PSIHOLOGIJA, 2015, Vol. 48(4), 361–377

UDC 303.64
159.9.072
DOI: 10.2298/PSI1504361H

Investigating response order effects in web surveys using eye tracking

Jan Karem Höhne[1] & Timo Lenzner[2]

[1]*Center of Methods in Social Sciences, University of Göttingen, Germany*
[2]*Department of Survey Design and Methodology, GESIS – Leibniz Institute for the Social Sciences, Germany*

Response order effects are a well-known phenomenon that can occur when answering survey questions with multiple response categories. Although various theoretical explanations exist, the empirical evidence is contradictory. Moreover, different scale types produce different effect sizes. In the current study, we investigate the occurrence and causes of response order effects in horizontal and vertical rating scales by means of eye tracking. We conducted an experiment (n = 84) with two groups and varied the scale direction so that the response scales either ran from agree to disagree or vice versa. The results indicate that response order effects in rating scales are relatively small and are more likely to occur in vertical than in horizontal rating scales. Moreover, our eye-tracking data reveal that respondents do not read all categories, nor do they pay equal attention to all categories; these data support the *survey satisficing theory* of response order effects (Krosnick, 1991).

Keywords: Eye tracking, rating scales, response behavior, response order effects, web survey

In quantitative social research, closed-ended survey questions are a common means of collecting data. Closed-ended means that the set of response categories of a question that the respondents can select from are given (Lavrakas, 2008). During the construction of such survey questions, essential questions arise with regard to their design, because it is well-known that the ways in which they are designed can have a profound effect on the responses they produce (Schwarz & Scheuring, 1992; Schwarz, Strack, & Hippler, 1991; Toepoel & Dillman, 2011a; Toepoel & Dillman, 2011b). Accordingly, psychologists and social scientists since Mathews (1929) have examined the influence of the design of questions and response formats on respondents' answers. Hence, there is an abundance of studies that deal with respondents' answers and how they are constructed (Bishop, 1990; Krebs & Hoffmeyer-Zlotnik, 2010; Rammstedt & Krebs, 2007; Schwarz, Hippler, Deutsch, & Strack, 1985; Schwarz, Bless,

Corresponding author: jhoehne@uni-goettingen.de

Bohner, Harlacher, & Kellenbenz, 1991; Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991; Schwarz, Grayson, & Knäuper, 1998; Toepoel & Dillman, 2011b; Tourangeau, Couper, & Conrad, 2004; Tourangeau, Couper, & Conrad, 2007; Tourangeau & Yan, 2007). For instance, many studies have shown that the order in which response categories are presented in closed-ended questions affects responses (Bishop & Smith, 2001; Krebs, 2012; Krebs & Hoffmeyer-Zlotnik, 2010; Malhotra, 2008; Mathews, 1929; Rammstedt & Krebs, 2007; Rugg & Cantrill, 1942; Yan & Keusch, 2015). This type of response bias is called a response order effect, which can be divided up into primacy and recency effects. Primacy effects refer to higher endorsements of response categories presented early in the list, while recency effects refer to higher endorsements of response categories presented late in the list (Schwarz & Hippler, 2004). In general, this suggests a selective memory or perception of response categories at the beginning or at the end of a response scale.

According to the current state of research, the reasons for the occurrence of response order effects in surveys are still unclear and several, divergent theoretical explanations exist. On the one hand, some researchers have suggested that these types of response biases are simply a consequence of questions' difficulty (Bishop & Smith, 2001). However, Mingay and Greenwell (1989) as well as Schuman and Presser (1996) found that response order effects appear even in short and simple questions. Hence, the difficulty of the questions cannot be the main factor. On the other hand, Rugg and Cantrill (1942) postulated the *opinion crystallization hypothesis*, according to which response order effects are the result of uncrystallized attitudes or opinions. This context led Bishop (1990) to explain that the effects of response order appear largely unrelated to how involved a respondent is with a particular issue. In addition, the effect direction – primacy vs. recency effects – depends largely on the survey mode used with primacy effects primarily occurring in self-administered surveys (e.g. mail and online) and recency effects primarily occurring in interviewer-administered surveys (e.g. face-to-face and telephone). This implies that there is a difference between a visual presentation form of the response categories, such as in self-administered surveys and an auditory presentation form of the response categories, such as in telephone surveys (Krosnick & Alwin, 1987). Moreover, the effect sizes of response order effects depend on the types of response formats used. Meaning that there is a substantial difference between categorical and rating scales. For example, Sudman, Bradburn, and Schwarz (1996) suggested that the frequency and size of response order effects seem more limited when rating scales are used than when categorical scales are used.

In the present paper, we examine the occurrence and causes of response order effects in horizontal as well as vertical rating scales. Unlike most former studies, we use eye-tracking methodology to examine our research questions. During eye tracking, respondents' eye movements are captured by infrared cameras while they read questionnaire instructions, survey questions, and response categories. It allows the (exact) eye location, fixation count, fixation duration, and fixation order to be recorded, and makes it possible to directly

investigate response behavior throughout a survey (Galesic, Tourangeau, Couper, & Conrad, 2008; Galesic & Yan, 2011; Geise, 2011). Hence, eye tracking is a suitable (new) technique to investigate hypotheses about response processes and respondents' behavior. For example, Lenzner, Kaczmirek, and Galesic (2011) tested different determinants of question comprehensibility – e.g. low-frequency words, vague and ambiguous noun-phrases, and complex syntactical structures. Kamoen, Holleman, Mak, Sanders, and van den Bergh (2011) investigated the cognitive burden of answering contrastive survey questions. Furthermore, Menold, Kaczmirek, Lenzner, and Neusar (2014) examined the influence of scale length – 5-point vs. 7-point response scales – and scale labeling – fully labeled vs. end-labeled response scales – based on the attention that (verbal) labels received. Although eye tracking is not yet frequently used in survey research, the advantages and potential of the investigation of response behavior and cognitive information processing during surveys are obvious.

To investigate the occurrence as well as the causes of response order effects, we first discuss several theoretical approaches. Afterwards, we describe the study design as well as our research hypotheses, the underlying sample, the eye-tracking equipment used, and the procedure of the study. Then we present the results of our study conducted to examine response order effects in horizontal and vertical rating scales. Finally, we discuss the practical implications of our findings and suggest perspectives for further research.

Theoretical Overview

Response order effects were first discovered by the German psychologist Hermann Ebbinghaus (1913) and can occur when answering questions with multiple response categories. These effects are dependent on the succession of response categories and can affect the response behavior of respondents. If response categories at the beginning of a response scale are selected more often, one speaks of primacy effects. If response categories at the end of a response scale are selected more often, one speaks of recency effects. However, as mentioned above, the causes of the emergence of such response order effects are unclear and the theoretical explanations are divergent. There are currently three main approaches to explaining how they occur: *memory limitation*, *cognitive elaboration*, and *survey satisficing*.

According to the *memory limitation hypothesis*, response order effects are the result of memory limitations, that is, respondents are not able to remember all given response categories (Smyth, Collins, Morris, & Levy, 1994). This is particularly to be expected when complex or relatively large amounts of information are presented in rapid succession and without visual aids (e.g. in telephone interviews), so that recency effects occur. Unfortunately, the *memory limitation hypothesis* cannot explain the emergence of primacy effects. Thus, memory limitations do not seem to be the only (or main) factor responsible for the emergence of response order effects (Sudman et al., 1996, p. 136).

Schwarz, Hippler, and Noelle-Neumann (1992) developed a model, which is known as the *cognitive elaboration model*. This approach is based on the

interaction of the serial position, the presentation mode, and the plausibility of response categories. The serial position refers to the place of a category on a response scale, that is, whether it is at the beginning, the middle or at the end of the response scale. The presentation mode can be distinguished in a visual format, such as self-administered questionnaires and face-to-face interviews, in which the response categories are presented on show cards, and an auditory format, such as telephone interviews or face-to-face interviews without show cards. Plausibility refers to the thoughts that a response category generates in respondents. It can therefore be assumed that a given category is more/less likely to be endorsed the more agreeable/disagreeable the thoughts are that it elicits (Schwarz & Hippler, 2004). As a function of these three factors the *cognitive elaboration model* makes the following predictions: if response categories are presented in a visual format and they elicit mainly agreeable thoughts, the model predicts primacy effects, because visual presentation facilitates the cognitive elaboration of response categories at the beginning. However, if response categories elicit more disagreeable thoughts, recency effects emerge. If response categories are presented in an auditory format and they elicit mainly agreeable thoughts, the model predicts recency effects, because auditory presentation facilitates the cognitive elaboration of response categories at the end. However, if the response categories elicit more disagreeable thoughts, primacy effects emerge. The *cognitive elaboration model* does not consider the influence of memory limitations, the complexity of response categories, and the cognitive ability and motivation of respondents (Bishop & Smith, 2001). Furthermore, contrast effects or a confirmation bias can impede the models' predictions (Sudman et al., 1996, pp. 141-142).

The *survey satisficing* approach developed by Krosnick and his colleagues (Krosnick, 1991; Krosnick & Alwin, 1987; Krosnick, Narayan, & Smith, 1996) distinguishes primarily between an optimizing and a satisficing response style. The second term is further classified into weak and strong satisficing. It is useful to think of optimizing and strong satisficing as the two ends of a continuum indicating the degrees of thoroughness with which the four response steps of the survey response process are performed (Krosnick & Presser, 2010, p. 266). The optimizing end implies an effortful and intensive cognitive response process that produces an optimal answer. In contrast, the satisficing end implies superficial and incomplete cognitive information processing that produces only a sufficient answer. The probability of satisficing depends on the task difficulty, respondent ability, and respondent motivation. Task difficulty depends largely on question characteristics, such as the familiarity of the words used or the complexity of the syntax (Graesser, Cai, Louwerse, & Daniel, 2006). Respondent ability is associated with the cognitive skills required to perform the survey response processes – question comprehension, information retrieval, judgment formation, and response building (Krosnick et al., 1996; Tourangeau, 1984; Tourangeau, Rips, & Rasinski, 2000). Respondent motivation is dependent on a number of different aspects. In principle, the motivation of a respondent varies with the

personal closeness to an issue and the perceived benefit of the survey (Krosnick, 1999). With respect to response order effects, satisficing theory presumes that they are a form of weak satisficing. Similar to the *cognitive elaboration model*, Krosnick and his colleagues predicted that a visual presentation of the response categories would lead to primacy effects and an auditory presentation to recency effects. Regarding primacy effects, there are two explanations on how they emerge (Krosnick, 1991; Krosnick, 1999; Krosnick & Alwin, 1987; Krosnick et al., 1996): on the one hand, it is assumed that respondents just choose the first adequate or reasonable response category, not bothering to read the subsequent ones. On the other hand, it is presumed that respondents consider all response categories, but they are not able to process the subsequent ones in the same way as the former ones, which leads to the earlier response categories being prioritized. The causes of recency effects are more difficult to understand because they are a function of response category processing and memory limitations (Krosnick & Presser, 2010). Due to the rapid presentation of information, respondents devote more cognitive processing time to subsequent categories, thus they are more likely to be selected. Unlike the initial response categories, the subsequent ones are not stored in long-term memory but rather in the short-term memory, from which they can be retrieved much more easily.

With respect to the special case of rating scales, respondents do not need to process all of the different substantive response categories to find the appropriate category, as is the case with categorical scales. For example, rating scales ranging from "agree strongly" to "disagree strongly" build an ordered response continuum, where the different response categories require less processing than is the case with categorical questions (Sudman et al., 1996, p. 157). It seems that ratings in visual as well as auditory presentation modes are generally shifted to the beginning of the response scale (Krosnick, 1991; Yan & Keusch, 2015). Furthermore, the relation between an incremental and decremental succession of the response scale and the primacy effect itself is rather unclear (Krebs & Hoffmeyer-Zlotnick, 2010; Toepoel, 2008). This basically means that primacy effects can also depend on the orientation of the response scale. Moreover, rating scales with a vertical arrangement of the response categories commonly show larger effects of the response order than rating scales with a horizontal arrangement of the response categories (Menold & Bogner, 2015). All in all, the investigation of response order effects in rating scales seems somewhat intricate.

Method

Design and Hypotheses

We conducted an eye-tracking experiment to investigate the impact of the response order on respondents' behavior while they completed an online survey. In order to do this, we changed the order of the underlying response scales in the experimental groups and tested horizontal and vertical arrangements of the response categories. The four items used were taken from the *Cross Cultural Survey of Work and Gender Attitudes* (2010) and were

answered on 6-point fully labeled rating scales ranging from "agree strongly" (= 1) to "disagree strongly" (= 6) in the positive/negative condition and from "disagree strongly" (= 1) to "agree strongly" (= 6) in the negative/positive condition. The question topics dealt with *competition* and *visibility* (the items and rating scales are listed in the appendix). According to the satisficing theory and the used scale type (rating scales), we expected to obtain primacy effects, that is, higher endorsements of categories at the beginning of the response scale. If the response orders influenced respondents' behavior in the postulated way, a larger amount of fixations and longer fixation times of the first response categories should be observed in the eye-tracking data. This argumentation is based on two hypotheses between eye fixations and cognitive processes (Just & Carpenter, 1980, p. 330): first, the *immediacy assumption*, which posits that the interpretations at all levels of processing are not deferred; they occur as soon as possible. And second, the *eye-mind assumption*, which posits that there is no appreciable lag between what is being fixated and what is being processed. Consequently, we assume that the fixation number and fixation time are directly related to the selection of a response category.

The respondents were randomly assigned to one of two experimental groups. The first group (n = 43) received two questions with a horizontal and two questions with a vertical response scale running from agree strongly to disagree strongly (positive/negative condition). The other group (n = 41) received also at first two questions with a horizontal and then two questions with a vertical response scale running from disagree strongly to agree strongly (negative/positive condition). With respect to our argumentation, we assumed that the direction of the response scale – *positive/negative* and *negative/positive* – affected respondents answers such that positive or negative response categories were chosen more frequently when appearing on the left or top half of the response scale (*hypothesis 1*). Additionally, we expected that positive or negative response categories would be fixated more frequently and longer when they are presented on the left or top half of the response scales (*hypothesis 2*). And finally, we hypothesized that the larger the amount of time the left or top half of the response scales was fixated, the more likely it would be that a response category is selected from this side (*hypothesis 3*).

Sample

This study was conducted in October and November of 2012 at the GESIS – Leibniz Institute for the Social Sciences in Mannheim (Germany) and was part of a larger study with several unrelated experiments (Lenzner, Kaczmirek, & Galesic, 2014; Neuert & Lenzner, 2015). We recruited n = 84 participants from the respondent pool maintained by the institute as well as by word of mouth. Due to technical difficulties, the eye movements of two participants could not be recorded accurately. Furthermore, the recorded eye fixations of seven participants were not satisfactory, because there was a systematic shift to the line below the one that was fixated. These participants were therefore excluded from the subsequent eye-tracking analyses. In total, 75 participants with satisfactory eye recordings remained. 53% of these participants were female and 47% were male. They were between 17 and 76 years old with a mean age of M = 35.7 and a standard deviation of SD = 14.6. 20% of the participants graduated from a lower secondary school, 12% from an intermediate secondary school, and 68% from a college preparatory secondary school or university. The bulk of the participants used a computer and the Internet every day or almost every day (89% and 88%, respectively) and 81% had participated in at least one online survey prior to this study. In addition, chi-square tests revealed no statistically significant differences between the two experimental groups with respect to these socio-demographic characteristics – gender ($\chi^2 = 1.78$; $df = 1$; $p = .67$), age ($\chi^2 = 1.15$; $df = 2$; $p = .56$), education ($\chi^2 = 1.58$; $df = 2$; $p = .45$), computer usage ($\chi^2 = .35$; $df = 1$; $p = .55$), Internet usage ($\chi^2 = .79$; $df = 1$; $p = .38$), and survey experience ($\chi^2 = .20$; $df = 1$; $p = .65$).

Eye-Tracking Equipment

Participants' eye movements were recorded by a Tobii T120 Eye Tracker, which allows for unobtrusive eye tracking, and the data were analyzed with the Tobii Studio 3.2.1 software. The T120 is accurate within 0.5° with less than 0.3° drift over time. It allows for head movement within a 30 x 22 x 30 cm volume centered up to 70 cm from the camera. The sampling rate is 120 Hz, meaning that 120 gaze data points per second are collected for each eye. To ensure that all fixations were unequivocally allocated to the response categories and answer boxes respondents had actually read, we used font sizes of 18 and 16 pixels and double-spaced text with line heights of 40 and 32 pixels for the question text and response categories, respectively. The screen resolution was set to 1280 by 1024 pixels. Before analyzing the eye-tracking data, we applied Tobii Studio's I-VT fixation filter in the default setting (*gap fill-in:* enabled, 75 ms; *eye selection:* average; *noise reduction:* disabled; *velocity calculator window length:* 20 ms; *I-VT classifier:* 30°/s; *merge adjacent fixations:* enabled, max time between fixations: 75 ms, max. angle between fixations: 0.5°; *discard short fixations:* enabled, minimum fixation duration: 60 ms) to identify "true" fixations in the raw data. As a sensitivity check, we repeated the analyses of the fixation times and counts on the response categories and the answer text using Tobii's ClearView fixation filter set to include only fixations that lasted at least 100 milliseconds and encompassed 20 pixels. The results were similar to the ones we obtained by applying the I-VT filter in the default setting and all of our conclusions remained unchanged. Before analyzing the eye-tracking data, we used the Tobii Studio 2.0.3 software to define so-called "areas of interest" (AOIs). These AOIs were created by drawing rectangles over the specific text feature words/phrases and over the question stems to quantify the gaze data on these regions and to obtain our dependent variables (i.e., response category fixation count and time).

Procedures

The participants were invited to the pretest laboratory of the institute and seated in front of the eye tracker. After completing a standardized calibration procedure, in which they were asked to follow a moving red dot on the screen with their eyes, they completed the online questionnaire. The calibration procedure was carried out by an experimenter who oversaw the experiment from a separate observer room next to the laboratory. The experimenter monitored respondents' eye movements on a computer monitor in real time. Respondents were instructed to read at a normal pace while trying to understand the questions as well as they could. Only one question at a time was displayed on the screen and the whole questionnaire took about 12 min to complete. For their participation in the whole study (including the cognitive interview), respondents received a compensation of 30 Euros.

Results

In order to reduce the number of statistical procedures, we add up the two competition items with a horizontal as well as the two visibility items with a vertical response scale. Hence, the postulated response order effects will be analyzed as an aggregate of the single items. There are two reasons for this strategy: first, there are no theoretical considerations why effects of the response order have to be analyzed at the item-level. And second, the results do not differ in effect sizes between item-level and item-aggregation.

Hypothesis 1

With regard to our first hypothesis, we investigated whether the direction of the response scale – *positive/negative* and *negative/positive* – influenced the response behavior of respondents so that categories at the beginning of the response scale were selected more often than at the end. In particular, we investigated whether the scale direction affects the univariate answer distributions. To this end, we calculated unpaired t-tests as well as Cohen's *d* as an indicator of the effect sizes between the two different response scale directions. Table 1 below displays the statistical results. Although our results show no significant differences in means between the two experimental groups, we can see that the differences in means between the items with a vertical arrangement of the response categories are much larger than between the items with a horizontal arrangement of the response categories. Considering Cohen's *d* this impression seems to be confirmed. Altogether, it appears that vertical rating scales generally produce larger response order effects than horizontal rating scales, as Menold and Bogner (2015) suggested.

Table 1
Means and standard deviations of horizontal and vertical rating scales running in two different directions

Scale	Condition	Mean / SD
Competition	Condition I (positive/negative)	3.46 / 1.21
(horizontal)	Condition II (negative/positive)	3.56 / .96
		d = .09
Visibility	Condition I (positive/negative)	3.31 / 1.01
(vertical)	Condition II (negative/positive)	3.06 / 1.07
		d = .24

Notations. For an appropriate statistical test, items were recoded to identical values from 1 "disagree strongly" to 6 "agree strongly". Therefore, a higher mean in the first condition (positive/negative) indicates a primacy effect.

Hypothesis 2

Regarding our second hypothesis – positive or negative response categories are fixated more frequently and for a longer amount of time when they are presented on the left or top half of the response scale – we compared the fixation number and fixation time on the first half of the response scale with those on the second half. To do so, we calculated several paired t-tests as well as Cohen's *d* as an indicator of the effect sizes. Table 2 below includes the results of the comparisons of means and the respective effect sizes. Except the fixation time for the two competition items in the first experimental group (positive/negative condition), all means turn out as postulated. The standard deviations show no considerable differences between the two scale halves. Statistically significant differences in means were only found for the visibility items with

a vertical arrangement of the response categories (except for the fixation time of the first experimental group). Table 2, additionally, shows that Cohen's *d* for these three visibility items vary between medium (*d* = .5) and large (*d* = .8) effect sizes (see Cohen, 1969). Interestingly, respondents fixated the first half of the response scale more frequently and for a longer amount of time when the response categories followed a vertical compared to a horizontal arrangement. The direction of the response scale thus had no further influence on the fixation number and time spent on the two scale halves.

Table 2
Means and standard deviations of the fixation number and time of the first and second half of horizontal and vertical rating scales

Scale	Condition	Response Categories	Mean / SD (Number)	Mean / SD (Time)
Competition	Condition I	First three	2.00 / 1.24	.54 / .35
(horizontal)	(positive/negative)	Last three	1.88 / 1.36	.62 / .52
			d = .09	*d* = .18
Visibility	Condition I	First three	2.51 / 1.46	.61 / .38
(vertical)	(positive/negative)	Last three	1.77 / 1.42	.49 / .45
			d = .51*	*d* = .274
Competition	Condition II	First three	2.02 / 1.22	.64 / .38
(horizontal)	(negative/positive)	Last three	1.80 / 1.07	.61 / .45
			d = .19	*d* = .06
Visibility	Condition II	First three	2.97 / 1.66	.76 / .48
(vertical)	(negative/positive)	Last three	1.52 / 1.21	.41 / .36
			d = 1.02***	*d* = .83***

*Notations. *p < .05; **p < .01; ***p < .001.* We calculated Cohen's *d* to determine the effect sizes between the two scale halves. The significance levels, however, are based on the results of the comparison of means. The rating scales in the first condition run from 1 "agree strongly" to 6 "disagree strongly" and in the second condition vice versa. Fixation times are measured in milliseconds.

Figure 1 below includes four heat maps for the first questions on competition and visibility for both experimental groups – *positive/negative* and *negative/positive* – for all respondents. Heat maps illustrate the allocation of attention for different areas of the stimuli and are based on the absolute fixation time for all respondents. The darker an area is marked, the higher is the fixation time. The heat maps reveal that, regardless of the arrangement of the response categories and the scale direction, especially the center – the middle categories – is fixated most intensively. This circumstance, however, is more distinct for horizontal rating scales. In addition, most respondents did not fixate on the last response category at the bottom of the scales, and hence did not read all categories, when answering the questions with vertical rating scales. This can be observed irrespective of the scale direction. Furthermore, these findings directly correspond to the statistical results presented in table 2 above and are in line with

the explanation of the *survey satisficing theory* for the emergence of primacy effects. However, it seems to be that rating scales with a vertical arrangement of the response categories in particular induce respondents to overlook the last response categories and to select the first adequate category instead of reading all available categories when answering survey questions.

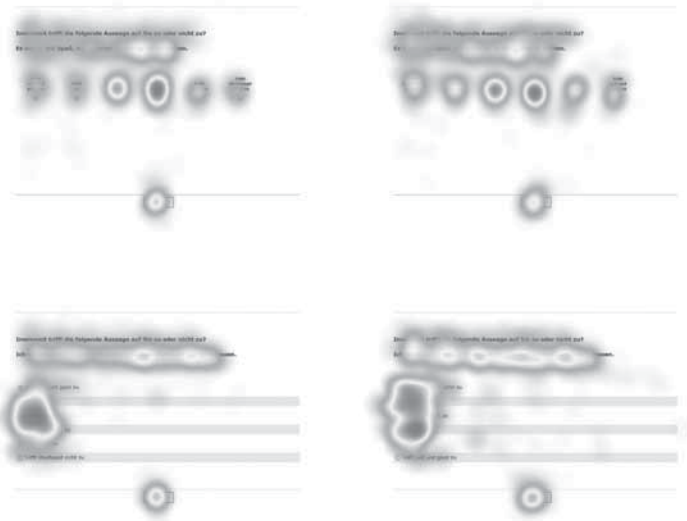

Figure 1. Heat maps of all respondents to the first question on competition and visibility for both experimental groups

Notations. The two heat maps on the left side correspond to the first experimental group (positive/negative condition) and the two heat maps on the right side correspond to the second experimental group (negative/positive condition).

Hypothesis 3

Based on the assumption that the fixation time corresponds directly with the duration of central processing (Just & Carpenter, 1980, p. 330), we were able to examine whether primacy effects are the consequence of processing earlier response categories more intensively than later ones. More explicitly, we tested whether the probability of choosing a response category from the first half of the response scale increases with the longer fixation times on this region. Figure 2 below shows that the longer the first half of the response scale was fixated, the more likely it was that a response category would be selected from this side. This can be observed irrespective of the scale direction and vertical or horizontal arrangement of the response categories. Hence, it appears that the position of a response category on the rating scale has a powerful effect on the relation between the fixation time and the selection of a response category.

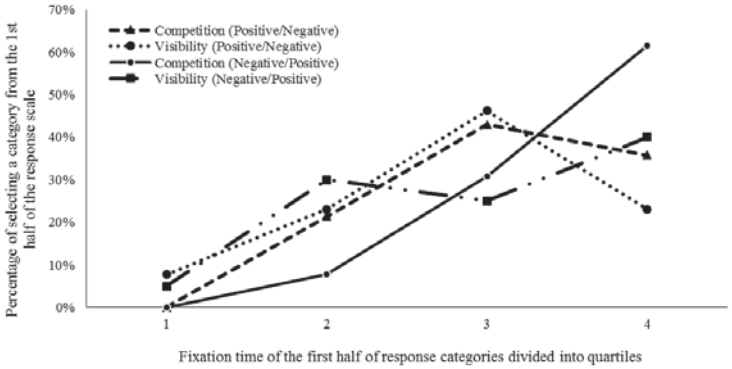

Figure 2. Relation between the amount of fixation time and the propensity of selecting a category from the first half of the rating scale

Additionally, we calculated several chi-square tests to determine whether there are significant differences between the fixation time on the first half of the response scale and the likelihood of selecting a response category from this side. The statistical results reveal significant differences irrespective of the scale direction and the arrangement of the response categories for all items, except for the two visibility items of the first experimental group – *competition positive/negative* ($\chi^2 = 8.93$; $df = 3$; $p = .03$), *visibility positive/negative* ($\chi^2 = 4.64$; $df = 3$; $p = .20$), *competition negative/positive* ($\chi^2 = 15.58$; $df = 3$; $p = .00$), and *visibility negative/positive* ($\chi^2 = 10.36$; $df = 3$; $p = .01$). Altogether, these empirical findings support the two postulated assumptions for the emergence of primacy effects in answering survey questions postulated by the *survey satisficing theory*: respondents either choose the first acceptable response category or process the earlier response categories more deeply than the later ones.

Discussion and Conclusion

The aim of our study was to examine the occurrence and causes of response order effects in horizontal as well as vertical rating scales by means of eye tracking. For this purpose, we assumed the following three hypotheses: (1) positive or negative response categories are selected more often when appearing at the beginning of the scale, (2) positive or negative response categories are fixated more intensively when presented first, and (3) the longer the first half of the scale is fixated, the more likely it is that a response category will be chosen from this side. Firstly, our data suggest that response order effects in rating scales are relatively small. In vertical rating scales, however, they are substantially larger than in horizontal rating scales. Secondly, we found empirical support that respondents fixated the first half of the response scales more intensively than the second half. In particular, this can be observed for rating scales with a vertical arrangement of the response categories. Thirdly, our study provides strong

evidence that the amount of time spent looking at the first half of the response scale correlates to the probability of selecting a response category from this side. In other words, the longer respondents fixate the first half of the response scale, the more likely they are to select one of these response categories. Hence, our findings support the explanation of the emergence of primacy effects postulated by the *survey satisficing theory* (Krosnick, 1991).

A special characteristic of rating scales is that, compared to categorical scales, they follow an ordered response continuum. This basically implies that respondents do not need to process all underlying substantive response categories, because they can (mentally) extrapolate this response continuum. As a result, rating scales seem to be less prone to response order effects than categorical questions. However, there is a considerable difference between rating scales with a vertical and those with a horizontal arrangement of response categories, meaning that vertical rating scales produce larger response order effects than horizontal rating scales; this is in line with our empirical findings. Although the items with a vertical arrangement of the response categories show relatively large effects of the response order, they are not statistically significant. A power analysis (*t*-test, $\alpha = .05$; $\beta = .20$) indicated that minimum sample sizes of $n_1 = 259$ and $n_2 = 247$ (G*Power 3; Faul, Erdfelder, Lang, & Buchner, 2007) would be required to detect any significant differences in means. However, such sample sizes are highly uneconomical in eye-tracking studies.

With respect to our second hypothesis, we observed two interesting aspects: first, although respondents generally fixated the first half of the rating scales more often and longer, there is a significant difference between horizontal and vertical rating scales, i.e. the first half of the scales with a vertical arrangement of the response categories will be fixated more intensively. Therefore, it can be assumed that the type of arrangement – either horizontal (from left to right) or vertical (from top to bottom) – has a high impact on the processing of the rating scale itself. Rayner (1998), for example, point out that the writing system affects the manner of perception. For readers of alphabetical orthographies – e.g. English, French, and German – the span of effective vision extends from 3-4 letters to the left of a fixation to 14-15 letters to the right of a fixation (Rayner & Pollastek, 2006). Thus, the perception while reading tends towards the direction of reading. By way of contrast, readers do not acquire processable information from subjacent lines, because their vertical perception is quite limited (Pollastek, Raney, LaGasse, & Rayner, 1993). Due to their counterintuitive arrangement regarding the writing system, the processing of vertical rating scales seems to be more difficult and burdensome for respondents. It is to assume that exactly this additional effort results in higher endorsements of the response categories presented at the beginning of vertical rating scales. For this reason, rating scales with a vertical arrangement of the response categories seem to be more prone to response order effects than rating scales with a horizontal arrangement, as Menold and Bogner (2015) suggested. Second, we found evidence that respondents mostly fixated the response categories in the center of the rating scales. However, this behavior is much more pronounced for horizontal than for vertical rating

scales and is consistent with our findings in terms of the first hypothesis, since the means are located in the middle of the scale and the differences in means are marginal. Moreover, the differences in the fixation number and time between the first and second half of the horizontal rating scales are either negligibly small or tend to the opposite direction. Referring to the *anchoring-and-adjustment heuristic* by Tversky and Kahnemann (1974), respondents make estimations according to an initial fixed starting point that is aligned with the final answer. This implies that respondents use the middle of horizontal rating scales as a reference point to carry out the rating task. However, due to the fact that such (mental) adjustments made to a reference point are frequently superficial, the final answers often tend towards the reference point. Therefore, our findings regarding horizontal rating scales suggest a response bias that is known as *error of central tendency*. Altogether, the *anchoring-and-adjustment heuristic* offers a reasonable explanation for the observed processing of horizontal rating scales and is in accordance with the reported results and the previous reasoning.

All in all, there are two limitations to this study. On the one hand, the sample size (n = 84) as well as the number of tested items (two questions with a horizontal and two questions with a vertical rating scale) was relatively small, which can be attributed to the eye-tracking experiment and the associated laboratory setting. In particular, this circumstance becomes important with respect to the results of our first hypothesis. It would be quite interesting to see whether the observed effect sizes of the horizontal as well as vertical rating scales would remain constant or change under different circumstances. And if the effect sizes change, what would the differences be between the two scales? On the other hand, our experimental design partially complicates the interpretation of the results, because it is conceivable that the order of the rating scales – horizontal and then vertical – had a further impact on the response behavior of respondents. It would therefore be advisable for further studies to use a more appropriate experimental design to guarantee that the scale order has no impact; this, however, would also require a larger sample size.

Our findings have theoretical and practical implications. From a theoretical point of view, we found empirical evidence for the emergence of primacy effects postulated by the *survey satisficing theory*. In the light of our findings, it seems that primacy effects – at least in vertical rating scales – are simply a consequence of selecting the first acceptable response category. Due to the fact that respondents do not need to process several substantive response categories (as rating scales build a closed response continuum), it seems extremely implausible that they are caused by the inability of respondents to process all response categories in the same way. This explanation is in accordance with the empirical findings of Galesic et al. (2008), who investigated, among others, response order effects in vertical rating scales by means of eye tracking. However, as previously mentioned this explanation applies only to rating scales with a vertical arrangement of the response categories. In answering rating scales with a horizontal arrangement of the response categories, we presumed above that respondents follow the *anchoring-and-adjustment heuristic* by Tversky and

Kahnemann (1974) and use the middle of the scales as an anchor to perform the rating task, so that the answers tend to the center. Unfortunately, this is just a theoretical consideration that requires a more appropriate experimental design with four groups that investigates identical questions with horizontal as well as vertical arrangements in both directions. With regard to practicality, our investigation of horizontal as well as vertical rating scales and their relation to response order effects can help to systematically improve the quality of survey data. Furthermore, our empirical findings can be used to enhance existing "guidelines" and "standards" of developing and constructing rating scales. A final practical recommendation that we can derive from our study is that vertical rating scales are much more prone to response order effects than horizontal rating scales, and should therefore be avoided where possible.

References

Bishop, G. F. (1990). Issue Involvement and Response Effects in Public Opinion Surveys. *Public Opinion Quarterly, 54,* 209–218.

Bishop, G. F., & Smith, A. (2001). Response-Order Effects and the Early Gallup Split-Ballots. *Public Opinion Quarterly, 65,* 479–505.

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

Ebbinghaus, H. (1913): *Memory: A contribution to experimental psychology.* New York: Teachers College.

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.

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-Tracking Data: New Insights on Response Order Effects and other Cognitive Shortcuts in Survey Responding. *Public Opinion Quarterly, 72*(5), 892–913.

Galesic, M., & Yan, T. (2011). Use of Eye Tracking for Studying Survey Response Processes. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet. Advances in applied methods and research strategies* (pp. 349–370). New York: Routledge.

Geise, S. (2011). Eye Tracking in Communication and Media Studies: Theory, Method, and Critical Reflection. *Studies in Communication and Media, 2*, 149-263

Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A Web Facility that Tests Question Comprehensibility. *Public Opinion Quarterly, 70*(1), 3–22.

Just, M. A., & Carpenter, P. A. (1980). A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review, 87*(4), 329–354.

Kamoen, N., Holleman, B., Mak, P., Sanders, T., & van den Bergh, H. (2011). Agree or Disagree? Cognitive Processes in Answering Contrastive Survey Questions. *Discourse Processes, 48*(5), 355–385.

Krebs, D. (2012). The Impact of Response Format on Attitude Measurement. In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, theories, and empirical applications in the social sciences. Festschrift for Peter Schmidt* (pp. 105–113). Wiesbaden: Springer VS.

Krebs, D., & Hoffmeyer-Zlotnik, J. H. (2010). Positive First or Negative First? Effects of the Order of Answering Categories on Response Behavior. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*(3), 118–127.

Krosnick, J. A. (1991). Response Strategies for Coping with the Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology, 5,* 213–236.

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology, 50,* 537–567.

Krosnick, J. A., & Alwin, D. F. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly, 51,* 201–219.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in Surveys: Initial Evidence. In M. T. Braverman & J. K. Slater (Eds.), *New directions for evaluation: Advances in survey research* (pp. 29–44). San Francisco, Calif.: Jossey-Bass Pub.

Krosnick, J. A., & Presser, S. (2010). Question and Questionnaire Design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 263–313). Bingley, UK: Emerald.

Lavrakas, P. J. (2008). Closed-ended Questions. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods - Volume 1* (p. 96). London: Sage Publications.

Lenzner, T., Kaczmirek, L., & Galesic, M. (2011). Seeing Through the Eyes of the Respondent: An Eye-tracking Study on Survey Question Comprehension. *International Journal of Public Opinion Research, 23*(3), 361–373.

Lenzner, T., Kaczmirek, L., & Galesic, M. (2014). Left Feels Right: A Usability Study on the Position of Answer Boxes in Web Surveys. *Social Science Computer Review, 32*, 743–764.

Malhotra, N. (2009). Completion Time and Response Order Effects in Web Surveys. *Public Opinion Quarterly, 72*(5), 914–934.

Mathews, C.O. (1929). The Effect of the Order of Printed Response Words on an Interest Questionnaire. *Journal of Educational Psychology, 30*, 128–134.

Menold, N., & Bogner, K. (2015). Gestaltung von Ratingskalen in Fragebögen. Retrieved on 25th June 2015 from http://www.gesis.org/fileadmin/upload/SDMwiki/Ratingskalen_MenoldBogner_012015_1.0.pdf

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How Do Respondents Attend to Verbal Labels in Rating Scales? *Field Methods, 26*(1), 21–39.

Mingay, D. J., & Greenwell, M. T. (1989). Memory Bias and Response-Order Effects. *Journal of Official Statistics, 5*(3), 253–263.

Neuert, C., & Lenzner, T. (2015). Incorporating Eye Tracking into Cognitive Interviewing to Pretest Survey Questions. *International Journal of Social Research Methodology.*

Pollastek, A., Raney, G.E., LaGasse, L., & Rayner, K. (1993). The Use of Information below Fixation in Reading and in Visual Search. *Canadian Journal of Experimental Psychology, 47*(2), 179-200.

Rammstedt, B., & Krebs, D. (2007). Does Response Scale Format Affect the Answering of Personality Scales? *European Journal of Psychological Assessment, 23*(1), 32–38.

Rayner, K. (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin, 124,* 372–422.

Rayner, K., & Pollastek, A. (2006). Eye-Movement Control in Reading. In M. J. Traxler, M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistic* (Vol. 2, pp. 613–658). Amsterdam: Elsevier.

Rugg, D., & Cantril, H. (1942). The Wording of Questions in Public Opinion Polls. *Public Opinion Quarterly, 5,* 52–78.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Thousand Oaks, CA: Sage Publications.

Schwarz, N., Bless, H., Harlacher, U., & Kellenbenz, M. (1991). Response Scales as a Frame of Reference: The Impact of Frequency Range on Diagnostic Judgments. *Applied Cognitive Psychology, 5,* 37–49.

Schwarz, N., Grayson, C. E., & Knäuper, B. (1998). Formal Features of Rating Scales and the Interpretation of Question Meaning. *International Journal of Public Opinion Research, 10*(2), 177–183.

Schwarz, N., & Hippler, H.-J. (2004). Response Alternatives: The Impact of their Choice and Presentation Order. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 41–56). Hoboken, N.J.: Wiley-Interscience.

Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F. (1985). Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments. *Public Opinion Quarterly, 49,* 388–395.

Schwarz, N., Hippler, H.-J., & Noelle-Neumann, E. (1992). A Cognitive Model of Response-Order Effects in Survey Measurement. In N. Schwarz, & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 187–202). New York: Springer.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating Scales: Numeric Values may Change the Meaning of Scale Labels. *Public Opinion Quarterly, 55,* 570–582.

Schwarz, N., & Scheuring, B. (1992). Selbstberichtete Verhaltens- und Symptomhäufigkeiten: Was Befragte aus Antwortvorgaben des Fragebogens lernen. *Zeitschrift für klinische Psychologie, 21*(2), 197–208.

Schwarz, N., Strack, F., & Hippler, H.-J. (1991). Kognitionspsychologie und Umfrageforschung: Themen und Befunde eines interdisziplinären Forschungsgebietes. *Psychologische Rundschau, 42,* 175–186.

Smyth, M. M., Collins, A. F., Morris, P. E., & Levy, P. (1994). *Cognition in action*. Hove: Erlbaum.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass Publishers.

Toepoel, V. (2008). *A Closer Look at Web Questionnaire Design*. Retrieved on 25th June 2015 from https://pure.uvt.nl/portal/files/1035674/Final_Thesis.pdf

Toepoel, V., & Dillman, D. A. (2011a). How Visual Design Affects the Interpretability of Survey Questions. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet. Advances in applied methods and research strategies* (pp. 165–190). New York: Routledge.

Toepoel, V., & Dillman, D. A. (2011b). Words, Numbers, and Visual Heuristics in Web Surveys: Is there a Hierarchy of Importance? *Social Science Computer Review, 29*(2), 193–207.

Tourangeau, R. (1984). Cognitive Sciences and Survey Methods. In T. Jabine, J. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines* (pp. 73–100). Washington DC: National Academic Press.

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly, 68*(3), 368–393.

Tourangeau, R., Couper, M. P., & Conrad, F. (2007). Color, Labels, and Interpretive Heuristics for Response Scales. *Public Opinion Quarterly, 71*(1), 91–112.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin, 133*(5), 859–883.

Tversky, A., & Kahnemann, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185,* 1124-1131

Yan, T., & Keusch, F. (2015). The Effects of the Direction of Rating Scales on Survey Responses in a Telephone Survey. *Public Opinion Quarterly, 79*(1), 145–165.

Appendix

Table A
Order of the items in the web questionnaire

Item	Item Content
1	I enjoy being in competition with others. *C*
2	It is important to me to perform better than others on a task. *C*
3	I would like to do something important where people look up to me. *V*
4	I find satisfaction in having influence over others. *V*

Notations. C = Competition and V = Visibility.

The rating scales were presented below the items and ran from agree strongly, agree, agree somewhat to disagree somewhat, disagree, disagree strongly in the first condition (positive/negative) and vice versa in the second condition (negative/positive). The competition items received a horizontal and the visibility items a vertical arrangement of the response categories.