expectation, respondents in condition 1 ("it depends" presented in the middle) and condition 3 ("it depends" presented at the top) produced significantly more and longer fixations than respondents in condition 2 ("it depends" presented at the bottom), as shown in Table 7. This result is again supported by Cohen's d, which indicates strong effects sizes (d > 0.70).

Table 7. Mean differences and effect sizes (in parentheses) of fixation count and time on the conceptual middle response option between the three experimental groups

Fixation count					
	Condition 1	Condition 2	F value ($df_1 = 2$)	df_2	p value
Condition 2	–3.50*** (0.99)		11.65	108	0.001
Condition 3	1.05 (0.19)	4.55*** (1.10)			
Fixation time (sec)					
Condition 2	–0.98** (0.79)		7.53	108	0.001
Condition 3	0.13 (0.09)	1.11** (0.81)			

Note. **p < 0.01; ***p < 0.001. Mean differences were calculated by subtracting column means from row means. Cohen's d (in parentheses) indicates the effect size. Condition 1: consistent order; condition 2: mildly inconsistent order; condition 3: strongly inconsistent order.

Discussion and Conclusion

Tourangeau et al. (2004) provided intriguing examples and theoretical explanations of how the *middle means typical* and the *left and top means first* heuristics can influence response behavior. However, the conclusions drawn from their studies on the underlying mechanisms are based on

so-called indirect data, such as response distributions and response times. The usefulness of such indirect data is somewhat limited because question processing and respondent behavior is not directly observed and studied (Galesic et al. 2008; Galesic and Yan 2011). For this reason, we conducted the experiments on *separating scale points from non-substantive responses* and *order of the response options* by means of eye-tracking methodology to examine how violations of the *middle means typical* and the *left and top means first* heuristics influence respondents' processing of survey questions. Table 8 provides a summary of our findings in relation to the research hypotheses.

Table 8. Summary of the findings for the hypotheses of the two experiments on *separating scale points from non-substantive responses* and *order of the response options*

Experiment 1: *Separating scale points from non-substantive responses*	
Hypotheses	Findings
H1.1	Supporting evidence
H1.2	Supporting evidence
H1.3	No supporting evidence
H1.4	No supporting evidence
Experiment 2: *Order of the response options*	
H2.1	Supporting evidence
H2.2	Supporting evidence
H2.3	No supporting evidence

In line with our expectation (see H1.1), we found that the visual midpoint received the same amount of attention, irrespective of whether it fell on the conceptual midpoint of the scale (conditions 2 and 3) or whether it fell on the lower side of the conceptual midpoint of the scale (condition 1). In addition, we found that respondents paid more attention to the visual midpoint ("too little") than to the conceptual midpoint ("about the right amount") of the response scale in codition 1 (see H1.2). However, we did not find evidence for the attention shift towards the side of the scale where the visual midpoint is located (see H1.3). There was also no evidence for more attention on non-substantive options when they are separated by a space or divider line from the substantive options (see H1.4). Our results indicate that respondents either do not utilized the *middle means typical* heuristic when answering the survey questions or only a comparatively small number of respondents is affected by the visual design changes. To investigate this issue further, future research could employ studies with more statistical power and investigate the effects among different respondent groups. It is also crucial to test questions that, for instance, vary with respect to the topic, verbalization, number of categories, scale labeling and polarity, and numerical labels. Altogether, more systematic research on the *middle means typical* heuristic and its implications is necessary to get a solid understanding of its importance.

For the experiment on *order of the response options* we found that both fixation counts and fixation times increase with the order discrepancies (see H2.1), replicating findings reported by Holbrook et al. (2000) and Tourangeau et al. (2004). In addition, we observed that the number of options read and the number of re-fixations increase with the degree of inconsistency (see H2.2). However, there was no evidence indicating that respondents fixated on the

conceptual midpoint ("it depends") least intensively when it was placed in the middle of the scale (see H2.3). In general, the *left and top means first* heuristic seems to be at work for web survey responding. Response options that are not presented in a logical order affect question processing in terms of eye fixations, increase effort in responding, and affect responses obtained. Therefore, we highly recommend presenting response options in a way consistent with the *left and top means first* heuristic to decrease response effort and to enhance survey responses.

We note three limitations associated with this study. First, as suggested in the result section, the intensity with which a response option is fixated (e.g., in terms of fixation count and time) does not only depend on how deeply respondents process it, but also on whether they actually select it. For instance, in the case of the experiment on *separating scale points from non-substantive responses* it is difficult to determine whether the option "too little" received more attention because of its function as visual midpoint or because it was the most frequently selected option. It would be beneficial if the visual midpoint was not the most popular option since this impedes the evaluation of its relevance in comparison to the other options by means of eye tracking. Second, more evidence on how respondents' actually interpret response options, such as the middle option, is necessary to better understand the mechanisms of the *middle means typical* and *left and top means first* heuristics. Unfortunately, this cannot be solely achieved by eye-tracking methodology and goes beyond the scope of this study. For this reason, we recommend that future research address this point by, for instance, combining eye-tracking methodology with cognitive interviewing (see Neuert and Lenzner 2017). Third, we only tested two questions on the experiments on *separating scale points from non-substantive responses* and *order of the response options*, respectively. The main reason is that we adopted the experimental designs and survey questions used by Tourangeau et al. (2004). However, it would be useful if future research tests the implications of both heuristics by employing multiple questions.

Finally, our review of the existing survey literature on visual question design strategies and interpretive heuristics (see, for instance, Christian and Dillman 2004; Couper, Tourangeau, and Kenyon 2004; Schwarz, Grayson, and Knäuper 1998; Schwarz et al. 1991; Smith 1995; Toepoel and Dillman 2011a, 2011b; Tourangeau, Rips, and Rasinski 2000; Tourangeau et al. 2004, 2007) indicates that most of these existing studies only address question processing and response behavior but do not address data quality, such as reliability and validity. Therefore, we recommend to investigate the consequences for data quality in future studies. This would also facilitate evaluating the relevance of interpretive heuristics for quantitative social research.

References

Bradburn, N., Sudman, S., & Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design – For Market Research, Political Polls, and Social and Health Questionnaires.* San Francisco, CA: John Wiley & Sons.

Christian, L. M. and Dillman, D. A. (2004), "The influence of graphical and symbolic language manipulations on responses to self-administered questions," *Public Opinion Quarterly*, 68, 57–80.

Cohen, J. (1969), *Statistical Power Analysis for the Behavioral Science*. New York, NY: Academic Press.

Couper, M. P., Tourangeau, R., and Kenyon, K. (2004), "Picture this! Exploring visual effects in web surveys," *Public Opinion Quarterly*, 68, 255–266.

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014), *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, Hoboken, NJ: Wiley.

Fleiss, J. L., Levin, B., and Paik, M. C. (2003), *Statistical Methods for Rates and Proportions*, Hoboken, NJ: Wiley.

Galesic, M., Tourangeau, R., Couper, M. P., and Conrad, F. G. (2008), "Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding," *Public Opinion Quarterly*, 72, 892–913.

Galesic, M., and Yan, T. (2011), "Use of eye tracking for studying survey response processes," in *Social and Behavioral Research and The Internet. Advances in Applied Methods and Research Strategies*, eds. M. Das, P. Ester, and L. Kaczmirek, pp. 349–370, New York, NY: Routledge.

Holbrook, A., Krosnick, J. A., Carson, R. T., and Mitchell, R. C. (2000), "Violating conversational conventions disrupts cognitive processing of attitude questions," *Journal of Experimental Social Psychology*, 36, 465–494.

Höhne, J. K. (2019), "Eye-tracking methodology: Exploring the processing of question formats in web surveys," *International Journal of Social Research Methodology*, 22, 199–206,

Höhne, J. K. and Lenzner, T. (2015), "Investigating response order effects in web surveys using eye tracking," *Psihologija*, 48, 361–377.

Höhne, J. K. and Lenzner, T. (2018), "New insights on the cognitive processing of agree/disagree and item-specific questions," *Journal of Survey Statistics and Methodology*, 6, 401–417.

Just, M. A. and Carpenter, P. A. (1980), "A theory of reading: From eye fixations to comprehension," *Psychological Review*, 87, 329–354.

Kamoen, N., Holleman, B., Mak, P., Sanders, T., and van den Bergh, H. (2011), "Agree or disagree? Cognitive processes in answering contrastive survey questions," *Discourse Processes*, 48, 355–385.

Kamoen, N., Holleman, B., Mak, P., Sanders, T., and van den Bergh, H. (2017), "Why are negative questions difficult to answer? On the processing of linguistic contrasts in surveys," *Public Opinion Quarterly*, 81, 613–635.

Lenzner, T. (2012), "Effects of survey question comprehensibility on response quality," *Field Methods*, 24, 409–428.

Lenzner, T., L. Kaczmirek, and M. Galesic (2011), "Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension," *International Journal of Public Opinion Research*, 23, 361–373.

Neuert, C. E. (2017), "Processing forced-choice vs. check-all-that-apply question formats: Evidence from eye tracking," *Field Methods*, 29, 383–394.

Neuert, C. E., and Lenzner, T. (2017), "Incorporating eye tracking into cognitive interviewing to pretest survey questions," *International Journal of Social Research Methodology*, 19, 501–519.

Rayner, K. (1998), "Eye movements in reading and information processing: 20 years of research," *Psychological Bulletin*, 124, 372–422.

Rayner, K., and Pollastek, A. (2006), "Eye-movement control in reading," in *Handbook of Psycholinguistic*, eds. M. Traxler and M. A. Gernsbacher, pp. 613–658, Amsterdam: Elsevier.

Schwarz, N. (1996), *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*, Mahwah, NJ: Lawrence Erlbaum Associates.

Schwarz, N., Grayson, C. E., and Knäuper, B. (1998), "Formal features of rating scales and the interpretation of question meaning," *International Journal of Public Opinion Research*, 10, 177–183.

Schwarz, N., and Hippler, H.-J. (1987), "What response scales may tell your respondents: Information functions of response alternatives," in *Social Information Processing and Survey Methodology*, eds by H.-J. Hippler, N. Schwarz, and S. Sudman, pp. 163–178, New York, NY: Springer.

Schwarz, N., Hippler, H.-J., Deutsch, B., and Strack, F. (1985), "Response scales: Effects of category range on reported behavior and comparative judgments," *Public Opinion Quarterly*, 49, 388–395.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., and Clark, L.n (1991), "Rating scales: Numeric values may change the meaning of scale labels," *Public Opinion Quarterly*, 55, 570–582.

Smith, T. W. (1995), "Little things matter: A sampler of how differences in questionnaire format can affect survey responses," in *Proceedings of the American Statistical Association: Survey Research Methods Section*, ed. American Statistical Association, pp. 1046–1051, New York, NY: Springer.

Toepoel, V., and Dillman, D. A. (2011a), "How visual design affects the interpretability of survey questions," in *Social and Behavioral Research and The Internet. Advances in Applied Methods and Research Strategies*, eds. by M. Das, P. Ester, and L. Kaczmirek, pp. 165–190, New York, NY: Routledge.

Toepoel, V., and Dillman, D. A. (2011b). "Words, numbers, and visual heuristics in web surveys: Is there a hierarchy of importance?," *Social Science Computer Review*, 29, 193–207.

Tourangeau, R., Couper, M. P., and Conrad, F. G. (2004), "Spacing, position, and order. Interpretive heuristics for visual features of survey questions," *Public Opinion Quarterly*, 68, 368–393.

Tourangeau, R., Couper, M. P., and Conrad, F. G. (2007), "Colors, labels, and interpretive heuristics for response scales," *Public Opinion Quarterly*, 71, 91–112.

Tourangeau, R., Rips, L. L., and Rasinski, K. (2000), *The Psychology of Survey Response*, New York, NY: Cambridge University Press.

Appendix A

Experimental questions (EQ) of the experiment on separating scale points from non-substantive responses (middle means typical)

EQ 1: Think of how much the federal government is doing to make sure women have the same job opportunities as men. Would you say the federal government is doing too much, about the right amount, or too little about this?

EQ 2: Think of how much the federal government is doing to provide day care centers for the children of working parents. Would you say the federal government is doing too much, about the right amount, or too little about this?

Response options to EQ 1 and EQ 2 are far too much, too much, about the right amount, too little, far too little, don't know, no opinion

Experimental questions (EQ) of the experiment on order of the response options (left and top means first)

EQ 1: Do you agree or disagree with the following statement? It is sensible to do exactly what the doctors say.

EQ 2: Do you agree or disagree with the following statement? I have to be very ill before I go to the doctor.

Response options to EQ 1 and EQ 2 are agree strongly, agree, it depends, disagree, disagree strongly (consistent order)

Note. The order of the questions corresponds to the presentation in Appendix A. The German translations of all questions including response options are available from the first author on request.

Appendix B

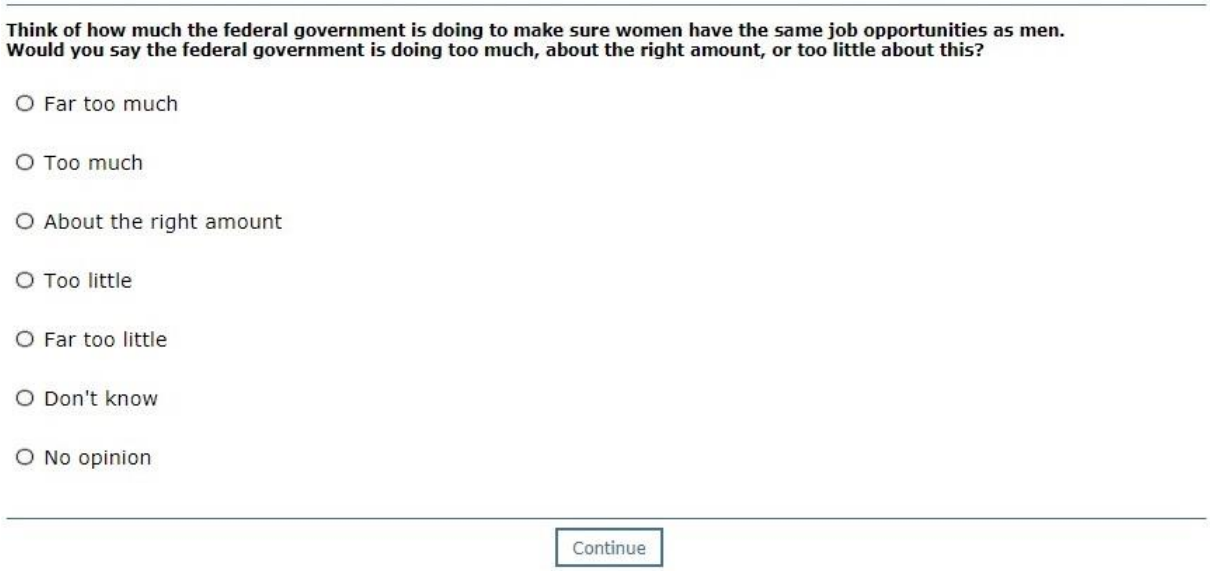

Figure 2a. First question of the experiment on *separating scale points from non-substantive responses* (non-separation condition)

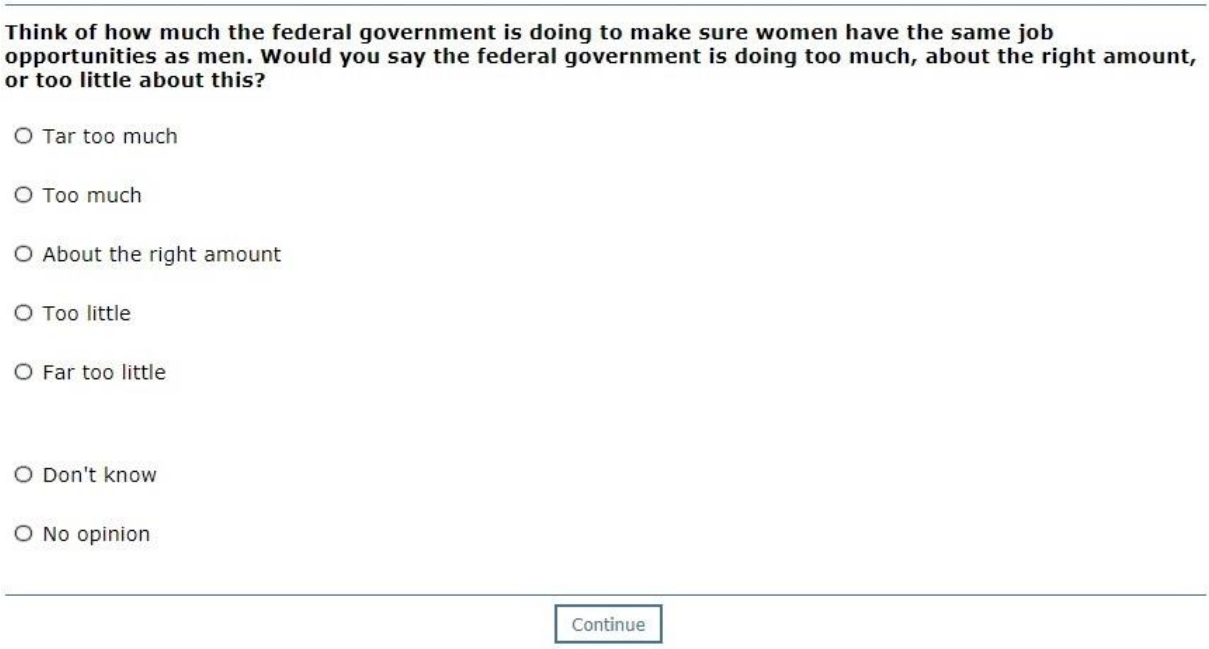

Figure 2b. First question of the experiment on *separating scale points from non-substantive responses* (space condition)

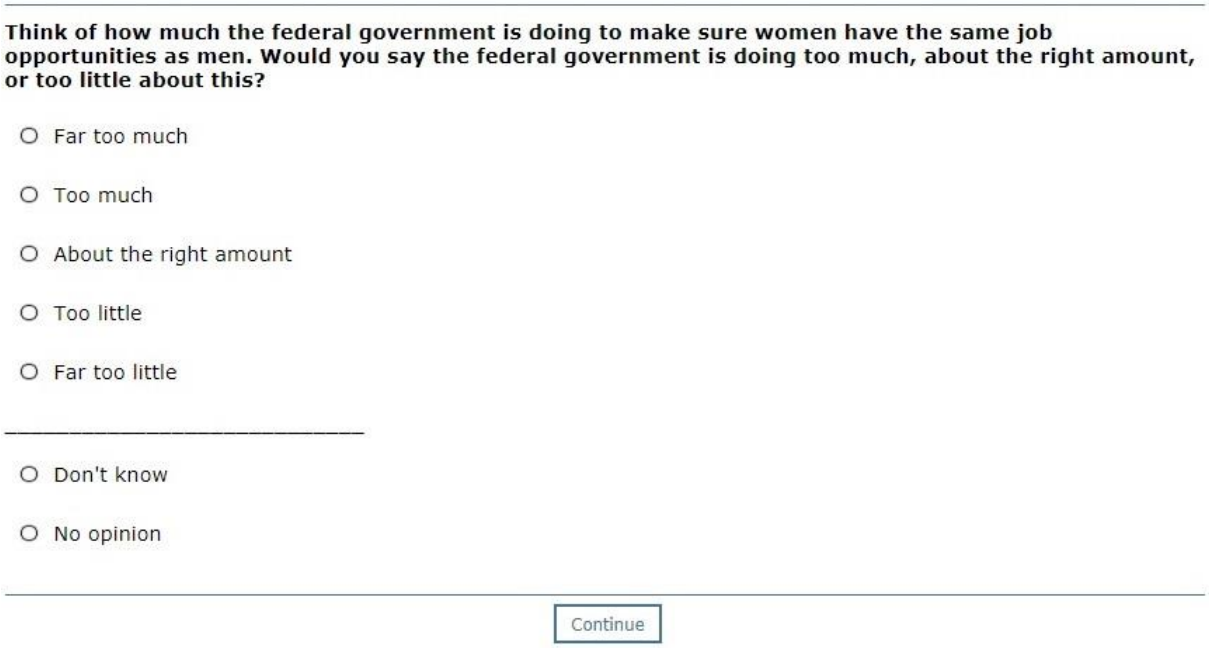

Figure 2c. First question of the experiment on *separating scale points from non-substantive responses* (line condition)

Figure 2d. First question of the experiment on *order of the response options* (consistent order condition)

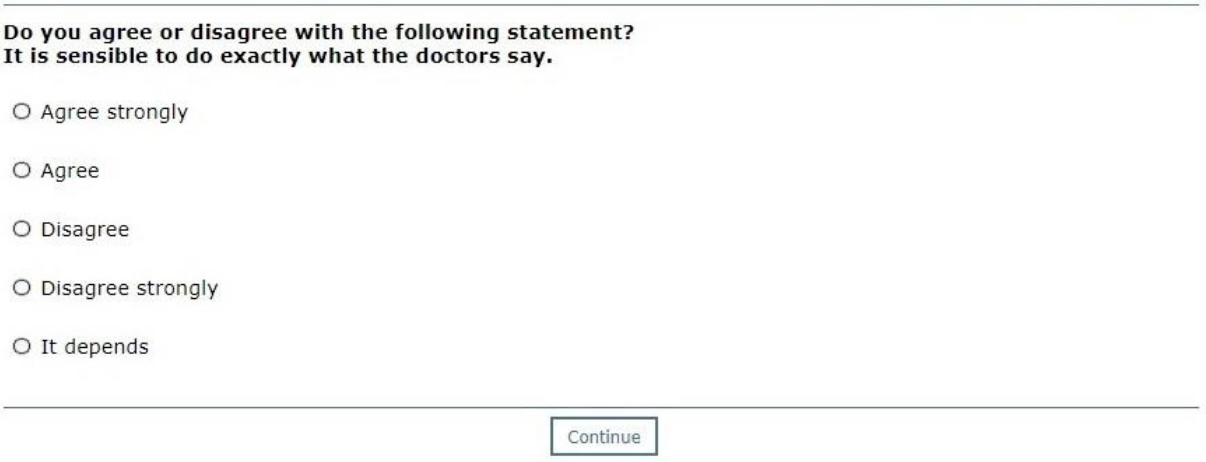

Figure 2e. First question of the experiment on *order of the response options* (mildly inconsistent order condition)

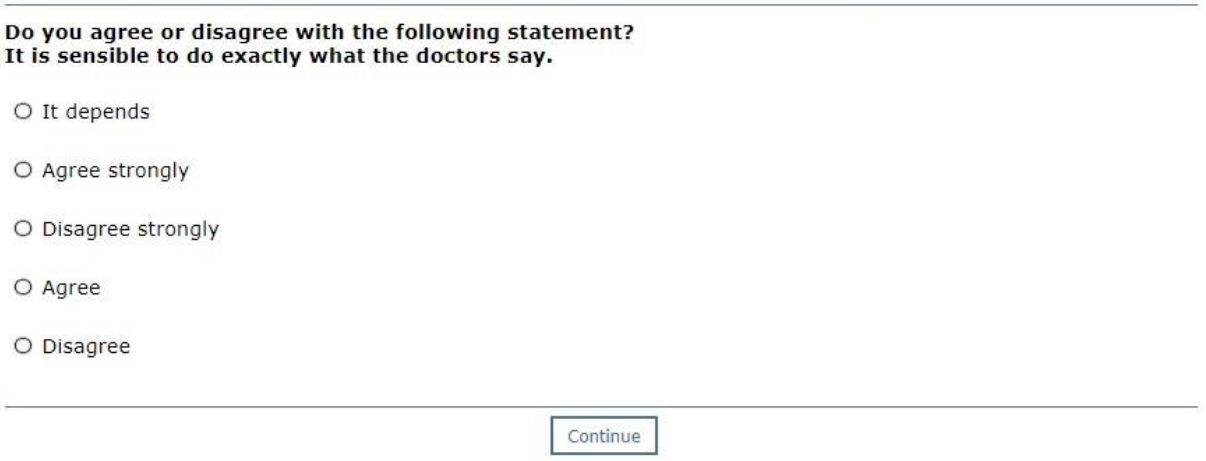

Figure 2f. First question of the experiment on *order of the response options* (strongly inconsistent order condition)

Appendix C

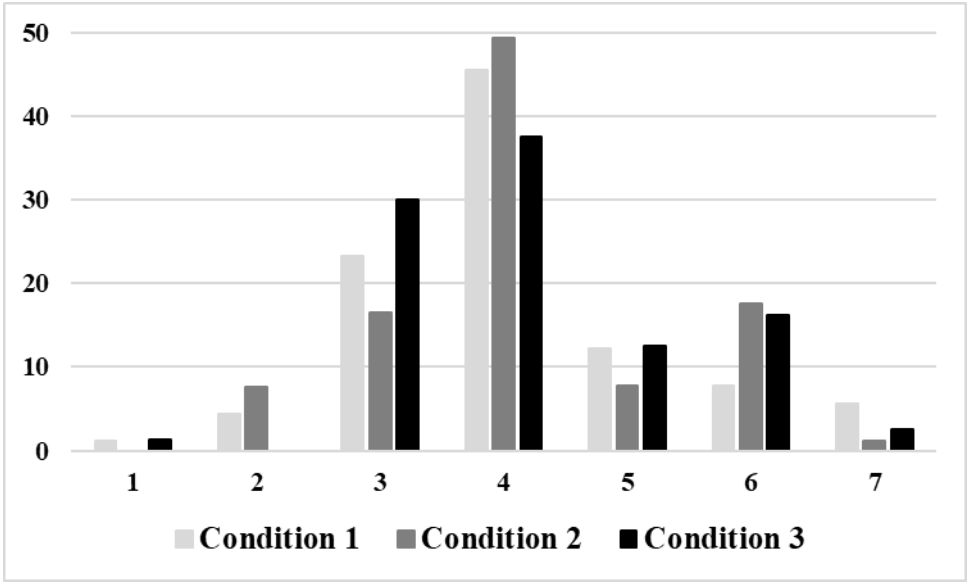

Figure 3a. Response distribution for the two questions of the experiment on *separating scale points from non-substantive responses*

Note. N = 131. Condition 1: non-separation; condition 2: space; condition 3: line. Response options: 1) far too much, 2) too much, 3) about the right amount, 4) too little, 5) far too little, 6) don't know, and 7) no opinion. Due to the small sample size we do not report any test statistics.

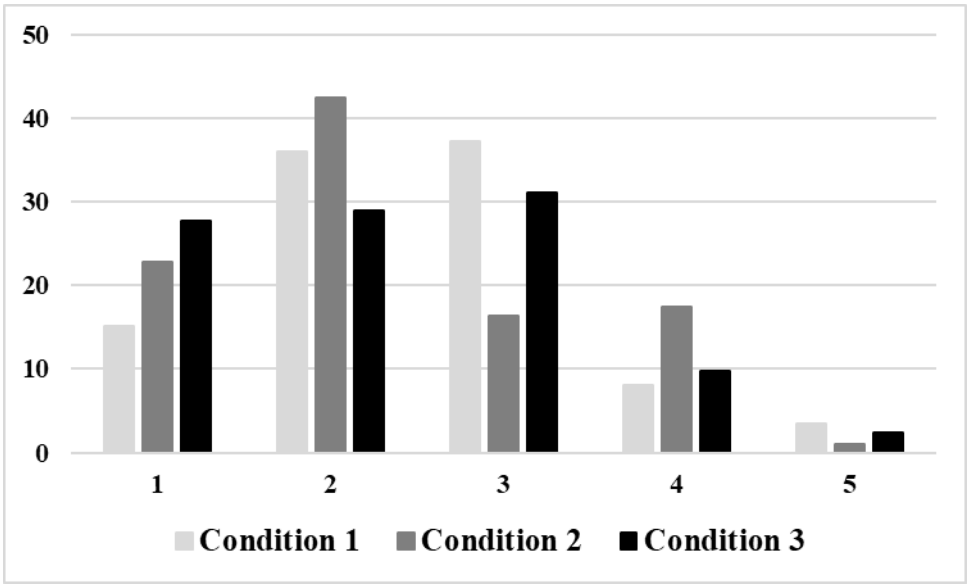

Figure 3b. Response distribution for the two questions of the experiment on *order of the response options*

Note. N = 131. Condition 1: consistent order; condition 2: mildly inconsistent order; condition 3: strongly inconsistent order. Response options: 1) agree strongly, 2) agree, 3) it depends, 4) disagree, 5) disagree strongly. Due to the small sample size we do not report any test statistics.

Appendix D

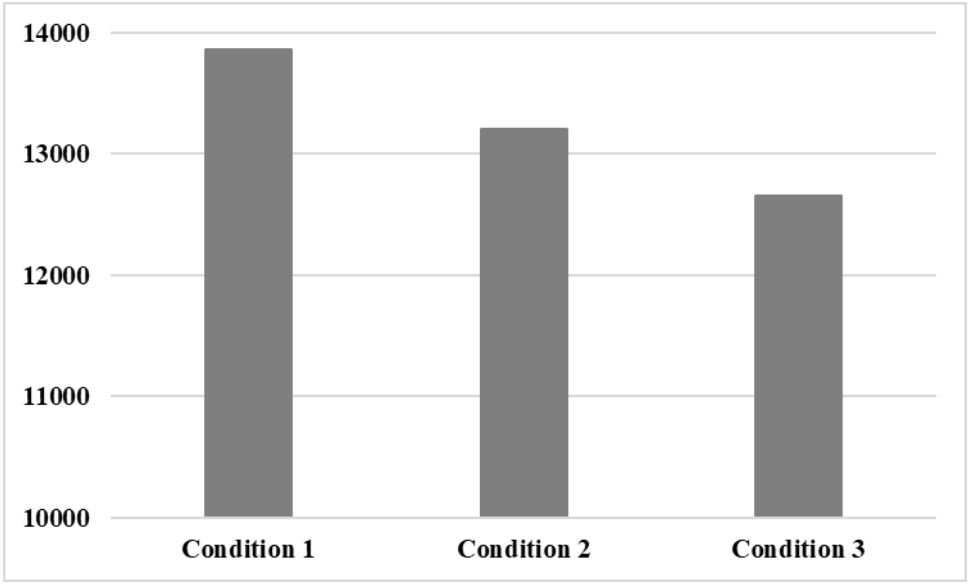

Figure 4a. Median response times (in milliseconds) for the two questions of the experiment on *separating scale points from non-substantive responses*

Note. N = 131. Condition 1: non-separation; condition 2: space; condition 3: line. The result of a Kruskal-Wallis test indicates no significant differences.

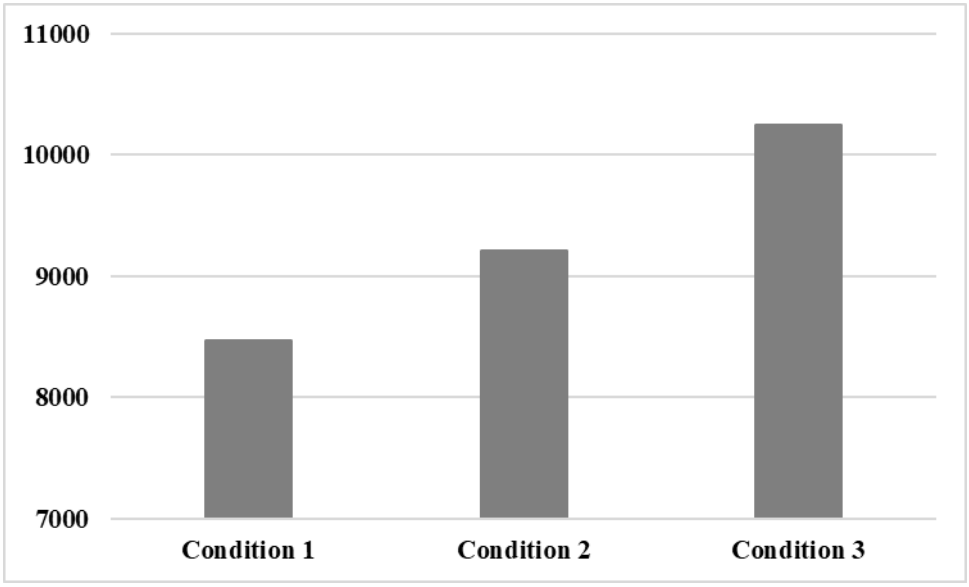

Figure 4b. Median response times (in milliseconds) for the two questions of the experiment on *order of the response options*

Note. N = 131. Condition 1: consistent order; condition 2: mildly inconsistent order; condition 3: strongly inconsistent order. The result of a Kruskal-Wallis test indicates no significant differences.