Using Eye-Tracking Methodology to Study Grid Question Designs in Web Surveys

Neuert, Cornelia; Roßmann, Joss; Silber, Henning

Empfohlene Zitierung / Suggested Citation:
Neuert, C., Roßmann, J., & Silber, H. (2023). Using Eye-Tracking Methodology to Study Grid Question Designs in Web Surveys. *Journal of Official Statistics*, *39*(1), 79-101. https://doi.org/10.2478/jos-2023-0004

Journal of Official Statistics, Vol. 39, No. 1, 2023, pp. 79–101, http://dx.doi.org/10.2478/JOS-2023-0004

Using Eye-Tracking Methodology to Study Grid Question Designs in Web Surveys

Cornelia E. Neuert[1], *Joss Roßmann*[1], *and Henning Silber*[1]

Grid questions are frequently employed in web surveys due to their assumed response efficiency. In line with this, many previous studies have found shorter response times for grid questions compared to item-by-item formats. Our contribution to this literature is to investigate how altering the question format affects response behavior and the depth of cognitive processing when answering both grid question and item-by-item formats. To answer these questions, we implemented an experiment with three questions in an eye-tracking study. Each question consisted of a set of ten items which respondents answered either on a single page (large grid), on two pages with five items each (small grid), or on ten separate pages (item-by-item). We did not find substantial differences in cognitive processing overall, while the processing of the question stem and the response scale labels was significantly higher for the item-by-item design than for the large grid in all three questions. We, however, found that when answering an item in a grid question, respondents often refer to surrounding items when making a judgement. We discuss the findings and limitations of our study and provide suggestions for practical design decisions.

Key words: Web surveys; response behavior; cognitive processing; question design; eye-tracking methodology.

1. Introduction and Background

The use of grid questions is popular in self-administered surveys, such as web surveys. In a grid question format, respondents receive a series of substantially related items that share the same response scale. The items are usually presented in rows, and the response entry fields are presented in columns (Liu and Cernat 2018). An alternative approach of presenting items sharing the same response scale is the item-by-item design, where items are presented as stand-alone questions (Couper et al. 2013). Between those two extreme points of presenting a series of items are design choices that break the series of target items in smaller groups; for instance, by presenting a set of ten target items in two grids with five items (e.g., Couper et al. 2001; Grady et al. 2019). Each of these formats has benefits and drawbacks.

From a survey designers' perspective, the grid question format is an efficient way to ask multiple questions with the same response scale in a time- and space-saving manner (Couper et al. 2001; Couper et al. 2013; Tourangeau et al. 2004). From a respondent's perspective, the survey length and, thus, completion time is perceived to be shorter, and so is the perceived burden of answering the survey (Heerwegh 2009). Also, grouping items into a grid allows respondents to compare their answers as the content is perceived as belonging conceptually together (Heerwegh 2009; Tourangeau et al. 2004). The latter,

[1] GESIS – Leibniz Institute for the Social Sciences, P.O. Box 12 21 55, 68072 Mannheim, Germany. Emails: cornelia.neuert@gesis.org, joss.rossmann@gesis.org and henning.silber@gesis.org

which goes back to the principle of proximity of Gestalt psychology (Jenkins and Dillman 1995), facilitates comparative judgments and increases consistency of answers compared to when each item is considered in isolation (Couper 2008). However, this may also have disadvantages, such as artificially high inter-item correlations (Silber et al. 2018) due to respondents consistently giving the same answer to each item (a form of satisficing called straightlining or nondifferentiation; Krosnick and Alwin 1988).

Compared to an item-by-item design, in which each item is usually presented on a new page, grid questions present a large amount of information on one page. The amount of information and the effort required to answer the items increases with the size of a grid. An increasing number of items in rows and more response entry fields in columns imply a larger matrix, making navigation more difficult (Couper et al. 2013; Grady et al. 2019). With a cognitively more demanding task, respondents may get more easily confused and distracted, thereby increasing their actual or perceived response burden (Couper et al. 2013; Liu and Cernat 2018).

According to the theory of survey satisficing, the complexity of grid formats might encourage respondents to minimize time and effort for answering them thoroughly (Couper et al. 2013; Krosnick 1991). Congruent with that assumption, a large body of experimental research has shown that grid questions can have negative effects on data quality, which has been shown by higher rates of missing or non-substantive answers (i.e., “don’t know”; Mavletova and Couper 2015; Roßmann et al. 2018; Toepoel et al. 2009), higher levels of non-differentiated answers (i.e., straightlining; DeBell et al. 2021; Roßmann et al. 2018; Tourangeau et al. 2004), and higher breakoff rates compared to item-by-item formats (Couper et al. 2013; Liu and Cernat 2018; Tourangau et al. 2004). Although nondifferentiation, item nonresponse, and similar response behaviors are generally viewed as undesirable response effects, there is the possibility that the less differentiated responses, and greater expressions of uncertainty (e.g., selecting “don’t know”) are closer to “truth”– that is, that grids actually help respondents to understand that their responses to individual items are (legitimately) close to each other, or recognize legitimate uncertainty; and that separating questions into an item-by-item format artificially magnifies differences between responses. While we acknowledge that this alternative interpretation is also plausible, we follow the general view that classifies the response behaviors as undesirable.

The faster completion times of grids compared to item-by-item question formats may also represent a form of superficial cognitive response processing and might increase measurement error (Couper et al. 2001; Peytchev 2005, cited in Couper et al. 2013; Roßmann et al. 2018; Tourangeau et al. 2004). Comparing response times between an item-by-item and a grid design, Roßmann et al. (2018) have shown that the response time for the first item did not differ between the two formats. This finding leads to the question of whether the longer response times within the item-by-item designs result from deeper cognitive processing of the item itself or from the response task of reading the question stem and the response scales each time in the item-by-item format.

In this study, we used eye-tracking methodology to gain more insights into the cognitive information process and response behavior when answering grid versus item-by-item question formats. Therefore, we employed an experiment with three questions. The respondents were randomly assigned to one of three question formats (item-by-item, two small grids, one large grid) while their eye movements were monitored.

2. Research Questions and Hypotheses

We investigated the following two research questions:

1. Does altering the question format affect the depth of cognitive processing when answering grid questions?
2. Does the differential depth of cognitive processing explain differences in response quality between the three different question designs?

Answering survey questions requires respondents to pass through four stages of cognitive processing (Tourangeau and Rasinski 1988; Tourangeau et al. 2000): (1) question comprehension, (2) retrieval of relevant information, (3) use of the information to arrive at a judgment, and (4) reporting of an answer within the response options provided. Each of these four stages of cognitive processing can be challenging for respondents, and thus, may contribute to the emergence of response effects. The theory of survey satisficing complements this framework by incorporating motivational components of the respondents (Krosnick 1991; Roßmann and Silber 2020). It states that respondents might alter their response behavior from complete and thorough execution of the four cognitive steps (i.e., optimizing) to less diligent or incomplete execution (i.e., satisficing) contingent on three factors: task difficulty, ability, and motivation. The higher the difficulty of answering a survey question and the lower a respondent's ability and motivation to perform the task, the higher is the chance of respondents showing satisficing response behavior (Krosnick 1991).

Referring to this theoretical framework, we propose specific hypotheses about how the design of grid questions can affect cognitive processing, and in consequence the survey response. While the design of grid questions could affect all four stages of cognitive processing, it seems likely that it mainly affects the processes of question comprehension and reporting of an answer. By using eye-tracking data, we can differentiate between different steps in the response process by observing response times and eye fixations for each part of the question (see Figure 1).

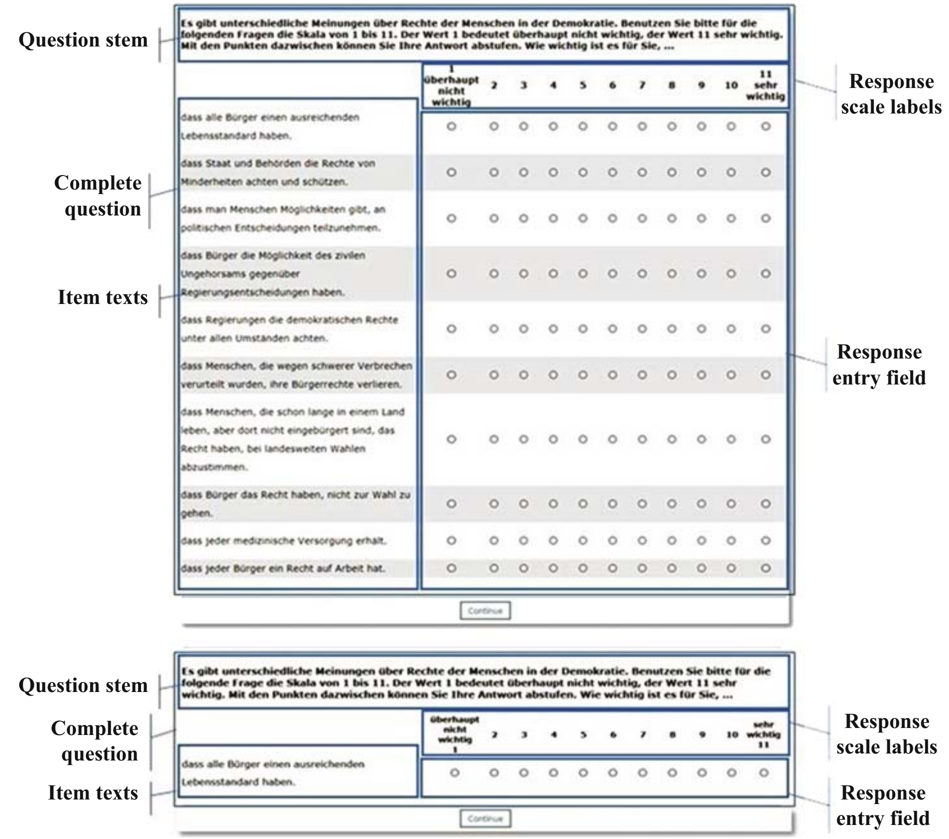
2.1. *Hypotheses for the Stage of Question Comprehension*

Comprehension includes such processes as attending to the question and instructions and identifying the information sought (Tourangeau et al. 2000). For grid questions, the comprehension stage requires respondents to attend to the question stem and the item texts. Particularly, the grid format promises efficient processing of the "question stem" compared to an item-by-item presentation. While in the former question design, respondents must attend to the question stem only once, they need to check whether the question is the same for each item in the item-by-item format. Thus, the more items are grouped on a single page of a questionnaire; the less the relative effort respondents need to invest in processing the question stem.

Hypothesis 1

"The fewer the number of items presented on a survey page; the more time is spent on processing the question stem on average across all target items."

Accordingly, the time spent processing the question stem should be highest for an item-by-item format, where the question stem is repeatedly displayed with each item, and



Es gibt unterschiedliche Meinungen über Rechte der Menschen in der Demokratie. Benutzen Sie bitte für die folgenden Fragen die Skala von 1 bis 11. Der Wert 1 bedeutet überhaupt nicht wichtig, der Wert 11 sehr wichtig. Mit den Punkten dazwischen können Sie Ihre Antwort abstufen. Wie wichtig ist es für Sie, ...

	1 überhaupt nicht wichtig	2	3	4	5	6	7	8	9	10	11 sehr wichtig
dass alle Bürger einen ausreichenden Lebensstandard haben.	○	○	○	○	○	○	○	○	○	○	○
dass Staat und Behörden die Rechte von Minderheiten achten und schützen.	○	○	○	○	○	○	○	○	○	○	○
dass man Menschen Möglichkeiten gibt, an politischen Entscheidungen teilzunehmen.	○	○	○	○	○	○	○	○	○	○	○
dass Bürger die Möglichkeit des zivilen Ungehorsams gegenüber Regierungsentscheidungen haben.	○	○	○	○	○	○	○	○	○	○	○
dass Regierungen die demokratischen Rechte unter allen Umständen achten.	○	○	○	○	○	○	○	○	○	○	○
dass Menschen, die wegen schwerer Verbrechen verurteilt wurden, ihre Bürgerrechte verlieren.	○	○	○	○	○	○	○	○	○	○	○
dass Menschen, die schon lange in einem Land leben, aber dort nicht eingebürgert sind, das Recht haben, bei landesweiten Wahlen abzustimmen.	○	○	○	○	○	○	○	○	○	○	○
dass Bürger das Recht haben, nicht zur Wahl zu gehen.	○	○	○	○	○	○	○	○	○	○	○
dass jeder medizinische Versorgung erhält.	○	○	○	○	○	○	○	○	○	○	○
dass jeder Bürger ein Recht auf Arbeit hat.	○	○	○	○	○	○	○	○	○	○	○

Es gibt unterschiedliche Meinungen über Rechte der Menschen in der Demokratie. Benutzen Sie bitte für die folgende Frage die Skala von 1 bis 11. Der Wert 1 bedeutet überhaupt nicht wichtig, der Wert 11 sehr wichtig. Mit den Punkten dazwischen können Sie Ihre Antwort abstufen. Wie wichtig ist es für Sie, ...

	überhaupt nicht wichtig 1	2	3	4	5	6	7	8	9	10	sehr wichtig 11
dass alle Bürger einen ausreichenden Lebensstandard haben.	○	○	○	○	○	○	○	○	○	○	○

Fig. 1. Illustration of the different parts of the question and Areas of Interest (AOI) for the analysis of eye-tracking data. Above is an example of the question parts/AOIs for the grid question design of Question 2, below for the item-by-item design.

lowest for the presentation in one large grid, where the question stem is displayed only once.

In contrast, grouping more items on a page should not lead to more efficient processing of *item texts* under the condition of optimizing response behavior, that is, if respondents are able and motivated to thoroughly read and answer the items. However, grouping many items on a survey page can increase the actual or perceived complexity and burden of a grid question (Couper et al. 2013; Liu and Cernat 2018). The higher complexity of grid questions may discourage respondents, and thus increase the chances that they alter their response strategy to satisficing. In this regard, it seems plausible that the likelihood of superficial or incomplete processing of item texts increases with each additional item that is grouped on a survey page. This can be further reinforced because respondents can use the previously answered items as sources of information and orientation which allows them to answer the following items in a similar way without thinking thoroughly about each single one.

Hypothesis 2

"The fewer the number of items presented on a survey page; the more time is spent on processing the item texts on average, across all target items."

Accordingly, the time spent processing the item texts should be highest for an item-by-item format, and lowest in one large grid.

2.2. *Hypotheses for the Stage of Reporting an Answer*

The stage of reporting an answer includes two groups of processes: mapping the answer onto the response options and editing the response (Tourangeau et al. 2000). Regarding grid questions, the reporting and response selection stage particularly concerns the processing of the (numeric or verbal) response scale labels and the response entry fields (i.e., the response options). As for the processing of the question stem in the comprehension stage, the grid format promises efficient processing of the "response scale labels" compared to an item-by-item presentation. In the former, respondents must attend to response scale labels only once, whereas they need to repeatedly check whether they have changed or not in the item-by-item format. Thus, the more items are grouped on a single survey page using a common set of response scale labels; the relatively less effort respondents need to invest in processing them per item; thereby increasing item-efficiency.

Hypothesis 3

"The fewer the number of items presented on a survey page; the more time is spent on processing the response scale labels on average, across all target items."

Accordingly, the time spent processing the response scale labels should be highest for an item-by-item format, and lowest for the presentation in one large grid.

For the "response entry fields", the higher complexity that results from the larger size of the grid most likely increases the chance that respondents experience navigational difficulties in reporting responses. In other words, the fewer items are presented on a survey page, the easier it should be for respondents to select the answer that applies to them from the available response entry fields.

Hypothesis 4

"The fewer the number of items presented on a survey page; the less time is spent on processing the response entry fields on average, across all target items."

Also, the grouping of items in grids may encourage respondents to edit their responses for inter-item consistency or other criteria. This would additionally increase the processing time of the response entry fields in a grid.

However, some respondents may be discouraged by the daunting size of the grid and alter their response behavior to satisficing. As a consequence, the likelihood of incomplete or careless processing of the response entry fields might increase with each additional item that is grouped on a survey page. Thus, satisficing in grids may to some extent offset the higher processing time that results from "response editing".

Besides fixation durations, it is also relevant to investigate survey responding that is related to satisficing, such as nondifferentiation (Couper et al. 2013; Roßmann et al. 2018; Zhang and Conrad 2014). In line with the assumption that presenting items together increases the likelihood that they are perceived and answered in the same context, previous research has found that respondents differentiated their answers more when answering item-by-item and less when the items were presented in grids (e.g., Roßmann et al. 2018).

Hypothesis 5

"Answering items in grids, compared to item-by-item formats, leads to less differentiation (e.g., more straightlining) in the responses, across all target items."

To better understand the response process when grid questions are answered, we also analyzed additional indicators of respondent behavior. First, we examined whether respondents answered the items in grids sequentially. We defined responding as sequential, when a respondent read and answered the first item, then read and answered the second item, then the following item, and so on until the last item on the page. Conversely, non-sequential responding involves skipping items or going back to previous items while reading through the list (see Figure 3 for examples). We also observed if respondents read all or several items on a survey page before starting to answer and whether respondents changed their response to an item after having read other items. For these different response behaviors, we proceeded exploratively and did not postulate hypotheses.

3. Methods

3.1. Experimental Design

In this study, we implemented a question format experiment with three questions. Each of these questions was presented either as a single large grid question with ten items on one page, as two small grids with five items on each of the two pages, or in an item-by-item design, in which each of the ten items was presented on a separate page. The respondents were randomly assigned to one of the three formats for each of the three questions (see Table A.1 in Online Appendix A for details on respondents' sociodemographic characteristics per question). Further, the questions were either presented with a five-point response scale or a 11-point response scale with labels at the end points and numbers in between. As the length of the scale was not the focus of the present research, and as the randomization regarding the response scale length was independent of the randomization regarding question format, we combined the two response scales for the comparison across formats presented here (Question 1 "Trust in Institutions": $\chi^2 = .46$; df $= 2, p = .978$; Question 2 "People's rights: $\chi^2 = .062$; df $= 2$, $p = .970$; Question 3 "BFI-10": $\chi^2 = .110$; df $= 2$, $p = .947$; see Tables A.2 and A.3 in Online Appendix A for an overview of the main results by response scale length).

3.2. Survey Questions

To ensure comparability between questions, we implemented three questions with ten items each. In surveys, grid questions with up to ten items are often used, and this number of items can still be presented on one page of a personal computer without scrolling (see Toepoel et al. 2009). Also, ten items could easily be split up into two almost equally sized grids with five items on each page. We selected published scales that differed in item text length. The first set of items asked about "trust in institutions" (Question 1; GLES 2019). The items are very short and state different institutions such as the European Commission or the Federal Constitutional Court (see Online Appendix B for question wordings of the three sets of questions). The second set of items asked about "people's rights in a democracy" (Question 2; ISSP Research Group 2016). The items consist of rather long sentences. The third set of items is the "BFI-10" (Question 3), a ten-item scale measuring the Big Five personality traits extraversion, agreeableness, conscientiousness, emotional stability, and openness (Rammstedt and John 2007). The items include complete sentences, but the statements are relatively short.

3.3. Participants

The study was conducted at GESIS – Leibniz Institute for the Social Sciences in Mannheim, Germany, between April and May of 2017. We recruited 132 respondents from the respondent pool maintained by the institute or by word of mouth. An equal share of women and men was recruited, but no quotas for other demographics such as age and education were set. However, the intention was to obtain a sample as diverse as possible. Technical difficulties prevented recording of eye movements for one respondent, and in each of the questions the eye-tracking data of 13 to 18 respondents were of no satisfactory quality as we observed shifts between the text on the screen and the eye gaze data. These respondents were excluded from the analyses, leaving 103 respondents with good quality of recordings in all three experimental questions and 125 respondents with good recordings in at least one question. Of those 125 respondents, 51% were female; 38% were between 18 and 24 years, 27% between 25 and 34 years, 11% between 35 and 44 years, 10% between 45 and 54 years, 8% between 55 and 64 years, and 5% were 65 years or older; 6% had a school-leaving certificate from lower secondary education after 9 years of education ("Volks-/ Hauptschulabschluss" – ISCED Level 244), 22% from lower secondary education after ten years of education ("Mittlere Reife/Realschulabschluss"– ISCED Level 244), and 71% from upper secondary education providing access to tertiary education ("Fachhochschulreife/Allgemeine Hochschulreife" – ISCED Level 344) or tertiary/university education ("Universitätsabschluss" – ISCED Level 64 or 74). More than one third of the participants (38%) had participated in at least one web survey during the last three months. To evaluate the effectiveness of random assignment and the sample composition across conditions, we conducted several χ^2-tests for the reported sociodemographic characteristics mentioned previously. Except for sex in Question 3, no significant differences between sociodemographic characteristics were observed (see online Table A.1). To ensure that this does not affect our conclusions; we included sex as covariate in the analyses of response times, fixation durations, and response differentiation for Question 3.

3.4. Eye-Tracking Equipment and Procedures

We used the Senso Motoric Instruments (SMI) RED250 mobile Eye Tracker to record participants' eye movements and "BeGaze" version 3.6.57 for data analysis. The RED250 mobile Eye Tracker was mounted on the bottom frame of a 22″ TFT desktop monitor (resolution 1280x1024). The documentation of the RED250 mobile describes its accuracy to be within 0.4° and its tracking range of 32x21 at 60 centimeters distance. 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 categories, respectively. The online questionnaire did not feature a "back" button.

Before the web survey started, respondents completed a calibration exercise (in which they followed black circles displayed at nine different points of the screen with their eyes). The questionnaire, which contained several experiments, took on average 30 minutes to complete. During this time, an experimenter stayed in the room next door to observe respondents' eye movements on a second computer screen for reasons of quality assurance. Participants were paid an incentive of EUR 20 for taking part in the study.

3.5. *Measures and Analytical Strategies*

We tested our hypotheses on the effects of question design on cognitive processing by analyzing indicators of cognitive effort measured by response times and eye-tracking data. For collecting *response times,* we used UCSP, Universal Client Side Paradata (Kaczmirek 2005). Response times were measured in milliseconds from the time a question appeared on the screen to the time respondents clicked on the next button to move on to the next question. For the small grid (two pages) and the item-by-item condition (ten pages), response times from the individual survey pages were summed up. Eye-tracking data provide information on the question answer process by recording where respondents look, for how long, and in what order while reading question stems, item texts, and response options (Galesic and Yan 2011; Romano Bergstrom and Schall 2014). Eye-tracking can be used as a proxy for depth of cognitive processing (Rayner 1998). The analysis of eye movements is based on two common assumptions (Just and Carpenter 1980; Rayner 1998). The first one, called the "immediacy assumption", states that objects fixated by the eyes are processed immediately (i.e., the mind follows the eye). The second one, called the "eye-mind assumption", states that the eye remains fixated on an object, as long as it is being processed (i.e., the eye follows the mind). Taken together, these two assumptions state that there is a close relationship between fixation duration and processing duration. A longer fixation duration indicates a longer response process. A long response process can be due to thorough consideration and recalling, but it can also indicate difficulties during the answer process. Those difficulties might arise from unknown or difficult terms, difficulties in arriving at an answer or selecting one of the response options (Galesic and Yan 2011; Kamoen et al. 2017; Neuert and Lenzner 2017). To measure "cognitive effort", we compare fixation durations on predefined areas of interest (AOI) to be able to compare these measures across the different question formats. Each question was conceptionally divided in five AOIs: (1) the complete question, (2) the question stem, (3) response scale labels, (4) item texts, and (5) response entry fields (see Figure 1). The AOIs on each individual page in the small grid and item-by-item format were summed up to compare fixation durations across formats. For response times and fixation durations, we excluded those respondents from the analyses who had response times below or above the mean plus/minus two standard deviations (see, e.g., Mayerl 2013).

To determine whether cognitive effort measured by response latencies and fixation durations were associated with the question format, we employed OLS regression models.

To study respondents' response behavior, we investigated how much respondents varied their answers to the items within the experimental questions. Nondifferentiation is found when respondents do not differentiate in their answers but give similar (or identical) responses to all items. The level of differentiation can be investigated by the probability of differentiation P_d (Krosnick and Alwin 1998), which indicates the variability of the responses. P_d is calculated as $P_d = 1 - \sum_{i=1}^{n} P_i^2$, where P_i is the proportion of the values rated on a given point of a response scale and n is the number of rating points. If $P_d = 0$, respondents answered all items by selecting the same response, while a higher P_d means that different response options were given. We also measured the coefficient of variation (CV) as an indicator of the extremity of the responses (McCarty and Shrum 2000). CV is computed as $CV = \frac{s}{\bar{x}}$, where s is the standard deviation and $\bar{x}$ is the mean of the responses over items. The CV indicates the distance between the responses given. A CV of zero indicates straightlining response behavior, while larger values indicate that respondents differentiated their answers to a greater extent. As a measure of perceived difficulty, respondents were asked after each experimental question to rate how difficult answering the question was on a fully labeled five-point scale ranging from "extremely difficult" to "not difficult at all." Analyses were conducted using Stata version 16.1.

To further analyze respondents' behavior, two student assistants coded the eye gaze videos with regard to the following response patterns: (1) sequential responding, that is, did respondents answer grid questions in a sequential order; answering one item following the next?; (2) how many items did respondents read before answering the first item?; (3) answer change, that is, did respondents change their response after having read other items? Agreement between the two raters was 95% and Cohen's Kappa (1960) was found to be .87, which is "almost perfect", according to Landis and Koch's (1977, 165) criteria. To make those response patterns comparable between the small and large grid format, we summed the results for the two small grids so that both numbers are based on ten items for each question. For sequential responding, we included two measures for the small grid: answering both pages of the small grid sequentially and answering at least one of the two pages sequentially. As the response behavior is only comparable across grid questions, which present several items on the same survey page, these analyses were restricted to the two grid formats.

4. Results

4.1. Overall Cognitive Effort – Response Times and Question Fixation Durations

Before considering fixation durations on specific parts of the questions, we compared response times and fixation durations for the complete question as indicators of overall cognitive effort by question format. Response times were measured from the time a question appeared on the screen to the time respondents clicked on the next button to move on to the next question. Hence, anything that respondents did in between is included in this indicator. In contrast, fixation duration corresponds to the time a respondent spent fixating the question which might therefore be a more accurate measure of cognitive question processing (Just and Carpenter 1980; Staub and Rayner 2007). With regard to response times (in seconds) for Questions 1 ("Trust in Institutions") and 2 ("People's rights"), there was the general trend observable that respondents needed the least amount of time when the items were presented in a large grid (ten items per page) followed by small grids (five-items per page), and the most amount of time when the questions were presented item-by-item. However, the differences were not statistically significant (see Table 1). For Question 3 ("BFI-10"), our analysis showed that response times were significantly shorter

Table 1. Means and standard errors of cognitive effort indicators.

Cognitive effort indicators	Large grid with ten items (1 page)		Small grid with 5 items each (2 pages)		Item-by-item (10 pages)		Test	
	M	*(SE)*	*M*	*(SE)*	*M*	*(SE)*	*F value*	*p*
Response times								
Question 1	49.8	30.6	55.2	32.2	57.5	31.8	1.16	.204
Question 2	94.2	49.5	101.9	51.2	107.9	57.3	1.69	.189
Question 3	58.7^{c}	28.7	54.9^{c}	28.7	71.1^{ab}	29.6	5.79	.001
Fixation durations – Complete question								
Question 1	38.3	25.8	37.8	27.6	44.9	27.6	2.06	.132
Question 2	73.2	42.1	78.7	43.7	79.7	49.8	0.62	.539
Question 3	46.6^{c}	26.0	41.5^{c}	26.7	54.4^{ab}	26.9	3.76	.013
Fixation durations – Question stem								
Question 1	3.9^{bc}	.62	7.5^{a}	.62	6.7^{a}	.65	9.45	.001
Question 2	8.2^{bc}	.87	12.2^{a}	.88	12.5^{a}	1.03	7.02	.001
Question 3	1.4^{c}	.27	1.9	.32	2.8^{a}	.28	5.86	.001
Fixation durations – Item texts								
Question 1	11.2	1.2	10.7	1.4	8.9	1.5	1.46	.238
Question 2	35.1	2.4	34.5	2.5	37.4	2.6	.35	.703
Question 3	18.4	1.1	15.4	1.2	18.1	1.2	1.30	.279
Fixation durations – Response scale labels								
Question 1	3.8^{c}	.47	4.5^{c}	.49	7.7^{ab}	.54	15.6	.001
Question 2	4.7^{bc}	.55	6.4^{ac}	.56	10.0^{ab}	.66	19.1	.001
Question 3	5.9^{c}	.92	6.9^{c}	1.1	12.0^{ab}	.97	8.03	.001
Fixation durations – Response entry fields								
Question 1	15.6	1.0	14.5	1.0	12.9	1.1	1.72	.184
Question 2	17.4^{c}	.89	18.4^{c}	.89	13.7^{ab}	.96	7.02	.001
Question 3	19.2 ^{bc}	1.0	15.1^{a}	1.1	13.3^{a}	1.0	5.56	.001

Note: Question 1 = Trust in institutions, Question 2 = People's Rights, Question 3 = BFI-10. Reported are estimated marginal means from linear regression models. For Question 3, we report estimated marginal means controlling for sex. Superscripts present a significant difference ($p < .05$) compared to (a) large ten-item grid, (b) small five-item grids, or (c) item-by-item presentation. To compare the response times and fixation durations across formats, response times and fixation durations from the individual survey pages in the small grid and in the item-by-item design were summed up.

for the two grid designs ($M_{large} = 58.7$; $M_{small} = 54.9$) than for the item-by-item design ($M_{single} = 71.1$; $F = 5.79$; $p = .001$).

Comparing fixation durations on the complete question across designs showed a similar relationship. Fixation durations were slightly higher for the item-by-item presentation than for the grid formats, however, the differences were not statistically significant for Questions 1 and 2. For Question 3, we found statistically significant differences for both the small ($M_{small} = 41.5$) and the large grid ($M_{large} = 46.6$; $t = 2.06$, $p = .041$) compared to the item-by-item format ($M_{single} = 54.4$; $t = 3.31$, $p = .001$).

4.2. Question Fixation Durations for the Stage of Question Comprehension

Regarding the stage of question comprehension, our results showed that cognitive processing of the "question stem" was lowest for the large grid in all three experimental questions, followed by the presentation in two small grids with five items each, and the presentation in the item-by-item-design in Questions 2 ("People's rights") and 3 ("BFI-10"), as expected in Hypothesis 1 (see Table 1). For Question 1 ("Trust in institutions"), fixation durations on the question stem were higher in the small grids than in the item-by-item format (although not significantly different).

Contrary to Hypothesis 2, which stated that the depth of processing of "item texts" is expected to decrease with the number of items presented on a survey page, we did not find any significant differences across designs.

4.3. Question Fixation Durations for the Stage of Reporting An Answer

Regarding the depth of processing of the "response scale labels", we found that fixation durations were significantly higher for the item-by-item format than for the presentation in both the large and the small grid in all three questions. This is in line with Hypothesis 3. However, the expected relation that depth of processing increases, the fewer items of the question are presented on a survey page, only holds true for Question 2 ("People's rights": $M_{large} = 4.7$; $M_{small} = 6.4$; $M_{single} = 10.0$; $F = 19.1$; $p = .001$). For Questions 1 ("Trust in institutions") and 3 ("BFI-10"), there were no statistically significant differences between the two grid designs.

Regarding the stage of reporting and response selection, we did not find that the fewer items presented on a survey page, the less time is spent processing the "response entry fields" as expected in Hypothesis 4. Respondents processed the response entry fields more extensively in the large grid design than in the item-by-item-design in Questions 2 ("People's rights") and 3 ("BFI-10"). For the small grids, findings were mixed. In Question 1 ("Trust in institutions"), we did not find statistically significant differences in fixation durations.

4.4. Observations of Response Behavior

We also investigated two indicators of satisficing, the probability of differentiation (P_d) and the coefficient of variation (CV). The results are shown in Table 2. Regarding the indicator of differentiation P_d, we did not find any significant differences across the three presentation formats.

Table 2. Indicators of satisficing (means)

Indicators of satisficing	Large grid with 10 items (1 page)		Small grid with 5 items each (2 pages)		Item-by-item (10 pages)		Test	
	M	*(SE)*	*M*	*(SE)*	*M*	*(SE)*	*F value*	*p*
P_d								
Question 1	.55	.030	.51	.031	.56	.031	0.94	0.39
Question 2	.63	.023	.61	.023	.66	.025	1.05	0.35
Question 3	.74	.016	.74	.016	.74	.016	0.32	0.81
CV								
Question 1	.22	.022	.22	.022	.21	.022	0.04	0.95
Question 2	.32	.016	.33	.017	.34	.019	0.19	0.83
Question 3	.46^c	.019	.42	.019	.40^a	.018	1.62	0.19

Note: Question 1 = Trust in institutions, Question 2 = People's rights, Question 3 = BFI-10. Reported are estimated marginal means from linear regression models. For Question 3, we report estimated marginal means controlling for sex. Superscripts present a significant difference ($p < .05$) compared to (a) large 10-item grid, (b) small 5-item grids, or (c) item-by-item presentation.

For the indicator CV, we also did not find significant differences across the three experimental conditions for Questions 1 (“Trust in institutions”) and 2 (“People’s rights”). For Question 3 (“BFI-10”), however, we observed that the variation of answers across the items was lower in the item-by-item design ($M_{single} = .40$) than in the large grid design ($M_{large} = .46$, $t = -2.07$, $p = .040$). Hence, Hypothesis 5, that answering items in grids, compared to item-by-item formats, leads to less differentiation in the responses cannot be confirmed in the current study. Figure 2 illustrates two different response styles, with the respondent on the left-hand side showing no differentiation in the responses selected (straightlining) while the respondent on the right-hand side differentiated the responses to a greater extent. Observing the gaze videos provided the interesting finding that respondents may actually spend more cognitive effort than one might initially expect when showing a straightlining response behavior.

Finally, we analyzed how respondents answered the items presented in the two grid formats to gain more knowledge about response patterns when answering grid question formats. For the small grid, we report the number of respondents who answered the ten items displayed on two pages sequentially, and the number of respondents who answered at least one page with five items in a sequential order (see Table 3). Comparing sequential responding for all ten items, we did not observe differences between the large and the small grid for Questions 1 (“Trust in institutions”: 37% large vs. 29% small) and Question 2 (“People’s rights”: 42% large vs. 40% small). However, for Question 3 (“BFI-10”), we found that half of the respondents in the large grid (51%) answered the items sequentially, while only 12% in the small grid condition did show this response behavior. When considering those respondents in the small grid condition who answered at least one page with five items in a sequential order, we observed that this amount is quite large, between 57% and 68%, but no statistically significant differences compared to the large grids were found.

Figure 3 illustrates both a sequential as well as a non-sequential response style. As shown by the eye movement patterns, the respondents on the left-hand side read and answered the items sequentially one by one; the respondents on the right-hand side first read all item texts and then moved to the response options and answered all items one after another or jumped back and forth on the question parts.

We also investigated whether respondents perceived the items in a grid as belonging conceptually together. The response behavior that respondents read all items before beginning to answer them only occurred in Question 1 (“Trust in institutions”: 16% large vs. 41% small). In Questions 2 and 3, respondents did not read through all items before selecting a response for the first item. Comparing how many items respondents read on average before they started answering, revealed that respondents read on average between one and three items in the large grid and between two and four items in the small grid condition (Question 1 “Trust in institutions”: 3.2 large vs. 4.4 small; Question 2 “People’s rights”: 1.2 large vs. 2.4 small; Question 3 “BFI-10”: 1.1 large vs. small 2.5 small). It must be noted that the average number of items read on the two separate pages of the small grids were summed up, resulting in a systematic overestimation compared to the large grid conditions.

Interestingly, the overall number of items read before beginning to answer is higher in the question asking for “trust in institutions” (Question 1) than in the other two questions. Notably, the question asking for “trust in institutions” had the shortest item text regarding the number of characters but was perceived as most difficult by respondents ($M = 2.11$)

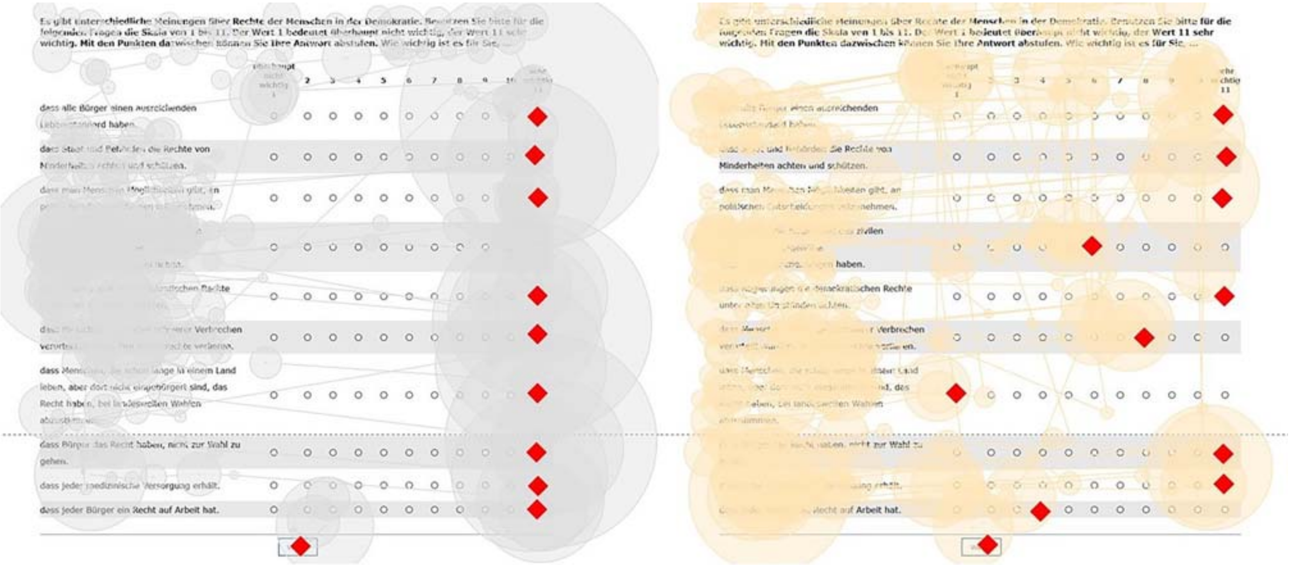

Es gibt unterschiedliche Meinungen über Rechte der Menschen in der Demokratie. Benutzen Sie bitte für die folgenden Fragen die Skala von 1 bis 11. Der Wert 1 bedeutet überhaupt nicht wichtig, der Wert 11 sehr wichtig. Mit den Punkten dazwischen können Sie Ihre Antwort abstufen. Wie wichtig ist es für Sie, ...
überhaupt nicht wichtig 1
2
3
4
5
6
7
8
9
10
sehr wichtig 11
dass alle Bürger einen ausreichenden Lebensstandard haben.
dass Staat und Behörden die Rechte von Minderheiten achten und schützen.
dass man Menschen Möglichkeiten gibt, an politischen Entscheidungen teilzunehmen.
haben.
demokratischen Rechte
Verbrechen
verurteilt
dass Menschen, die schon lange in einem Land leben, aber dort nicht eingebürgert sind, das Recht haben, bei landesweiten Wahlen abzustimmen.
dass Bürger das Recht haben, nicht zur Wahl zu gehen.
dass jeder medizinische Versorgung erhält.
dass jeder Bürger ein Recht auf Arbeit hat.

Fig. 2. Gaze Plots of two different respondents (large grid condition) showing either straightlining behavior or not.

Note: The figure shows the gaze plots of two different respondents answering Question 2. Each gaze plot displays the eye movements of one respondent. The circles indicate fixations, and the lines between the circles indicate saccades. The size of the circle is proportional to the fixation time, meaning larger circles indicate longer fixations. The squares indicate the mouse clicks of the respondents. The respondent on the left-hand side showed straightlining behavior; the respondent on the right-hand side differentiated responses to a greater extent.

Table 3. Response behavior for grid formats, by question and grid size.

Question	1 – Trust in Institutions		2 – People's rights		3 – BFI-10	
	Large grid with 10 items (1 page)	Small grid with 5 items each (2 pages)	Large grid with 10 items (1 page)	Small grid with 5 items each (2 pages)	Large grid with 10 items (1 page)	Small grid with 5 items each (2 pages)
% Sequential responding						
– Yes, all 10 items	37.2 (16)	28.6 (12)	42.2 (19)	39.5 (17)	51.2 (22)	12.2 (5)
	χ^2=.718; df=1		χ^2=.066; df=1		χ^2=14.61; df=1**	
– Yes, at least on 1 page with 5 items	37.2 (16)	57.1 (24)	42.2 (19)	60.5 (26)	51.2 (22)	68.3 (28)
	χ^2=3.39; df=1		χ^2=2.93; df=1		χ^2=2.56; df=1	
Mean number of items read before beginning to answer						
	3.2 (43)	4.4 (42)	1.2 (45)	2.4 (43)	1.1 (43)	2.5 (41)
	F(1,83)=.412		*F*(1,86)=6.011**		*F*(1,82)=27.083**	
% All items read before beginning to answer						
– No	83.7 (36)	59.5 (25)	100 (45)	97.7 (42)	100. (43)	97.6 (40)
– Yes	16.3 (7)	40.5 (17)	0 (0)	2.3 (1)	0 (0)	2.4 (1)
	χ^2=6.139*; df=1		χ^2=1.059; df=1		χ^2=1.061; df=1	
% Answer change						
– No	72.1 (31)	78.6 (33)	71.1 (32)	81.4 (35)	81.4 (35)	87.8 (36)
– Yes	27.9 (12)	21.4 (9)	28.9 (13)	18.6 (8)	18.6 (8)	12.2 (5)
	χ^2=.479; df=1		χ^2=1.280; df=1		χ^2=.659; df=1	
n	43	42	45	43	43	41

Note: * $p < .05$; ** $p < .01$; Parenthetical entries are cell sizes. For sequential responding, we report both the comparison of the large grid with (1) reading all ten items of the small grids (summed up) and with (2) answering at least one page of two sequentially.

compared to Question 2 (M = 1.54, $p < .001$), and Question 3 (M = 0.81, $p < .001$). This finding can be interpreted as an indication that respondents used the information from the item texts of several other items to answer each item of the grid question.

Finally, we observed whether respondents tended to change their answers given to one item after having read other items in the grid. Across all three experimental questions between 16% and 25% of respondents changed at least one answer after having read the following items in the related grid (Question 1 "Trust in institutions": small 21% vs. 28% large; Question 2 "People's rights": small 19% vs. 29% large; Question 3 "BFI-10": small 12% vs. 19% large). Together with the findings on sequential reading and reading the item text of several other items before starting to answer, this might indicate that respondents used the additional information provided by the remaining items in a grid to give a response by applying the "near means related" heuristic (Tourangeau et al. 2004).

5. Discussion

5.1. Summary and Discussion of Findings

This study investigated the depth of cognitive processing when answering different grid question or item-by-item formats. We implemented three questions with ten items each in an experiment and tracked respondents' eye movements while they answered a web survey in

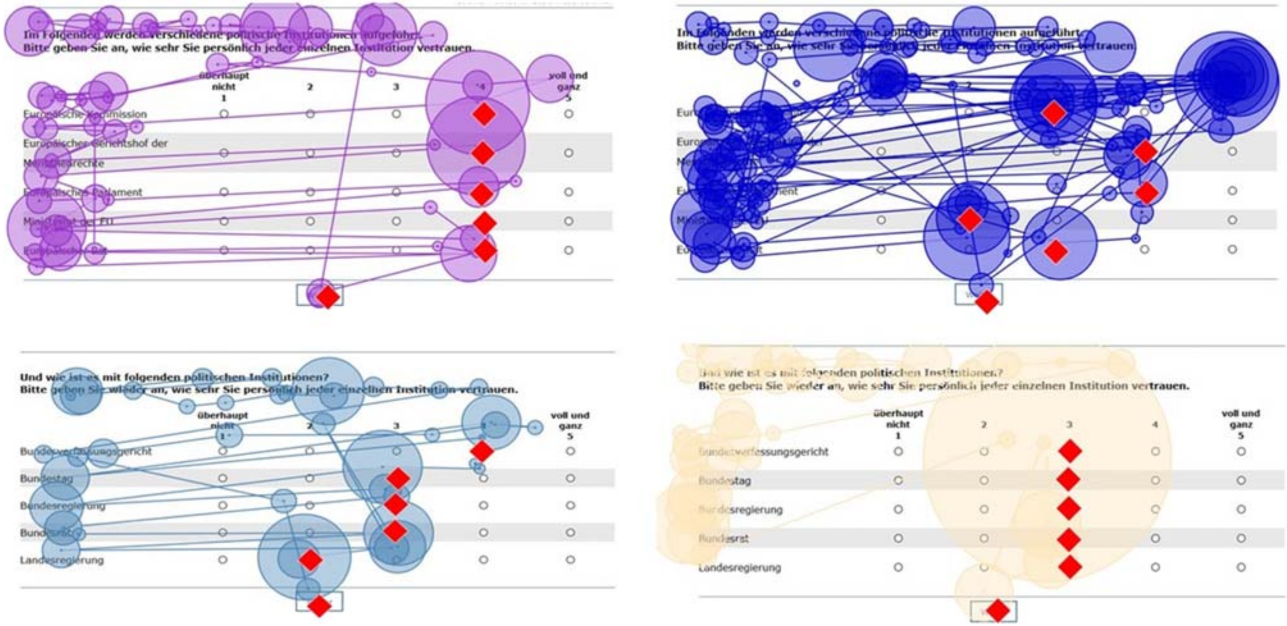

Im Folgenden werden verschiedene politische Institutionen aufgeführt.
Bitte geben Sie an, wie sehr Sie persönlich jeder einzelnen Institution vertrauen.
überhaupt nicht 1
2
3
4
voll und ganz 5
Europäische Kommission
Europäischer Gerichtshof der Menschenrechte
Europäisches Parlament
Ministerrat der EU
Europäischer Rat
Und wie ist es mit folgenden politischen Institutionen?
Bitte geben Sie wieder an, wie sehr Sie persönlich jeder einzelnen Institution vertrauen.
Bundesverfassungsgericht
Bundestag
Bundesregierung
Bundesrat
Landesregierung

Fig. 3. Gaze Plots of different respondents (small grid condition) representing sequential (left-hand side) and non-sequential response styles (right-hand side).

Note: The figure shows the gaze plots of four different respondents answering Question 1. Each gaze plot displays the eye movements of one respondent. The circles indicate fixations, and the lines between the circles indicate saccades. The size of the circle is proportional to the fixation time, meaning larger circles indicate longer fixations. The two respondents on the left-hand side read and answered the items sequentially one by one; the respondents on the right-hand side showed non-sequential responding: above in the form of jumping back and forth on the question parts; below in the form of first reading all item texts and then moving to the response options and answering them in a row.

which the presentation was varied. For each question, we randomly assigned respondents to three question formats (item-by-item, small grid, or large grid). The eye-tracking data studied showed that the previous finding of longer response times (e.g., Callegaro et al. 2009; Couper et al. 2001; Roßmann et al. 2018; Toepoel et al. 2009) in the item-by-item format could be attributed to more extensive processing of the question stem and the response scale labels compared to the grid formats (Hypothesis 1 and 3). In contrast, we did not find differences with respect to processing of the item texts (Hypothesis 2). This indicates that respondents do not spend more time on the item texts in either format but need time to adjust to the new question context when the items are presented in an item-by-item format. Specifically, they have to read the same question stem and response scale labels multiple times, since they are presented to them with each item. Those findings suggest that the item-by-item format increases response burden compared to the grid formats.

With regard to the response process stages of reporting and response selection (Tourangeau et al. 2000), we observed that fixation durations on the area of the response entry fields were significantly longer in the large grid than in the item-by-item presentation in two out of three questions (Hypothesis 4). Since those two questions had longer item texts, one possible explanation might be that navigating within a grid is more difficult for long items than when the items are presented item-by-item on separate pages. Hence, selecting and reporting a response seems less burdensome for respondents in the item-by-item design.

By using eye-tracking methodology, we were able to observe the behavior of respondents more directly while they were answering the grid questions. These analyses suggest that respondents apply the “near means related” heuristic (Tourangeau et al. 2004; Silber et al. 2018), which is grounded in the proximity principle from Gestalt psychology (Koffka 1935; Wertheimer 1923). According to the principle of proximity, placing objects close to each other will let them be perceived as a group, and hence as not only physically but also conceptually related (Dillman et al. 2014). Consequently, items presented in the grid format were likely perceived, processed, and answered in the same context. Presenting multiple items together on a page can facilitate respondents’ cognitive processing. If the respondent is not familiar with the topic or when the meaning of the question is not clear, respondents might try to capture the content using the surrounding items to improve their understanding of the question (Krosnick and Presser 2010).

Consistent with the “near means related” heuristic, many respondents in our survey did not respond to the grid questions sequentially but instead read multiple items before answering the first item. They also changed their answers later after reading other questions, suggesting that they reconsidered their answers after answering other items. Such response behavior was more pronounced for items presented in one large grid than for items presented in two small grids but was also visible there. For related items, grouping them may improve measurement (Krosnick and Presser 2010), for example, by increasing the consistency of responses among items and inter-item correlations (Couper 2008; Heerwegh 2009; Toepoel et al. 2009). In contrast, the grouping may also have negative effects on measurement. Although we did not observe differences regarding non-differentiation in our study (Hypothesis 5), previous research has consistently shown that a separate presentation like in an item-by-item design reduced undesired response effects, such as non-differentiation or item nonresponse (e.g., Roßmann et al. 2018; Toepoel et al.

2009). Also, higher inter-item correlations may be due to measurement error (Peytchev 2005, cited in Couper et al. 2013).

Across the three questions, we also observed some differences regarding response behavior, which might be related to the content of those questions. While the item texts of Question 1 only featured the names of institutions, the item texts of Questions 2 and 3 consisted of full sentences. The shorter item texts of Question 1 may have led more respondents to read multiple or even all item texts before answering the first item. Since our study did not experimentally vary the item length, further research is needed to understand the relationship between response behavior and item length in grid formats.

With respect to response burden, we found that both the item-by-item format and the grid format entail different burdensome elements. The item-by-item format requires respondents to adjust to a new context for each item. Specifically, presenting the question stem and response scale labels on each page makes respondents undergo repetitive reading tasks, which lowers response efficiency. In contrast, the grid format increases the complexity because respondents are confronted with multiple items on a single survey page. This may increase response burden due to navigational difficulties in the process of reporting the responses. With respect to response behavior, this study showed that respondents do not answer the items necessarily from top to bottom and in the presented order. Instead, some respondents read several items before beginning to answer the first item. Some respondents also changed their responses after they had read the following items. The extent of this behavior depended on the item texts of the questions. Finally, we observed that some respondents selected the identical responses for all items in a grid (i.e., straightlining) but still read the item texts attentively. With respect to strong satisficing, we expect that respondents skip the question comprehension stage altogether and provide responses haphazardly. Yet, the observation in the current study also fosters the notion of weaker forms of satisficing, in which respondents attend to the question stem and item text, but then decide to simplify the perceivably difficult task of reporting accurate and meaningful answers, for instance, by resorting to the "near means related" heuristic (Tourangeau et al. 2004; Silber et al. 2018). Hence, it might be worthwhile to investigate this and similar response behaviors in more depth with respect to how respondents arrive at selecting the same answer to all questions of an item sequence. Due to the low number of straightlining respondents in the lab setting, this study did not allow us to investigate this pattern further.

5.2. Limitations and Avenues for Future Research

Our study has several limitations. The most important is external validity since we designed our investigation as a lab study. Filling-out a questionnaire in a lab situation, in which the eye movements of respondents are recorded, may not perfectly reflect the behavior of respondents in a common survey interview environment. Also, the participants might have been rather engaged as they were willing to participate in a lab experiment and received an incentive of EUR 20 for their participation. Hence, the differences in cognitive processing might be less pronounced than in studies conducted in common survey settings. For instance, we did not find differences with respect to nondifferentiation, even though many previous studies have shown such differences (e.g., Mavletova et al. 2018; Roßmann et al. 2018; Tourangeau et al. 2004). Likewise, fixation durations may have been overestimated, and the number of respondents engaging in response behaviors such as non-sequential responding or answer changes underestimated. Yet, those or similar response behaviors can be expected to occur outside the lab as well, since previous research found answer patterns that were likely due to answering items in the same context when they were presented in a grid (e.g., Couper et al. 2001; Toepoel et al. 2009; Tourangeau et al. 2004). Second, also attributable to being a lab study, is the comparatively smaller sample size than, for instance, in many online experiments. However, recruiting and testing more participants in an eye-tracking study would be laborious and expensive. A third limitation that should be addressed by additional research is that the scales we used were all endpoint-labeled with numbers in between. How the processing of the scales differs across grid formats when using fully labeled scales would be worthwhile examining in a follow-up study. Also, investigating the generalizability of our findings with a different number of items, such as eight or six items for the large grid and four or three for the small grid, is an avenue for future research. Fourth, we decided to place the two grids and every single item on separate pages in a so-called paging design. Future studies could explore whether similar results are obtained if they are presented on the same page in a scrolling design (see, e.g., Liu and Cernat 2018). A possible outcome of using the scrolling design could be that the items in the single item or the smaller grid formats might be more often answered in the same context due to the visual presentation on the same page. Finally, as the questions in our study were answered on a desktop PC screen only due to the eye-tracking system used, this study does not address the issue of responding on mobile devices, which seems to be another worthy avenue for future research.

5.3. Increasing Relevance of Mobile Devices in Web Surveys

Given the increasing number of respondents using smartphones or other mobile devices in answering web surveys (Gummer et al. 2019), design decisions on using grid versus item-by-item presentation have become increasingly important. This is especially true in the context of decisions regarding whether to use layouts that adapt to the device used by respondents (adaptive or responsive layouts) or to optimize layouts for use on a specific device (e.g., mobile first layouts). When grids are presented at full size on the small(er) screens of smartphones, this may require horizontal scrolling and zooming. Previous research has found that answering grids on smartphones compared to grids on personal computers increases breakoff rates and stimulates undesired response behaviors like straightlining (see Antoun et al. 2018 for a systematic review). In mobile-first unified designs and responsive designs, survey software often automatically adapts grids to screen size by converting them into a series of single items. Thus, some respondents will see the set of items as grids if they use a personal desktop or laptop computer, while other respondents will see them in an item-by-item format. This might result in systematic mode differences and measurement error. The same might apply to mixed-mode surveys, for example, paper versus web questionnaires (De Leeuw et al. 2018; Dillman et al. 2014). Respondents in our study answered grids solely on a desktop PC. Thus, we suggest that future studies could use eye-tracking methodology to investigate how adaptive (or responsive) layouts impact cognitive processing and response behavior in web surveys with multiple devices, and in particular, the suitability of grids on mobile devices.

5.4. Recommendations

Our findings have practical implications for researchers deciding between grids or item-by-item designs. In our study, both designs appear to come with format-specific limitations, which directly affect response burden. To ensure that the difference in question presentation format between respondents does not lead to potential measurement error, especially if smartphone participation in web surveys continues to rise, we recommend using the smallest screen as the basis of the format decision, which is in favor of the item-by-item design (see also Antoun et al. 2018; Liu and Cernat 2018; Mavletova et al. 2018). Another argument for the item-by-item design is that items in grid questions do not meet web content accessibility guidelines (WCAG; W3C 2018), which suggests that each question should be entirely understandable on its own. Yet, if a consistent use of item-by-item presentation is not possible, for instance due to restrictions in the available questionnaire length, we would recommend to break up larger grids into (as in our case two) smaller grids (see also Dillman et al. 2014) as they did not show substantial disadvantages compared to a large grid question, and they seem to be easier to navigate (this is in line with Grady et al. (2019) who recommend a small to medium grid size). Though, for some surveys, grid formats might be the best alternative (e.g., brand image research; Brosnan et al. 2021). This decision may depend on factors such as the question type, the complexity of the information, and the question content. For example, grid questions may help respondents quickly understand their response task for multiple items at once and thereby increase response efficiency.

6. Conclusion

This study showed that it takes respondents longer to answer a question in the item-by-item format than in the grid format because the former shows the question stem and the response scale repetitively, and respondents need to process both multiple times. The differences in the visual presentation and the shorter response times of grid questions did not result in more satisficing response behavior than in the item-by-item format, which might have been due to the lab setting in which participants are likely to be quite engaged. Finally, by using eye tracking, we were able to observe specific response styles (i.e., reading a few items before answering a grid and answer changes) when a question was presented in the grid format. An area for future research would be to investigate whether items presented in a grid format are more likely to be processed and answered in one context than when presented in an item-by-item format within a scrolling design, and whether these differences in cognitive processing and responding have a substantial impact on substantive analyses with the items is an area for future research.

7. References

Antoun, C., J. Katz, J. Argueta, and L. Wang. 2018. "Design Heuristics for effective Smartphone Questionnaires." *Social Science Computer Review* 36(5): 557–574. DOI: https://doi.org/10.1177/0894439317727072.

Brosnan, K., B. Grün, and S. Dolnicar. 2021. "Cognitive load reduction strategies in questionnaire design." *International Journal of Market Research* 63(2): 125–133. DOI: https://doi.org/10.1177%2F1470785320986797.

Callegaro, M., J. Shand-Lubbers, and J.M. Dennis. 2009 "Presentation of a Single Item versus a Grid: Effects on the Vitality and Mental Health Scales of the SF-36v2 Health Survey." 64th Annual Conference of the American Association for Public Opinion Research (AAPOR), May 14, 2009: 5887–5897. Hollywood, Florida. Available at: http://www.asasrms.org/Proceedings/y2009/Files/400045.pdf (accessed July 2022).

Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and psychological measurement* 20(1): 37–46. DOI: https://doi.org/10.1177/001316446002000104.

Couper, M.P. 2008. *Designing effective Web surveys*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511499371.

Couper, M.P., R. Tourangeau, F.G. Conrad, F.C. Zhang. 2013. "The Design of Grids in Web Surveys." *Social Science Computer Review* 31(3): 322–345. DOI: https://doi.org/10.1177/0894439312469865.

Couper, M.P., M.W. Traugott, and M.J. Lamias. 2001. "Web Survey Design and Administration." *Public Opinion Quarterly*: 65(2): 230–253. DOI: https://doi.org/10.1086/322199.

DeBell, M., C. Wilson, S. Jackman, and L. Figueroa. 2021. "Optimal Response Formats for Online Surveys: Branch, Grid, or Single Item?" *Journal of Survey Statistics and Methodology* 9(1): 1–24. DOI: https://doi.org/10.1093/jssam/smz039.

De Leeuw, E.D., Z.T. Suzer-Gurtekin, and J.J. Hox. 2018. "The Design and Implementation of Mixed-mode Surveys." In *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC*), edited by T.P. Johnson, B. Pennell, I.A.L. Stoop, and B. Dorer: 387–409. Hoboken: Wiley.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-mode Surveys: the Tailored Design Method* (4th edition). Hoboken: Wiley.

Galesic, M., and T. Yan. 2011. "Use of Eye Tracking for Studying Survey Response Processes." In *Social and Behavioral Research and the Internet*, edited by M. Das, P. Ester, and L. Kaczmirek: 349–370. New York: Taylor and Francis.

GLES. 2019. "Longterm-Online-Tracking, Cumulation 2009–2017 (GLES)." GESIS Data Archive, Cologne. ZA6832 Data file Version 1.1.0, DOI: https://doi.org/10.4232/1.13416.

Grady, R.H., R.L. Greenspan, and M. Liu. 2019. "What Is the Best Size for Matrix-Style Questions in Online Surveys?" *Social Science Computer Review* 37(3): 435–445. DOI: https://doi.org/10.1177/0894439318773733.

Gummer, T., F. Quoß, and J. Roßmann. 2019. "Does Increasing Mobile Device Coverage Reduce Heterogeneity in Completing Web Surveys on Smartphones?" *Social Science Computer Review* 37(3): 371–384. DOI: https://doi.org/10.1177/0894439318766836.

Heerwegh, D. 2009. "Mode Differences Between Face-to-Face and Web Surveys: an Experimental Investigation of Data Quality and Social Desirability Effects." *International Journal of Public Opinion Research* 21(1): 111–121. DOI: https://doi.org/10.1093/ijpor/edn054.

ISSP Research Group (2016): International Social Survey Programme: Citizenship II – ISSP 2014. GESIS Data Archive, Cologne. ZA6670 Data file Version 2.0.0. DOI: https://doi.org/10.4232/1.12590.

Jenkins C.R, and D.A. Dillman. 1995. "Towards a Theory of Self-Administered Questionnaire Design." In *Survey Measurement and Process Quality,* edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin: 165–196. New York Wiley.

Just, M.A., and P.A. Carpenter. 1980. "A Theory of Reading: From Eye Fixations to Comprehension." *Psychological Review* 87: 329–354. DOI: https://doi.org/10.1037/0033-295X.87.4.329.

Kaczmirek, L. 2005. "A Framework for the Collection of Universal Client Side Paradata (UCSP)." Available at: http://kaczmirek.de/ucsp/ucsp.html (accessed January 2021).

Kamoen, N., B. Holleman, P. Mak, T. Sanders, and H. van den Bergh. 2017. "Why are Negative Questions Difficult to Answer? On the Processing of Linguistic Contrasts in Surveys." *Public Opinion Quarterly* 81(3): 613–635. DOI: https://doi.org/10.1093/poq/nfx010

Koffka, K. 1935. *Principles of Gestalt psychology*. New York: Harcourt.

Krosnick, J.A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5(3): 213–236. DOI: https://doi.org/10.1002/acp. 2350050305.

Krosnick, J.A., and D.F. Alwin. 1988. "A Test of the Form-Resistant Correlation Hypothesis. Ratings, Rankings, and the Measurement of Values." *Public Opinion Quarterly* 52(4): 526–538. DOI: https://doi.org/10.1086/269128.

Krosnick, J.A., and S. Presser. 2010. "Question and Questionnaire Design." In *Handbook of Survey Research*, edited by P.V. Marsden, and J.D. Wright: 263–314. Emerald Group Publishing.

Landis, J.R., and G.G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33(1): 159–174. DOI: https://doi.org/10.2307/2529310.

Liu, M., and A. Cernat. 2018. "Item-by-Item versus Matrix Questions: A Web Survey Experiment." *Social Science Computer Review* 36(6): 690–706. DOI: https://doi.org10.1177/0894439316674459.

Mavletova, A., and M.P. Couper. 2015. "A Meta-Analysis of Breakoff Rates in Mobile Web Surveys." In *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies*, edited by D. Toninelli, R. Pinter, and P. de Pedraza: 81–98. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bar.f.

Mavletova, A., M.P Couper, and D. Lebedev. 2018. "Grid and item-by-item formats in PC and mobile web surveys." *Social Science Computer Review* 36(6): 647–668. DOI: https://doi.org/10.1177%2F0894439317735307.

Mayerl, J. 2013. "Response Latency Measurement in Surveys. Detecting Strong Attitudes and Response Effects". *Survey Methods: Insights from the Field*. Retrieved December 17, 2020, from https://surveyinsights.org/?p = 1063. DOI: https://doi.org/10.13094/SMIF-2013-00005.

McCarty, J.A., and L.J. Shrum. 2000. "The Measurement of Personal Values in Survey Research: A Test of Alternative Rating Procedures." *Public Opinion Quarterly* 64(3): 271–298. DOI: https://doi.org/10.1086/317989.

Neuert, C.E., and T. Lenzner. 2017. “Incorporating Eye Tracking into Cognitive Interviewing to Pretest Survey Questions.” *International Journal of Social Research Methodology* 19(5): 501–519. DOI: https://doi.org/10.1080/13645579.2015.1049448.

Peytchev, A. 2005. “How Questionnaire Layout Induces Measurement Error.” Paper presented at the 60th annual meeting of the American Association for Public Opinion Research, May, 2005. Miami Beach, FL,USA. Available at: http://www.websm.org/db/12/3636/Bibliography/Causes%20of%20Context%20Effects:%20How%20Questionnaire%20Layout%20Induces%20Measurement%20Error/.

Rammstedt, B., and O.P. John. 2007. “Measuring Personality in One Minute or Less: A 10-Item Short Version of the Big Five Inventory in English and German.” *Journal of Research in Personality* 41: 203–212. DOI: https://doi.org/10.1016/j.jrp. 2006.02.001.

Rayner, K. 1998. “Eye Movements in Reading and Information Processing: 20 Years of Research.” Psychological Bulletin 124: 372–422. DOI: https://doi.org/10.1037/0033-2909.124.3.372.

Romano Bergstrom, J., and A. Schall. 2014. *Eye Tracking in User Experience Design*. San Francisco, CA: Morgan Kaufmann.

Roßmann, J., T. Gummer, and H. Silber. 2018. “Mitigating Satisficing in Cognitively Demanding Grid Questions: Evidence from Two Web-Based Experiments.” *Journal of Survey Statistics and Methodology* 6(3): 376–400. DOI: https://doi.org/10.1093/jssam/smx020.

Roßmann, J., and H. Silber. 2020. “Satisficing and Measurement Error.” In *SAGE Research Methods Foundations*, edited by P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug, and R.A. Williams. London: SAGE Publications. DOI: https://dx.doi.org/10.4135/9781526421036912794.

Silber, H., J. Roßmann, and T. Gummer. 2018. “When Near Means Related: Evidence from Three Web Survey Experiments on Inter-Item Correlations in Grid Questions.” *International Journal of Social Research Methodology* 21(3): 275–288. DOI: https://doi.org/10.1080/13645579.2017.1381478.

Staub, A., and K. Rayner. 2007. “Eye movements and on-line comprehension processes.” In *The Oxford Handbook of Psycholinguistics*, edited by G. Gaskell: 327–342. Oxford, UK: Oxford University Press.

Toepoel, V., M. Das, and A. van Soest. 2009. “Design of Web Questionnaires: The Effects of the Number of Items per Screen.” *Field Methods* 21(2): 200–213. DOI: https://doi.org/10.1177/1525822X08330261.

Tourangeau R., M.P. Couper, and F. Conrad. 2004. “Spacing, Position, and Order. Interpretive Heuristics for Visual Features of Survey Questions.” *Public Opinion Quarterly* 68(3): 368–393. DOI: https://doi.org/10.1093/poq/nfh035.

Tourangeau, R., and K. Rasinski. 1988. “Cognitive Processes Underlying Context Effects in Attitude Measurement.” *Psychological Bulletin* 103(3): 299–314. DOI: 10.1037/0033-2909.103.3.299.

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

Wertheimer, M. 1923. *Laws of organization in perceptual forms: A source book of Gestalt psychology*. London: Routledge.

W3C: World Wide Web consortium. 2018. “Web Content Accessibility Guidelines (WCAG) 2.1.” Available at: https://www.w3.org/TR/WCAG21/ (accessed June 2021).

Zhang, C., and F. Conrad. 2014. “Speeding in Web Surveys: The Tendency to Answer very Fast and its Association with Straightlining.” *Survey Research Methods* 8(2): 127–135. DOI: https://doi.org/10.18148/srm/2014.v8i2.5453.

Received January 2021
Revised June 2021
Accepted August 2022