Comparing the Performance of Agree/Disagree and Item-Specific Questions across PCs and Smartphones

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

Melanie Revilla
Universitat Pompeu Fabra (Spain)

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

Abstract

The use of agree/disagree (A/D) questions is a common technique to measure attitudes. For instance, this question format is employed frequently in the Eurobarometer and International Social Survey Programme (ISSP). Theoretical considerations, however, suggest that A/D questions require a complex processing. Therefore, many survey researchers have recommended the use of item-specific (IS) questions, since they seem to be less burdensome. Parallel to this methodological discussion is the discussion around the use of mobile devices for responding to surveys. However, until now, evidence has been lacking as to whether the use of mobile devices for survey response affects the performance of established question formats. In this study, implemented in the Netquest panel in Spain (N = 1,476), we investigated the cognitive effort and response quality associated with A/D and IS questions across PCs and smartphones. For this purpose, we applied a split-ballot design defined by device type and question format. Our analyses revealed longer response times for IS questions than AD questions, irrespective of the device type and scale length. Also, the IS questions produced better response quality than their A/D counterparts. All in all, the findings indicate a more conscientious response to IS questions compared to A/D questions.

Keywords: agree/disagree questions, asking manner, device type, item-specific questions, paradata, SurveyFocus, web surveys

Introduction and Background

Over the last couple of years, a growing number of scientific contributions and empirical studies have compared the performance of agree/disagree (A/D) and item-specific (IS) questions (Fowler, 1995; Höhne & Krebs, 2017; Höhne & Lenzner, 2017; Höhne, Schlosser, & Krebs, 2017; Kunz, 2017; Kuru & Pasek, 2016; Lelkes & Weiss, 2015; Liu, Lee, & Conrad, 2015; Saris, Revilla, Krosnick, & Shaeffer, 2010). A/D questions usually start with a request for an

This document is a preprint and thus it may differ from the final version: Höhne, Jan K., Revilla, Melanie, & Lenzner, Timo (2018). Comparing the performance of agree/disagree and item-specific questions across PCs and smartphones. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 14, 109–118. DOI: 10.1027/1614-2241/a000151

answer (e.g., "Do you agree or disagree with the following statement?"), accompanied by a statement (e.g., "I am interested in politics.") including a response scale in which the categories are based on an agreement/disagreement continuum (e.g., "agree strongly" to "disagree strongly"). In contrast, IS questions usually consist of an interrogative request for an answer (e.g., "How interested are you in politics?") including a response scale that matches the underlying content dimension (e.g., "very interested" to "not at all interested").

Regarding A/D questions, it is expected that their statements and response scales can be used to measure a latent variable, such as political interest, provided that the statement represents a characteristic of the latent variable. From a theoretical perspective, however, the measurement of attitudes by means of the A/D question format exhibits an ordinal level, even if the latent concept is continuous. Furthermore, it is supposed that ordering respondents on an agreement/disagreement continuum is equivalent to ordering them on an actual content dimension (Krosnick & Presser, 2010). However, these assumptions have been challenged in the survey literature and several problems associated with the A/D question format, such as acquiescence, have been discussed (Converse & Presser, 1986; Höhne & Lenzner, 2015, 2017; Krosnick, 1991; Revilla, Saris, & Krosnick, 2014; Saris et al., 2010). In the following, we will provide a brief theoretical overview on A/D as well as IS questions and their implications for attitude measurement (for a more detailed discussion, see Höhne et al., 2017 or Saris et al., 2010).

Several researchers have argued that respondents must accomplish specific mental tasks when answering AD questions, which implies that a complex cognitive process occurs between reading and responding (Carpenter & Just, 1975; Fowler, 1995; Höhne & Lenzner, 2017; Höhne et al., 2017; Revilla et al., 2014; Saris et al., 2010). First, respondents must comprehend the literal meaning of the statements and response categories. Second, they must identify the pragmatic meaning of the survey question. Third, they must perform a mental positioning task to place themselves on the scale of interest. Finally, they must translate their mental decision onto the underlying response categories. Since IS questions directly express the underlying content dimension in the stem and response categories, the pragmatic meaning (task 2) is clearer, and the translation of the mental judgment (task 4) is simpler.

According to Converse and Presser (1986), A/D questions are prone to response bias – a systematic distortion of the cognitive response process (Groves et al., 2004). Höhne et al. (2017), for instance, state that A/D questions promote boredom and weariness because the "asking manner" does not change – they always employ identical response categories (e.g., running from "agree strongly" to "disagree strongly"). Respondents must repeat the same response task without any variation. Also, responses to A/D statements are not concerned with an actual content dimension (e.g., "interest in politics"), since these statements use an indirect asking manner. Hence, A/D questions seemingly dismay respondents and restrain them from paying much attention to and/or using much effort for their responses. IS questions, in comparison, usually change the asking manner continuously and are based on a direct interrogation, which might promote a more thoughtful and well-considered response process.

From a cognitive psychological point of view, the A/D question format demands a complex and elaborate response process. However, respondents do not have to read the response categories repeatedly, since the asking manner does not change, which promotes a superficial rather than a conscientious response process. Although easier to process, the IS question format demands a continuous reconsideration of the underlying content dimension due to a direct and changing asking manner, which encourages respondents to perform a more active and intensive response. Thus, it can be expected that responding to IS questions is more elaborate than responding to A/D questions.

Overall, the evidence concerning A/D and IS questions in the survey literature is somewhat mixed, and most studies do not take mobile devices into account. Saris et al. (2010), for example, investigated the measurement quality of A/D and IS questions using data from the European Social Survey (ESS), which is based on face-to-face interviews. The authors found a higher reliability and validity for the IS question format. Lelkes and Weiss (2015) and Liu et al. (2015) compared A/D and IS questions in terms of their reliability and validity, as well as extreme response style, using data from the American National Election Study (ANES), which is based on a face-to-face or web mode (PC only), but found no differences between them. Höhne et al. (2017) investigated both question formats with paradata using a student sample, but due to technical difficulties, they excluded mobile devices. These authors found that IS questions are cognitively more demanding but produce better response quality (i.e., less speeding) than A/D questions. To our knowledge, only one study by Kunz (2017) – based on a student sample with no random assignment of devices – has compared A/D and IS questions across different devices. This author also found that IS questions were associated with higher cognitive effort and better response quality (i.e., less speeding and less primacy effects).

In recent years, the use of mobile devices, particularly smartphones[1], to complete surveys has increased steadily. For instance, Revilla, Toninelli, Ochoa, and Loewe (2016) have shown that in a variety of countries, the share of respondents who participate in web surveys using smartphones has been increasing continuously. One of the main reasons for this development over the last years is a "skyrocketing" proportion of smartphone owners accompanied by an increase in (high-speed) mobile Internet access. Theoretically, smartphones enable respondents to participate in surveys independent of locality (Mavletova, 2013), even though previous research has found that in practice most smartphone respondents complete surveys at home (Mavletova & Couper, 2013; Toninelli & Revilla, 2016a). However, smartphone respondents are surrounded more often by third parties (Mavletova & Couper, 2013; Toninelli & Revilla, 2016a), and thus the different contexts of survey participation – compared to the web or other survey modes – may affect the quality of their responses. Toninelli and Revilla (2016b), for instance, have shown that smartphone respondents report more distractions due to their environment and/or multitasking behavior, such as watching TV and communicating (i.e., talking and chatting) with other people during survey completion.

In addition to contextual differences, the use of smartphones may affect survey processing negatively due to smaller screen sizes, lower processing power, and intricate input capabilities, such as the lack of a keyboard (Mavletova, 2013; Toninelli & Revilla, 2017). Hence, it can be assumed that participating in surveys on smartphones can affect the process of responding to survey questions differently than participating with other more established devices (Fuchs, 2008). Thus, it cannot be expected that survey design in general and question design in

[1] In this article, we only focus on PCs and smartphones for two reasons. First, respondents using tablets to respond to surveys behave quite similar to respondents using PCs (Couper & Peterson, 2017). Second, Revilla et al. (2016) have shown that in the Netquest panel in Spain – the panel and country in which the data were collected – smartphones were used much more frequently than tablets for responding to surveys. Furthermore, smartphones would be preferred prospectively by respondents if survey completion is adapted entirely to mobile devices.

particular for PC devices (e.g., a desktop or laptop computer) and smartphones have the same impact on response behavior.

To gain a better understanding of the performance of A/D and IS question formats when answering with PCs and smartphones, we investigated the associated cognitive effort and response quality using a split-ballot design. The data were collected in 2016 using the Netquest (Spain) access panel. Whereas cognitive effort is measured by client-side response times in milliseconds, response quality is assessed by speeding and primacy effects (Höhne et al., 2017; Kunz, 2017; Malhotra, 2008). In contrast to previous research, we focused on a different question topic, conducted our research in a different country, used questions with different scale lengths, did not use a student sample, and randomly assigned respondents to the device type. Hence, our research constitutes a substantial extension of the current state of research.

Research Hypotheses

The use of the Internet to collect survey data enables the collection of client-side paradata, such as response times, to describe and investigate the response behavior of respondents during survey completion. By analyzing respondents' answers (i.e., looking for response patterns), it also is possible to draw conclusions about response quality. In the following, we explain and justify our research hypotheses on cognitive effort and response quality before providing a summary of them in Table 1.

Cognitive Effort

Researchers have assumed that a close connection exists between response time and cognitive processing, which implies that the shorter/longer the time to answer a question, the lower/higher the cognitive effort must be (Conrad, Couper, Tourangeau, & Zhang, 2017; Höhne et al., 2017; Lenzner, Kaczmirek, & Lenzner, 2010; Yan & Olson, 2013). Thus, the length of time a respondent takes to answer a survey question – e.g., presented in the A/D or IS format – informs about the level of elaboration (Mayerl & Urban, 2008, pp. 22–24). In line with the asking manner concept (Höhne et al., 2017), we expect that IS questions generally produce longer response times than their A/D counterparts. In addition, we expect to observe higher differences between A/D and IS questions for PCs than for smartphones. Indeed, since smartphones enable respondents to complete a survey whenever and wherever they want, it can be assumed that they are more distracted during survey completion, which may negatively affect their response (Mavletova, 2013; Mavletova & Couper, 2013; Toninelli & Revilla, 2016a). Furthermore, we expect that the differences between the two question formats are more pronounced in the 7-point compared to the 5-point scales because the former requires a more complex mapping process due to the higher number of scale points (Krosnick & Presser, 2010).

Response Quality

In the present study, we follow Converse and Presser (1986) and Krosnick (1991) as we evaluate response quality in terms of satisficing response behavior, rather than in terms of reliability and validity. For this purpose, we use speeding and primacy effects as indicators of primarily poor response quality (Höhne et al., 2017; Kunz, 2017; Malhotra, 2008).[2] We define speeding as responding too fast, which implies that respondents are not able to process the questions properly – i.e., perform the required response tasks (e.g., mapping process) without using any cognitive shortcuts (Conrad et al., 2017). As shown by previous research, speeding is associated frequently with further types of satisficing response behavior (Callegaro, Yang, Bhola, Dillman, & Chin, 2009; Conrad et al., 2017; Höhne et al., 2017; Kunz, 2017; Malhotra, 2008). We define primacy effects as selecting response categories at the beginning of a scale without considering and/or processing the subsequent categories (Krosnick, 1991).

Since A/D questions are tiring and tend to attenuate respondents' motivation due to an indirect and unchanging asking manner, we postulate a better response quality (i.e., less speeding behavior and less primacy effects) for the IS questions compared to their A/D counterparts. In line with the hypotheses on response times, we expect to find higher differences for PCs than smartphones – due to the device-related issues mentioned previously – and for 7-point than 5-point response scales (see Revilla et al., 2014).

Table 1. Research hypotheses on response times and response quality

Response Times	IS questions produce longer response times than A/D questions. (H1)
	The differences between A/D and IS questions are higher in PCs than in smartphones. (H1a)
	The differences between A/D and IS questions are higher in 7-point than in 5-point response scales. (H1b)
Response Quality	IS questions produce better response quality than A/D questions. (H2)
	The differences between A/D and IS questions are higher in PCs than in smartphones. (H2a)
	The differences between A/D and IS questions are higher in 7-point than in 5-point response scales. (H2b)

Method

Data Collection

The data collection by using the Netquest access panel was conducted in Spain from 15th September to 3rd October 2016. Netquest (www.netquest.com) is an online fieldwork company operating in the USA, the main countries of Europe, and Latin America. The panel in Spain exists since 2005 and counts more than 203,500 active panelists (status in May 2018), as defined in the ISO26362 norm. Netquest has arrangements with a variety of websites and implements satisfaction surveys with the users of these websites. At the end of the surveys, if respondents match Netquest's targets, they are invited to join the panel. To avoid duplicate registrations, invitations can be used only once. Incentives are provided to respondents to register for the panel and for each survey completed (proportional to the estimated length of the survey; see Revilla, 2017, for more details).

In addition to respondents' answers to the survey questions, Netquest also collects several types of client-side paradata. For example, response times in milliseconds (i.e., the time elapsing between the question presentation on the screen and the time the page was submitted

[2] We also looked at item non-response and non-differentiation (see Revilla & Couper, 2018a) but the occurrence of these response behaviors was quite low (see also the chapter "Analytical Strategy"). In addition, we compared the inter-item correlations between the A/D and IS questions. However, we did not find substantial differences.

by clicking "Next") and SurveyFocus (Höhne & Schlosser, 2018; Höhne et al., 2017), which gathers the activity of the web-survey page (i.e., detecting whether the window that hosts the web survey is also the active or processed one, see Callegaro, 2013).

Population of Interest and Sample

For the purpose of the present study, the population of interest was not the general population per se but the group of panelists from the Netquest panel who were 18 years or older, had Internet access through both PC and smartphone, and had not been invited to install a tracker on their devices.[3] We used cross quotas for gender and age to get a sample representative of this population.

The decision to focus on this population was based on the topic of the survey, several practical reasons (e.g., Netquest does not provide Internet access to their panelists), and the fact that Netquest wants to make practical decisions based on the empirical results of this survey. To date, the literature does not provide evidence that differences between A/D and IS questions are affected by the population of interest. Hence, we expect that our findings also are applicable to more general populations.

In total, 5,907 panelists were invited to take part in the survey out of which 3,051 (52%) started it. A total of 1,623 respondents (i.e., 53% of those who started and 28% of those who were invited) answered the first "real" survey question, following the quota and filter questions. 1,428 (i.e., 47% of those who started and 24% of those who were invited) were screened out due to one of the following reasons: they did not connect through the required device type (and never came back through the required one), they were excluded based on the filters or quotas, or they dropped out during the first four questions on age, gender, education, and Internet access. A further 127 respondents were excluded because they switched to another device during the survey, which did not match the required type. An additional 5 respondents were excluded because they did not pass several basic quality checks. Finally, 15 participants dropped out after the first four questions. In total, 1,476 panelists successfully completed the entire survey.

These panelists were between 18 and 94 years old with a mean age of 36 years (standard deviation of 12 years). Of these participants, 61% were female. 1% had not graduated or had graduated only from a primary school, 26% had graduated from a secondary school, and 73% had graduated from a college preparatory school or university.

Study Design

To test our research hypotheses on cognitive effort and response quality, we applied a split-ballot design with four experimental groups defined by device type (PC vs. smartphone) and question format (A/D vs. IS), which provided a 2-by-2 study design, as displayed in Table 2.

[3] Netquest's panelists can install a tracker on their devices to gather the URLs of the webpage that they visit. In return, they receive additional incentives.

Table 2. Experimental design defined by device type and question format

Experimental Group	Device Type	Question Format	Number of Completions
1	PC	A/D	321
2	PC	IS	321
3	Smartphone	A/D	418
4	Smartphone	IS	416

In a first step, we randomly assigned the participants to a device (PC or smartphone) before the start of the survey using profiling information, which was provided by the online fieldwork company. The email invitation informed the participants about the device that they had to use to complete the survey. We only invited respondents who owned both a PC and a smartphone according to the profiling information, but we also checked that they have both devices by means of a filter question at the beginning of the survey. If they did not have access through both devices, they were redirected to a profiling module and screened out of the study. If they tried to participate through a different device than the one assigned, this action was automatically detected, and they were blocked by a message asking them to use the required device.

In a second step, we randomly assigned participants to the A/D or IS question format. The experiment was part of a larger study with different unrelated experiments (see Revilla & Couper, 2018a; Revilla & Couper, 2018b; Revilla, Couper, & Ochoa, 2018). All experiments were independently randomized to avoid carryover effects.

We used an optimized survey design because the layout was adapted to the screen and window size, which improved readability and avoided horizontal scrolling (Couper & Peterson, 2017; Mavletova & Couper, 2015).

Survey Questions

The survey contained about 70 questions. Respondents were able to skip questions, which is not the general procedure that Netquest uses. Therefore, we prepared an extra introductory page that explained that respondents can skip questions, but also that their answers were scientifically important.

We employed 12 questions in the second half of the questionnaire that we had adapted from personality test surveys. Based on these 12 questions, we developed A/D and IS counterparts preserving the question content (see Figure 1). The participants answered six A/D or IS questions with a 5-point response scale, and then another six A/D or IS questions with a 7-point response scale. The questions were designed in Spanish, which was the mother tongue of approximately 93% of the respondents. The full web questionnaire in Spanish can be accessed at *ww2.netquest.com/respondent/glinn/mobile2016*.

We used vertically arranged and end-point labeled scales with a decremental response category order (i.e., running from positive to negative) and numerical labels. In addition, we presented each question on a separate page and respondents received an initial introduction, adapted for A/D questions to explicitly explain that they had to decide to what extent they agree or disagree with the statements. To avoid question-order effects, we randomized the question order of the 5-point and 7-point response scale questions, respectively.

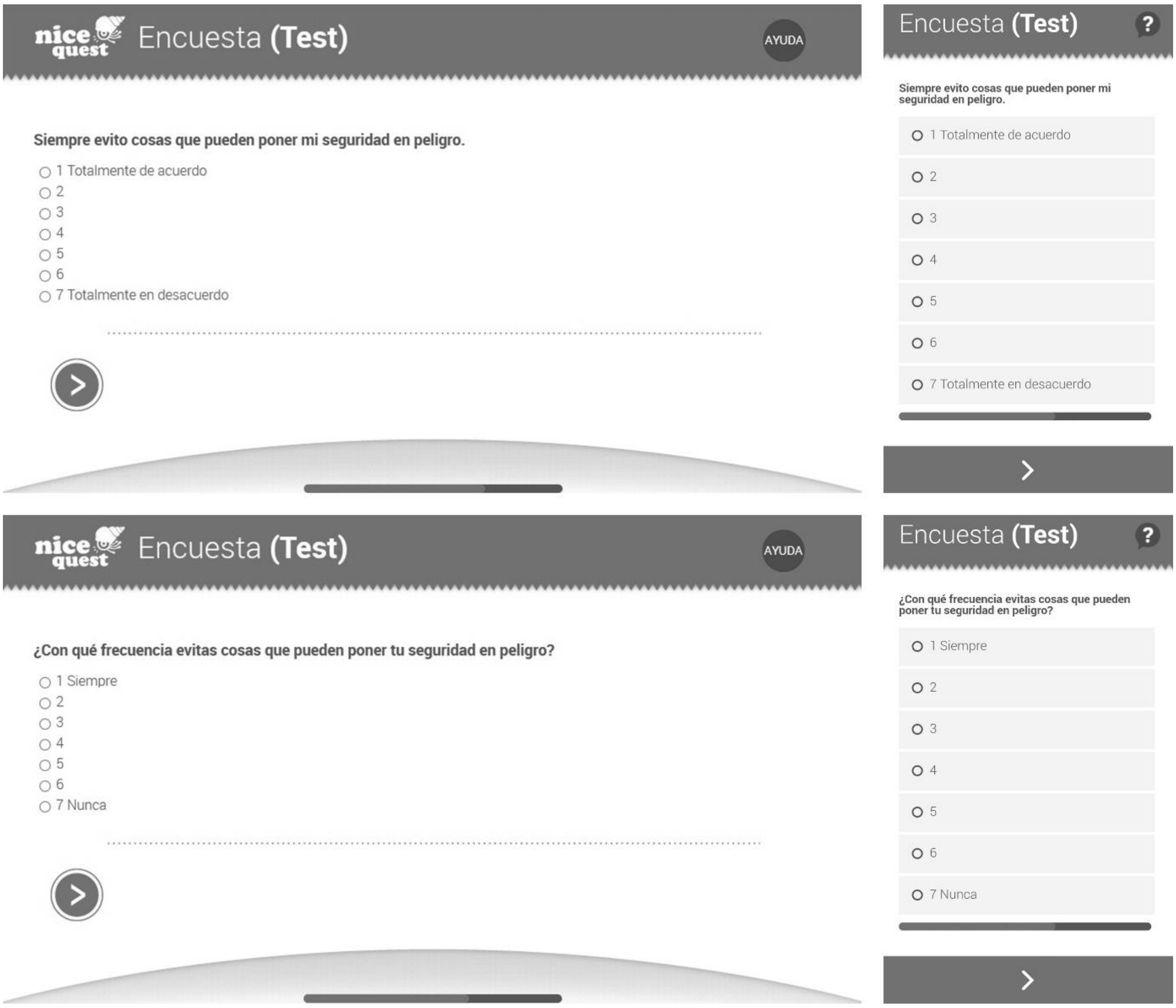

Figure 1. Screenshots of one A/D question and one IS question with a 7-point response scale for PCs and smartphones

Note. While the A/D question is presented in the upper part (PC on the left and smartphone on the right), the IS question is presented in the lower part (PC on the left and smartphone on the right). English translations of all questions including response categories are available from the first author on request.

Analytical Strategy

Cognitive Effort

To deal with response time outliers, we employed a two-step procedure using the paradata "SurveyFocus (SF)"[4] (Höhne & Schlosser, 2018; Höhne et al., 2017). First, we excluded as outliers all respondents who left the web-survey page (e.g., switched between browser tabs) for a certain time.[5] For the remaining respondents, we applied a distribution-sensitive outlier definition (Hoaglin, Mosteller, & Tukey, 2000) that excluded all response times below or above the median plus/minus the upper and lower quartile range multiplied by 3. We also tested the upper and lower one percentile (Lenzner et al., 2010) as thresholds, but the results did not change. Moreover, we conducted all the analyses with and without the log transformation of the response-time data. Again, the results were unchanged, and thus we report the untreated solution.

[4] Netquest adapted the SF tool to meet the purposes of our study (e.g., the application to mobile devices, which was not available that time).

[5] At the page-level, discontinuously processing respondents are about 2%.

We tried to keep the syllable numbers as similar as possible for the two question formats because the length of a question could influence a respondents' processing time (Baddeley, 1992). Nevertheless, the A/D and IS questions slightly differed in their syllable numbers, which was unavoidable without formulating artificial-sounding questions. Following Lenzner et al. (2010), we corrected for length differences between the A/D and IS questions by dividing response times by the syllable number. Thus, we report response times per syllable. As suggested by Ferreira and Clifton (1986), we also divided response times by the number of characters, but the main conclusions did not change. However, we did not adjust for the baseline reading speed (Couper & Kreuter, 2013).

Although not reported in the result section, we investigated the number of answer changes, which can be an indicator of cognitive effort (Stern, 2008). However, we did not find any significant differences between device types and question formats, irrespective of the scale length.

Response Quality

We used speeding and primacy effects as indicators of low response quality. Regarding speeding, we used the 15^{th} percentile of all response times – after outlier definition[6] – and compared the proportion of speeders between the experimental groups. Hence, we considered speeding as a relative phenomenon (Conrad et al., 2017; Höhne et al., 2017; Malhotra, 2008). As a robustness check, we tested other thresholds, such as the 5^{th}, 10^{th}, 20^{th}, and 25^{th} percentile. Again, the main conclusions did not change. Regarding primacy effects, we used the number of responses to the first response category.[7] The reason we used this strategy was that the responses on the first half of the scale – i.e., the first two categories (5-point scales) and the first three categories (7-point scales), respectively – were comparatively high. Thus, we decided to compare the average number of responses to the first category for the 5-point and 7-point scale questions. In addition to these two response quality indicators, we checked for item non-response and non-differentiation. The occurrence of these response biases was negligibly small (about 3% item non-response, and about 1% non-differentiation) and did not vary substantially across the experimental groups. Thus, we do not report them more precisely in the results section.

General Consideration

The analysis of paradata can be conducted on different aggregation levels. Since substantial differences do not exist at the question-level and to reduce the number of statistical tests and efficiently summarize the results, we conducted the statistical analyses for the six questions with 5-point and 7-point response scales, respectively. We used Stata version 13 to conduct the data preparation and analyses. For response times, we calculated one-way analyses of variance (ANOVAs) using the Bonferroni α-inflation correction procedure for equal variances to deal with the problem of multiple comparisons and Cohen's d (Cohen, 1969) as a measure of effect size. In line with our analytical strategy for response quality (see above), we calculated chi-

[6] We also tested for speeding without a previous outlier exclusion, but the results were unchanged, irrespective of the percentile used.

[7] It must be mentioned that the primacy effects in the A/D questions also could be an indicator of acquiescence, but due to the study design, these two response biases are not readily distinguishable. For the sake of convenience, we simply speak of primacy effects.

square tests for speeding and unpaired t-tests for primacy effects. To test the second order hypotheses on response times and response quality (H1a and b as well as H2a and b, respectively), we conducted significance tests across the conditions of interest for means and proportions using the Z-statistic.

Results

Cognitive Effort

For the following statistical analyses, response times – as indicator of cognitive effort – are the primary dependent variable.

Table 3 provides the results of the one-way analyses of variance (ANOVAs). In line with our previous expectation (see H1), the results reveal significant differences in average response times for the questions with 5-point and 7-point response scales. More precisely, the IS questions produce consistently higher response times than their A/D counterparts. This finding is also supported by Cohen's d that indicated small to medium effect sizes. However, contrary to our expectation (see H1a), the differences between the question formats are not more pronounced for PCs than smartphones for the 5-point scale questions. Instead, they are slightly higher for smartphones (Z = .09), which also is indicated by Cohen's d. In contrast, the response time differences for the 7-point scale questions are slightly higher for PCs than smartphones, but not statistically significant (Z = -.54). Accordingly, Cohen's d produces higher coefficients for PCs than smartphones.

Table 3. Mean differences of response times per syllable for the six aggregated A/D and IS questions with 5-point and 7-point response scales for PCs and smartphones

Experimental Comparison	5-Point Scales				7-Point Scales			
	Mean Difference	Effect Size	(df_1 = 3)	F Value	Mean Difference	Effect Size	(df_1 = 3)	F Value
A/D (PC) – IS (PC)	-2.01*	.22	df_2 = 1469	8.09***	-1.87*	.23	df_2 = 1460	9.47***
A/D (Smartphone) – IS (Smartphone)	-2.10**	.24			-1.43*	.20		
A/D (PC) – A/D (Smartphone)	-1.03	.12			-1.67*	.22		
IS (PC) – IS (Smartphone)	-1.11	.12			-1.22	.15		

Note. *p < .05; **p < .01; ***p < .001. Mean differences: first group minus second group. Cohen's d states the effect sizes.

A comparison of the findings for the questions with 5-point and 7-point response scales (see Table 3) did not result in higher response time differences for the 7-point scale questions (see H1b). In contrast, the 5-point scale questions cause continuously larger differences for PCs (Z = -.14) and smartphones (Z = -.83). The Cohen's d coefficients are quite similar. Altogether, supporting evidence does not exist for a more difficult mapping process due to the response scale length, irrespective of the device type.

An examination of the average response times of the two devices regarding the A/D and IS question formats has found higher response times for respondents using a smartphone compared to those using a PC. However, this finding is only significant for the 7-point A/D questions. Nevertheless, this finding corresponds to previous web studies comparing device types (Couper & Peterson, 2017).

Response Quality

To evaluate the response quality of A/D and IS questions, we used speeding and primacy effects. Table 4 provides the statistical results for the two response quality indicators for the six aggregated questions with 5-point scales and the six aggregated questions with 7-point scales.

Table 4. Response quality of the six aggregated A/D and IS questions with 5-point and 7-point response scales for PCs and smartphones

		Speeding		Primacy Effects	
Device Type	Question Format	5-Point Scales	7-Point Scales	5-Point Scales	7-Point Scales
PC	A/D	10.75	9.97	.82	1.45
	IS	7.32	7.48	.53	1.39
		$\chi^2(1) = 5.09$*	$\chi^2(1) = 2.77$	$t(640) = 4.87$**	$t(640) = .70$
Smartphone	A/D	7.55	6.95	.84	1.51
	IS	4.92	5.88	.61	1.50
		$\chi^2(1) = 5.20$*	$\chi^2(1) = .82$	$t(832) = 4.08$**	$t(832) = .03$

Note. *$p < .05$; **$p < .01$

Speeding

As previously hypothesized (see H2), the proportion of speeders was more distinct within the A/D compared to the IS question format. This finding applies to the questions with 5-point and 7-point response scales as well as to PCs and smartphones. However, this finding is only statistically significant for the 5-point scale questions in PCs and smartphones. For the 7-point scale questions in PCs and smartphones, the differences only tend toward the expected direction. In correspondence with our expectation (see H2a), the differences are slightly higher in PCs than smartphones for the 5-point scale questions (Z = .11) and 7-point scale questions (Z = .20). However, contrary to our expectation (see H2b), the differences are somewhat more pronounced for questions with 5-point compared to 7-point scales for PCs (Z = .13) and smartphones (Z = .23).

Primacy Effects

The second response quality indicator reported in this study is primacy effects (i.e., attraction to the first response category of the response scale). Table 4 shows that the average number of responses to the first category is significantly higher for the A/D compared to the IS question format (see H2). However, this finding only applies to the 5-point scale questions in both device types. For the questions with 7-point response scales, the differences are negligibly small. As expected (see H2a), the differences are higher in PCs compared to smartphones for the 5-point scale questions (Z = .12) and the 7-point scale questions (Z = .50), although these differences are not statistically significant. In contrast to our expectation (see H2b), the differences are significantly larger for the questions with 5-point compared to 7-point response scales for PCs (Z = 1.96) and smartphones (Z = 2.20). Hence, the scale length seems to matter to the response quality of the A/D and IS questions in terms of primacy effects.

Discussion and Conclusion

The purpose of this study was to investigate the performance of A/D and IS questions across different device types regarding the level of cognitive effort – assessed by response times – and response quality – assessed by speeding and primacy effects. The empirical findings suggest that IS questions consistently cause longer response times compared to their A/D counterparts, since they trigger a more well-considered response, which also is in line with the concept of asking manner.

We could not find supporting evidence for the expected differences in response times of the questions with 5-point and 7-point scales, which indicates that the mapping effort seems to be similar. However, the study design partially complicates the interpretation of the results, since it is possible that the order – 5-point and then 7-point scale questions – had an impact on the results. More precisely, respondents might be already familiar with the end-point labeled response scale format, so the increase of scale points did not affect their mental translation difficulty (for a more detailed discussion, see Krosnick & Presser, 2010). Thus, it would be desirable if future research applies a design that guarantees that the question order has no impact.

We investigated the response quality of A/D and IS questions in terms of speeding and primacy effects. In line with our expectation, we found that speeding occurs more often for the A/D compared to the IS question format, which can be assessed as further evidence that A/D questions cause a more superficial processing. Similar to Höhne and Krebs (2017), we found that the respondents who answered A/D questions were more likely to tend toward the beginning of the scale (i.e., to the first response category) compared to respondents who answered IS questions. However, this finding only applies to the 5-point response scale questions, which indicates that the scale length matters.

Similar to the differences in response times, we could not find the expected differences in response quality, except for the primacy effects between 5-point and 7-point response scale questions. This lack of supporting evidence for our expectation applies irrespective of the device type.

Our study has three limitations that could be addressed by future research. First, we did not randomly assign respondents to the 5-point and 7-point response scale questions, which may have confounded our results. However, we did randomly assign respondents to the device type and question format. Second, our target population was Netquest's panelists, which impedes the generalizability of our findings to other target populations (e.g., with a lower level of education and/or less survey experience). Third, although previous research has shown that IS questions produce better data quality than A/D questions in cross-national settings (Saris et al., 2010), it remains open whether this finding also applies across different devices, since this study was conducted only in Spain.

Finally, our findings contribute to quantitative social science research in a theoretical and a practical way. Theoretically speaking, we were able to show that respondents do not seem to expend more cognitive effort when responding to the A/D compared to the IS question format. Of course, from a psychological perspective, A/D questions force respondents to conduct an elaborate and intricate response process (see Carpenter & Just, 1975), although this does not necessarily mean that respondents conscientiously carry out all the required response steps. Optimal responses to A/D questions seem to emerge rather rarely due to an indirect and constant asking manner, which provokes a superficial response process. Thus, we argue that the asking manner is a significant characteristic of question formats, and so we encourage researchers to take it into account when evaluating them. Practically speaking, when designing survey instruments, our findings suggest that IS questions should be preferred over A/D questions. Most importantly, this finding seems to apply to different device types, such as PCs and smartphones. In correspondence with former studies (Saris et al., 2010), our empirical findings, and the asking manner concept, we recommend that researchers employ IS rather than A/D questions, since the former evoke a more thoughtful response process.

References

Baddeley, A. (1992). Working Memory. *Science, 255*, 556–559.

Callegaro, M. (2013). Paradata in Web Surveys. In F. Kreuter (Ed.), *Improving Surveys with Paradata. Analytic Uses of Process Information* (pp. 261–280). Hoboken, NJ: John Wiley & Sons.

Callegaro, M., Yang, Y., Bhola, D. S., Dillman, D. A., & Chin, T. Y. (2009). *Response latency as an indicator of optimizing in online questionnaires*. *Bulletin de Méthodologie Sociologique*, 103, 5–25.

Carpenter, P. A., & Just, M. A. (1975). Sentence Comprehension: A Psycholinguistic Processing Model of Verification. *Psychological Review, 82(1)*, 45–73.

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

Conrad, F. G., Couper, M. P., Tourangeau, R., & Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods, 11(1)*, 45–61.

Converse, J. M., & Presser, S. (1986). *Survey Questions: Handcrafting the Standardized Questionnaire*. Beverly Hills, CA: Sage.

Couper, M. P., & Kreuter F. (2013). Using Paradata to Explore Item Level Response Times in Surveys. *Journal of the Royal Statistical Society, 176*, 271–286.

Couper, M. P., & Peterson, G.J. (2017). Why do web surveys take longer on smartphones? *Social Science Computer Review, 35*, 357–377.

Ferreira, F., & Clifton, C. (1986). The Independence of Syntactic Processing. *Journal of Memory and Language, 25*, 348–368.

Fowler, F. (1995). *Improving Survey Questions: Design and Evaluation*. Thousand Oaks, CA: Sage.

Fuchs, M. (2008). Mobile Web Surveys: A Preliminary Discussion of Methodological Implications. In F. G. Conrad & M. F. Schober (Eds.), *Envisioning the Survey Interview of the Future* (pp. 77–94). New York, NY: John Wiley.

Groves, R. M., Fowler, F. L., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology*. Hoboken, NJ: John Wiley & Sons.

Hoaglin, D. C., Mosteller, F. & Tukey, J. W. (2000). *Understanding Robust and Exploratory Data Analysis*. New York, NY: John Wiley & Sons.

Höhne, J. K. & Krebs, D. (2017). Scale Direction Effects in Agree/Disagree and Item-Specific Questions: A Comparison of Question Formats. *International Journal of Social Research Methodology, 21,* 91–103.

Höhne, J. K. & Lenzner, T. (2015). Investigating Response Order Effects in Web Surveys using Eye Tracking. *Psihologija*, *48*, 361–377.

Höhne, J. K. & Lenzner, T. (2017). New Insights on the Cognitive Processing of Agree/Disagree and Item-Specific Questions. *Journal of Survey Statistics and Methodology*. DOI: 10.1093/jssam/smx028

Höhne, J. K. & Schlosser, S. (2018). Investigating the Adequacy of Response Time Outlier Definitions in Computer-based Web Surveys using Paradata "SurveyFocus". *Social Science Computer Review, 36,* 369–378.

Höhne, J. K., Schlosser, S., & Krebs, D. (2017). Investigating Cognitive Effort and Response Quality of Question Formats using Paradata. *Field Methods, 29,* 365–382.

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., & 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.

Kunz, T. (2017). Evaluation of Agree-Disagree Versus Construct-Specific Scales in a Multi-Device Web Survey. *Paper presented at the General Online Research conference*, Berlin, Germany.

Kuru, O., & Pasek, J. (2016). Improving Social Media Measurement in Surveys: Avoiding Acquiescence Bias in Facebook Research. *Computers in Human Behavior, 57,* 82–92.

Lelkes, Y. & Weiss, R. (2015). Much Ado about Acquiescence: The relative Validity and Reliability of Construct-Specific and Agree-Disagree Questions. *Research and Politics*.

Lenzner, T., Kaczmirek, L., & Lenzner A. (2010). Cognitive Burden of Survey Questions and Response Times: A Psycholinguistic Experiment. *Applied Cognitive Psychology, 24,* 1003–1020.

Liu, M., Lee, S., & Conrad, F. G. (2015). Comparing Extreme Response Styles between Agree-Disagree and Item-Specific Scales. *Public Opinion Quarterly*, *79*(4), 952–975.

Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public Opinion Quarterly, 72(5),* 914–934.

Mavletova, A. (2013). Data quality in PC and Mobile Web Surveys. *Social Science Computer Review, 31*, 725–743.

Mavletova, A., & Couper, M P. (2013). "Sensitive Topics in PC Web and Mobile Web Surveys: Is There a Difference?", *Survey Research Methods*, *7*(3), 191–205.

Mavletova, A., & Couper, M P. (2015). "A Meta-Analysis of Breakoff Rates in Mobile Web Surveys." In D. Toninelli, R. Pinter, and P. de Pedraza (Eds.), *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies* (pp. 81–98). London, UK: Ubiquity Press.

Mayerl, J., & Urban, D. (2008). *Antwortreaktionszeiten in Survey-Analysen: Messung, Auswertung und Anwendungen*. Wiesbaden: VS Verlag.

Revilla, M. (2017). Analyzing the Survey Characteristics, Participation, and Evaluation across 186 Surveys in an Online Opt-In Panel in Spain. *Methods, Data, Analyses, 11*, 135–162.

Revilla, M., & Couper, M. P. (2018a). Comparing Grids with Vertical and Horizontal Item-by-Item Formats for PCs and Smartphones. *Social Science Computer Review, 36*, 349–368.

Revilla, M., & Couper, M. P. (2018b). Testing Different Order-By-Click Question Layouts for PC and Smartphone Respondents. *International Journal of Social Research Methodology*. DOI: 10.1080/13645579.2018.1471371

Revilla, M., Couper, M. P., & Ochoa, C. (2018). Willingness of Online Panelists to Perform Additional Tasks. *Methods, Data, Analyses*. DOI: 10.12758/mda.2018.01

Revilla, M., Saris, W., & Krosnick, J. A. (2014). Choosing the Number of Categories in Agree-Disagree Scales. *Sociological Methods & Research, 43*(1), 73–97.

Revilla, M., Toninelli, D., Ochoa, C., & Loewe, G. (2016). Do Online Access Panels Really Need to Allow and Adapt Surveys to Mobile Devices? *Internet Research, 26*(5), 1209–1227.

Saris, W. E., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response Options. *Survey Research Method, 4*, 61–79.

Stern, M. J. (2008). The Use of Client-side Paradata in Analyzing the Effects of Visual Layout on Changing Responses in Web Surveys. *Field Methods*, *20*, 377–398.

Toninelli, D., & Revilla, M. (2016a). “Smartphones vs PCs: Does the Device Affect the Web Survey Experience and the Measurement Error for Sensitive Topics? A Replication of the Mavletova & Couper’s 2013 Experiment.” *Survey Research Methods, 10*(2), 153–169.

Toninelli, D, & Revilla, M. (2016b). “Is the Smartphone Participation Affecting the Web Survey Experience?”. *Proceedings of the 48th Scientific Meeting of the Italian Statistical Society*. Salerno. ISBN: 9788861970618.

Toninelli, D., & Revilla, M. (2017). “The Role Played by the Device Screen Size and by the Questionnaire Optimization within the Mobile Survey Participation.” *Paper presented at the General Online Research conference*, Berlin, Germany.

Yan, T., & Olson, K. (2013). Analyzing paradata to investigate measurement error. In F. Kreuter (Ed.), *Improving Surveys with Paradata: Analytic Uses of Process Information* (pp. 73–96). Hoboken, NJ: John Wiley.