QUESTION FORMAT, RESPONSE EFFORT, AND RESPONSE QUALITY – A METHODOLOGICAL COMPARISON OF AGREE/DISAGREE AND ITEM-SPECIFIC QUESTIONS

Dissertation
zur Erlangung des Doktorgrades
der Sozialwissenschaftlichen Fakultät
der Georg-August-Universität Göttingen

vorgelegt von
Jan Karem Höhne M.A.
geboren in Frankfurt am Main

Göttingen, August 2018

Betreuungsausschuss:

Erstbetreuer

Prof. Dr. Steffen-M. Kühnel

Weitere Betreuer

Prof. Dr. Jürgen H.P. Hoffmeyer-Zlotnik

Prof. Jon A. Krosnick, Ph.D.

Tag der mündlichen Prüfung: 30. November 2017

CONTENT

List of Tables

List of Figures

Acknowledgement

First, I would like to thank my academic supervisor Prof. Dr. Steffen-M. Kühnel (University of Göttingen) for his great support, critical advice, and trust in me and my research. I also am thankful to my academic supervisor Prof. Dr. Jürgen H.P. Hoffmeyer-Zlotnik (University of Gießen) for his help and the thoughtful supervision. My special gratitude goes to my academic supervisor Prof. Jon A. Krosnick, Ph.D. (Stanford University) for his supportive and insightful feedback, inspiring discussions, and inviting me to work with him and his Political Psychology Research Group.

The research for this thesis could not have been conducted without my coauthors and their excellent help and kind willingness to share their experience with me. I am grateful to Prof. Dr. Dagmar Krebs (University of Gießen), Dr. Timo Lenzner (GESIS – Leibniz Institute for the Social Sciences), Prof. Dr. Melanie Revilla (RECSM-Universitat Pompeu Fabra), and Stephan Schlosser (University of Göttingen). I also am grateful to Prof. Dr. Willem E. Saris (RECSM-Universitat Pompeu Fabra) for his valuable suggestions on Study II and to Carlos Ochoa (Netquest) for running Study IV and his recommendations.

Additionally, I would like to thank Dr. Veronika Andorfer (IQTIG), Florian Berens (University of Göttingen), Dr. Salima Douhou (City, University of London), Nourhan Elsayed (University of Mannheim), Dr. Sebastian Lundmark (Stanford University), Dr. Cornelia E. Neuert (GESIS – Leibniz Institute for the Social Sciences), Dr. Henning Silber (GESIS – Leibniz Institute for the Social Sciences), and my colleagues at the Center of Methods in Social Sciences (University of Göttingen) and the Political Psychology Research Group (Stanford University) for all the interesting discussions, their support, and the very good time.

I have benefited greatly from the faith and support of my friends. In particular, I would like to thank Amke Bils and Melanie Kipp for being my personal backup. I also am grateful to the "Eckensteherbande"[1], my friends from the University of Göttingen, Johanna Engel, and Fabian Hientz.

Finally, I would like to thank my family for their love and belief. Most importantly, I thank my lovely Mother, who passed away too early. I dedicate this PhD thesis to her. Many thanks!

[1] The term refers to the study "Street Corner Society: The Social Structure of an Italian Slum" published by Whyte (1943) that my university friends and I got to know during our first bachelor semester at the University of Gießen.

1. Introduction

Quantitative social research has a long tradition. One line of development can be identified in the early work of John Graunt and Sir William Petty as representatives of political arithmetic. As described by Schnell, Hill, and Esser (2013), political arithmetic was concerned with the (causal) explanation of social phenomena by means of what we call quantitative methodology and statistical procedures. The data at this time were not based on actual surveys but on birth registers and death records. For instance, Graunt's (1662; cited after Schnell, Hill, & Esser, 2013) book "Natural and Political Observations upon the Bills of Mortality" attempted to draw conclusions about the population size and growth of London.

An early documented empirical study – in the proper sense of quantitative social research – that employed an actual survey was conducted by Karl Marx in 1880; it investigated the working conditions of the French laboring class. The questionnaire entitled "Enquête Ouvrière" was published in the journal "La Revue Socialiste" and also was distributed via work societies and social organizations. Of course, this is only one example of the early studies that employed quantitative research methods.

In the 1930s and 1940s, several survey research-oriented organizations and institutes were established. For instance, the Gallup Organization located in Washington, DC was founded in 1935, the "National Opinion Research Center (NORC)" located at the University of Chicago was founded in 1941, the "Roper Center for Public Opinion Research" located at Cornell University was founded in 1947, and the "Institut für Demoskopie Allensbach (IfD)" located in Germany also was founded in 1947. During that time,

survey researchers started to investigate the effects of survey questions on response behavior, such as the order of response categories (Mathews, 1929; Rugg & Cantril, 1944) and the wording of questions (Rugg, 1941), and identified error sources that could affect survey responses (Deming, 1944). For instance, the Gallup Organization conducted many split-ballot experiments on question forms and context (see Bishop & Smith, 2001, p. 484). In addition, psychologists and social scientists proposed a number of methods to measure respondents' attitudes (Guttman, 1944; Likert, 1932; Osgood, Suci, & Tannenbaum, 1957; Thurstone, 1929).

In the 1960s and 1970s, a "psychological revolution" occurred in which cognitive science was applied to several behavioral research fields, such as survey research (Willis, 2008, p. 104). Survey researchers began to adopt theoretical approaches from psychology to explain the differences in response behavior and occurrences of response bias. The climax of this development was reached in 1983 and 1984 with the establishment of "Cognitive Aspects of Survey Methodology (CASM)" as an interdisciplinary research field that combines survey research and psychology (Schwarz, 2007).[2] Since the establishment of CASM, a huge number of edited volumes (see de Leeuw, Hox, & Dillman; 2008; Hippler, Schwarz, Sudman, 1987; Krebs & Schmidt, 1993; Kreuter, 2013; Lyberg et al., 1997; Marsden & Wright, 2010; Presser et al., 2004; Schwarz & Sudman, 1996; Sirken et al., 1999; Vannette & Krosnick, 2017) and monographs (see Biemer & Lyberg, 2003; Bradburn, Sudman, Wansink, 2004; Callegaro, Lozar Manfreda, & Vehovar, 2015; Fowler, 1995; Groves et al., 2004; Saris & Gallhofer, 2014; Sudman, Bradburn, & Schwarz, 1996; Tourangeau, Rips, & Rasinsky, 2000) have been published. Furthermore, several models of the cognitive response process were proposed (see Cannell, Miller, &

[2] See also chapter 2.2 "Response Effort".

Oksenberg, 1981; Carpenter & Just, 1975; Krosnick, 1991; Strack & Martin, 1987; Tourangeau, 1984; Tourangeau, Rips, & Rasinski, 2000).

Although the last three decades have been characterized by major scientific advances and a huge gain of knowledge, open questions still remain with respect to question and questionnaire design. For instance, Fowler's (1995) monograph "Improving Survey Questions: Design and Evaluation" has pointed to several methodological flaws associated with the agree/disagree (A/D) question format, and has argued that employing the item-specific (IS) question format represents a simpler, more direct, and more informative method (p. 56-57).[3] Two conclusions can be drawn from Fowler's argument: (1) Responding to A/D questions requires more effort than responding to IS questions. (2) Responses to A/D questions are characterized by lower response quality than responses to IS questions. To date, these assumptions partially lack empirical evidence and, thus, this thesis intends to contribute to closing this substantial research gap.[4]

To investigate the response effort and response quality associated with the A/D question format and the IS question format, the present thesis discusses the theoretical and methodological background of these formats. This includes a presentation of the theoretical approach developed in this thesis to explain the differences of these two question formats. In addition, the present thesis outlines the current research gap that has motivated the research questions under consideration. Then, the conceptual specifications, general research designs, and characteristics of the empirical studies will be expounded. The main

[3] Fowler (1995) did not use the term item-specific (IS) question format; rather, he used the term "direct rating task". In addition, some researchers speak of a construct-specific (CS) question format. This thesis, however, follows the terminology used by Saris et al. (2010) and, thus, uses the term item-specific (IS).
[4] See chapter 3. "General Research Questions", which outlines the current state of research.

body presents the empirical studies conducted for this thesis.[5] Finally, a general discussion and conclusion reflect the overall empirical findings in the light of the theoretical approach developed and also consider the practical implications and perspectives for future research.

[5] Versions of three of these chapters are already published in "Journal of Survey Statistics and Methodology" (chapter 5.2), "Field Methods" (chapter 5.3), and "International Journal of Social Research Methodology" (chapter 5.5). A version of chapter 5.1 is in press at "International Journal of Social Research Methodology" and a version of chapter 5.4 is in press at "Methodology: European Journal of Research Methods for the Behavioral and Social Sciences".

2. Theoretical Background

The first part of this thesis outlines the current state of research and represents a reflection and refinement of the theory on A/D and IS question formats. The structure and organization of this part fall under the title “Question Format, Response Effort, and Response Quality”.

2.1 Question Format

The field of quantitative social research encompasses a variety of disciplines, such as economics, political science, psychology, and sociology. All of these disciplines have something in common, namely the measurement of people’s attitudes toward particular issues. The goal of measuring attitudes is to draw conclusions about topics that include electoral behavior, political outcomes, social and demographic developments, media exposure, belief systems, public health, and consumer behavior. Attitudes, thereby, can be seen as representations of the beliefs, evaluations, feelings, impressions, and values held by a person toward an object (Eagly & Chaiken, 1998; Eaton & Visser, 2008; Fazio & Olson, 2003; Tourangeau, Rips, & Rasinsky, 2000). For instance, a person’s attitude toward a political candidate is a representation of the extent of positive impressions (i.e., “favor” tendencies) or negative impressions (i.e., “oppose” tendencies) held by person toward the candidate.

However, social psychological research indicates that people do not necessarily have a set of predetermined attitudes in mind but may build them “on the spot” when they are required to do so (Bassili & Fletcher, 1991; Converse, 1964; Eagly & Chaiken, 1998; Fazio & Olson, 2003; Saris & Gallhofer, 2014; Strack & Martin, 1987; Tourangeau, Rips,

& Rasinsky, 2000).[6] The way attitudes are built – either by drawing on pre-existing ones or by creating them instantly – seems to depend on the specific attitude in question and its strength.[7]

Attitudes represent latent variables or constructs that are operationalized and gathered by means of manifest variables or survey questions.[8] In order to measure attitudes, scientists generally use questionnaires, sets of (standardized) questions that are used to collect individual data about one or more topics (Trobia, 2008, p. 652). The structure of survey questions used to measure attitudes can be characterized by a question stem – including instructions and/or definitions – and a response scale (Roe, 2008). The stem represents the first part of the survey question and conveys the content. Typically, it is formulated as a question or a statement, depending on the question format used.[9]

Response scales used to measure attitudes generally have an ordinal scale but, in statistical analyses, they are frequently treated as continuous measures. According to Stevens (1946, p. 679), "the ordinal scale arises from the operation of rank-ordering. Since any 'ordering-preserving' transformation will leave the scale form invariant, this scale has the structure of what may be called the isotonic or order-preserving group". In other words, ordinal measures distinguish between the characteristics of objects with respect to equality (i.e., equal vs. unequal) and rank these characteristics (i.e., one value is greater

[6] For a comprehensive discussion of attitudes and their foundation, structure, and function, see Aronson, Wilson, and Akert (2014).
[7] A detailed theoretical discussion of the concept of "attitude strength" is beyond the scope of this chapter. Therefore, interested readers are referred to further literature (see Bauman, 2008).
[8] Groves et al. (2004, p. 43) speak of "measurements" or "survey measurements" to take their possible diversity into account (e.g., interviewer observations or blood pressure measurements).
[9] Saris and Gallhofer (2014) provide a comprehensive discussion of different types of so-called "requests for an answer" in light of linguistic and survey methodological considerations.

than the other one). However, they do not allow to draw conclusions with certainty about the exact distance or interval between the characteristics of the object.[10]

Most commonly, ordinal scales are based on so-called "closed-ended" response scales. Closed-ended here means that respondents can select from a specified set of response categories (Lavrakas, 2008a; Oldendick, 2008). Krosnick and Presser (2010) argue that, in order to ensure appropriate measurement, that such response scales should meet certain criteria: (1) The response categories should cover the entire response continuum and run from one side of the continuum to the other, (2) the meanings of adjacent response categories should not overlap, (3) all response categories should elicit a precise and stable understanding, (4) all respondents should agree in their interpretations of each response category. Violations of these methodological requirements might affect respondents answers in an undesirable way and, thus, reduce response quality (see chapter 2.3 "Response Quality" for a definition of response quality).

Höhne and Krebs (2018) recommend four additional design aspects that researchers must consider when designing response scales: (1) Deciding whether the response scale contains a midpoint (using an even or uneven number of categories), (2) determining the scale length (specifying the number of scale points), (3) deciding to what extent the response scale is verbally and/or numerically labeled (employing fully or partially labeled scales), (4) specifying the direction of the response scale (applying a decremental, i.e., from positive to negative, or incremental, i.e., from negative to positive, response order).

[10] According to Kühnel (1993), it is assumed that in the special case of attitude questions, the position of a respondent lies on a latent (metric) continuum, but the measurement of this position by means of response categories only represents an imprecise measurement of the latent variable or construct. In other words, if a respondent is asked to indicate his or her political interest using a rating scale, it can be expected that the choice of the response category depends on his or her location on the continuum.

Furthermore, researchers must decide between a unipolar or bipolar response scale (Kennedy, 2008) and a horizontal or vertical arrangement of the response categories (Höhne & Lenzner, 2015).[11] Although these design aspects have been discussed intensively in the survey literature, there is no consensus with respect to their implementation.[12]

There is an abundance of research on the development of appropriate question formats to measure attitudes (Guttman, 1944; Likert, 1932; Osgood, Suci, & Tannenbaum, 1957; Parducci, 1983; Thurstone, 1929; Tourangeau, Rips, & Rasinsky, 2000). Since the publication of Likert's (1932) well-known article on the measurement of attitudes, the A/D question format has grown in popularity.[13] This measurement technique enables researchers to measure both the respondent's evaluation of the object to be rated – e.g., "agree" vs. "disagree" – and the strength of the evaluation – e.g., "strongly" vs. "somewhat" (Liu, Lee, & Conrad, 2015). In particular, the A/D question format usually starts with a pre-request, followed by an indirect statement and a response scale where the categories are based on an agreement/disagreement continuum. See the following example of an A/D question:

Do you agree or disagree with the following statement?

I am interested in politics.

Agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, disagree strongly

[11] The polarity of the response scale depends on the dimensionality or content of the specific construct (Kennedy, 2008).

[12] For a comprehensive description of response scale characteristics and their influence on response quality, see DeCastellarnau (2017) and Krosnick (1999).

[13] Misleadingly, A/D questions are frequently referred to as "Likert questions" or "Likert items". However, Likert (1932) only tested 5-point fully labeled "approval/disapproval" response scales.

The A/D question format allows researchers to transform each statement into a request for an answer, irrespective of the content (Saris & Gallhofer, 2014).[14] Thereby, it is expected that the A/D statements and response scales can be used to measure a latent variable or construct, provided that the statement represents a characteristic of that variable or construct. Furthermore, it is assumed that ordering responses on an agreement/disagreement continuum is equivalent to ordering them on the actual content dimension (Krosnick & Presser, 2010, p. 278).

Major national and international surveys, such as the American National Election Study (ANES), the Eurobarometer, and the International Social Survey Programme (ISSP) use the A/D question format. As a consequence, many empirical findings in the social sciences and adjacent research fields are based on A/D questions. The reasons for the popularity of this question format are twofold (Saris et al., 2010). First, A/D questions streamline questionnaires by allowing researchers to inquire about different topics (e.g., social inequality and achievement motivation) using the same response scale for all questions. Second, A/D questions save space in self-administered surveys and save time in both self- and interviewer-administered surveys, particularly if grids are used.

Despite these methodological benefits, A/D questions have been criticized for several methodological drawbacks, the most prominent of which is their proneness to different kinds of response bias (Baumgartner & Steenkamp, 2001; Converse & Presser, 1986; Höhne & Lenzner, 2015; Holbrook, 2008; Krosnick, 1991; Krosnick, Narayan, & Smith, 1996; Schuman & Presser, 1981; van Vaerenbergh & Thomas, 2013).[15] Due to A/D questions' methodological disadvantages, survey researchers, including Fowler (1995) and

[14] The term "request for an answer" considers that surveys frequently do not use requests but statements. Nonetheless, respondents are expected to provide an answer (Saris & Gallhofer, 2014).
[15] See chapter 2.3 "Response Quality" for a discussion of response bias and how this term is used in this thesis.

Krosnick and Presser (2010), recommend the use of the IS question format. Unlike the A/D question format, the IS question format usually consists of an interrogative request for an answer and a response scale that directly matches the underlying dimension of the latent variable or construct. See the following example of an IS question:

How interested would you say you are in politics?
Very interested, fairly interested, somewhat interested, hardly interested, not at all interested

According to Fowler (1995, p. 57), the IS question format represents a simpler, more direct, and more informative way of asking survey questions than the A/D question format. Furthermore, he identified four methodological drawbacks that are generally associated with the A/D question format but not with the IS question format:

1. Insufficient anchoring of the question stem at one end of the content continuum.[16]
2. Complex cognitive processing (this is addressed by Study I, II, III, IV, and V in this thesis).
3. Low discriminatory power of adjacent response categories (this is addressed by Study III and IV in this thesis).
4. Proneness to response bias (this is addressed by Study III, IV, and V in this thesis).

[16] This point is not addressed in this thesis for feasibility reasons. However, it should be addressed by future studies using a combination of cognitive interviewing and eye-tracking methodology (for a comprehensive methodological discussion, see Neuert & Lenzner, 2016).

The first drawback identified by Fowler (1995) is that the question stems of A/D questions are frequently not clearly anchored at one end of the content continuum, making it difficult to interpret responses unambiguously. For instance, in the example of the A/D question on political interest given above, the respondents could disagree for different reasons: either because they are not interested in politics or because they are not only interested in but also passionate about politics (i.e., just being interested in politics does not adequately represent their attitude toward politics). This characteristic of A/D questions impedes data analysis and interpretation, as respondents holding different views are forced into the same response categories. It could also negatively affect response quality, because if respondents perceive the ambiguity and difficulty involved in responding to such A/D questions meaningfully, they may decide that resolving this problem is too wearisome and then may offer a non-substantive response or even skip the question. According to Saris and Gallhofer (2014), the IS question format normally does not suffer from the problem of insufficient anchoring, as long as the response categories are exhaustive and appropriately specify both ends of the content continuum (i.e., the categories represent fixed reference points).[17]

A second problem associated with the A/D question format that Fowler (1995) identified is that responding to such questions seems to be relatively complex and intricate. In theory, answering A/D questions requires respondents to accomplish the following mental tasks (see Carpenter & Just, 1975; Fowler, 1995; Fowler & Cosenza, 2008; Krosnick & Presser, 2010; Pasek & Krosnick, 2010; Saris & Gallhofer, 2014; Saris et al., 2010): (1) Comprehending the literal meaning of the question (i.e., what do the individual words mean?), (2) identifying the pragmatic meaning (e.g., the intensity of interest in politics),

[17] For an introduction to adverbial intensifiers and their cognitive and communicative implications for survey responses, see Cliff (1959) and Schaeffer and Charng (1991).

(3) placing themselves on that dimension (i.e., how interested are they in politics?), (4) calculating where on that dimension the stem of the question lies (i.e., where does "interested" lie on a continuum ranging from "very interested" to "not at all interested"?), (5) evaluating the distance between their own position on the dimension and the position of the question stem, (6) translating this judgement into the A/D response categories. In the IS question format, identifying the pragmatic meaning (task 2) and mapping an answer onto the response scale (task 6) seems to be less effortful. Moreover, performing task 5 is usually not required when answering IS questions (Höhne & Lenzner, 2017). Hence, it is reasonable to assume that answering questions in the IS format is a less complex endeavor than answering questions in the A/D format (for a more detailed discussion of the effort involved in responding to A/D and IS questions, see chapter 2.2 "Response Effort").

The third problem associated with A/D questions is that their response categories have a low discriminatory power, since they are ordinarily analyzed in terms of "agree" vs. "disagree" (Fowler, 1995). This means that A/D questions yield limited information about the content of interest (Liu, Lee, & Conrad, 2015). In contrast, the response categories of IS questions directly address the scale of interest, making it possible to better preserve information about the content of interest.

The fourth and most serious problem associated with A/D questions is that, as many studies have shown, such questions are prone to response bias (Baumgartner & Steenkamp, 2001; Converse & Presser, 1986; Höhne & Lenzner, 2015; Holbrook, 2008; Krosnick, 1991; Krosnick, Narayan, & Smith, 1996; Schuman & Presser, 1981; van Vaerenbergh & Thomas, 2013). Although there are several theoretical explanations for the occurrence of response bias in A/D questions, the reason as to why this question format causes such bias has not been conclusively identified. Krosnick (1991) and Krosnick and

Alwin (1987) provide a convincing argument; they posit that respondents are not always willing or motivated to expend the effort necessary to answer a survey question in the optimal way. Instead, respondents try to shortcut the response process, for example, by agreeing with statements presented to them in the A/D question format (Krosnick, 1991). Additionally, Höhne, Schlosser, and Krebs (2017) argue that A/D questions promote states of boredom and weariness and a superficial cognitive processing due to an indirect (i.e., A/D statements do not refer to the dimension of interest) and unchanging manner of asking (i.e., response categories are always based on an agreement/disagreement continuum). In consequence, they can dismay respondents and discourage them from putting the effort required to respond meaningfully. IS questions, in contrast, contain a direct question and commonly change the manner of asking, thereby presumably preserving respondents' attention and/or motivation.[18]

To summarize: Respondents do not have to read the A/D response categories repeatedly since the manner of asking is invariant. Moreover, the indirectness of the A/D statements may promote superficial rather than meaningful response behavior or impede responding altogether. In contrast to A/D questions, IS questions are, in theory, simpler to respond to because they are more direct. However, IS questions are more cognitively demanding than A/D questions as they usually require respondents to continuously reconsider the underlying content dimension (i.e., reading the question including response categories). Thus, the IS question format joys a comparative advantage that the A/D question format lacks, namely encouraging respondents to perform a more active and intensive

[18] IS questions do not necessarily have to change the manner of asking if all questions address the same content dimension (e.g., frequency or intensity). In this special case, IS questions can be presented in grid presentation mode (see Study III and V).

responding (Höhne & Krebs, 2018; Höhne & Lenzner, 2017; Höhne, Revilla, & Lenzner, 2018; Höhne, Schlosser, & Krebs, 2017).

2.2 Response Effort

Having discussed the measurement of attitudes in general and the characteristics of A/D and IS questions in particular, the goal is now to take a closer look at response effort and its facets. In the 1960s and 1970s, survey researchers incorporated cognitive psychological theories into survey research to understand the processes underlying survey responses. This early scientific development laid the foundation for a new interdisciplinary research field, known today as "Cognitive Aspects of Survey Methodology (CASM)".[19] During this period, survey researchers also started to investigate the burden associated with survey participation (see Bradburn, 1978; Frankel, 1981; Noelle-Neumann, 1976; Sharp & Frankel, 1983; Tourangeau, 1984), proposing a variety of different concepts and definitions to study the phenomenon: Respondent burden (see Bradburn, 1978; Frankel, 1981; Graf, 2008; Sharp & Frankel, 1983; Wenemark et al., 2010), task difficulty (see Krosnick, 1991; Krosnick, Narayan, & Smith, 1996; Tourangeau, 1984), response burden (see Haraldsen, 2004; Hoogendoorn & Sikkel, 1998; Jäckle, 2008; Jones, 2012; Rolstad, Adler, & Rydén, 2011; Yan et al., 2016), cognitive burden (see Lenzner, Kaczmirek, & Lenzner, 2010), cognitive satisficing (see Hoogendoorn, 2004), respondent effort (see Revilla & Couper, 2018a), respondent fatigue (see Ben-Nun, 2008), and subject burden (see Rolstad, Adler, & Rydén, 2011).[20] In the following, an overview of the most established concepts – respondent burden and task difficulty – will be provided[21], accompanied by a definition of the notion of response effort and its methodological localization.

[19] CASM was established by two major conferences. The first one, titled "Advanced Research Seminar on Cognitive Aspects of Survey Methodology", took place in the U.S. in 1983 (see Jabine et al., 1984). The second one, titled "Conference on Social Information Processing and Survey Methodology", took place in West Germany in 1984 (see Sudman, Bradburn, & Schwarz, 1996).

[20] Many of these concepts are based on relatively unclear definitions. Furthermore, they are frequently interchangeably used.

[21] A discussion of each concept mentioned above is beyond the scope of this chapter. Therefore, interested readers are referred to the literature cited in the text.

Respondent burden: Bradburn (1978) seems to have made the first attempt to develop a concept for the burden caused by survey participation. Referred to as respondent burden, the concept consists of four facets that interact with one another and affect the perception of burden across respondents in different ways: (1) The length of the interview, (2) the amount of effort required of the respondent, (3) the amount of stress on the respondent, (4) the frequency with which the respondent is interviewed.

The first facet of respondent burden, interview length, can be measured using several (divergent) strategies (Bradburn, 1978; Haraldsen, 2004)[22]: the total time of the interview, the number of pages, the number of questions, or even the number of words per question.[23] However, the most commonly used measure is the total time of the interview (e.g., minutes) to determine the length of an interview (Galesic, 2006; Graf, 2008; Groves et al., 1999; Jäckle, 2008; Sharp & Frankel, 1983; Stocké & Langfeldt, 2004). Irrespective of the measure used, it is assumed that interview length is positively correlated with respondent burden.[24]

The second facet of respondent burden identified by Bradburn (1978) respondent effort is the sum of actions that a respondent has to undertake to participate in and/or complete the survey. For instance, it includes the way of survey completion (e.g., respondents have to complete the survey in a lab or schedule a phone call), the difficulty of answering questions (e.g., identifying the pragmatic meaning or mapping an answer), and

[22] According to Bradburn (1978), the term "interview length" can refer to a single interview (i.e., in a cross sectional study) but also to the total length of repeated interviews (i.e., in a longitudinal study).
[23] Although it can be assumed that these parameters are related to each other and might be interchangeable, from a conceptual point of view, they address different aspects. For instance, the number of questions can differ greatly from page to page. Especially, in web-based surveys (e.g., paging vs. scrolling surveys, see Peytchev et al., 2006). Thus, the number of questionnaire pages seems to be a less precise measure.
[24] There are many studies, which investigate the relation between interview length (i.e., total time of interview) and several indicators of primarily low response quality, such as non-substantial responses (see Galesic, 2006; Rolstad, Adler, & Rydén, 2011).

the questionnaire design (e.g., navigating through the instrument).[25] Like interview length, respondent effort is positively correlated with respondent burden.

The third facet of respondent burden, namely respondent stress, refers to the amount of personal discomfort that a respondent undergoes during the course of an interview (Bradburn, 1978, p. 38). Discomfort is strongly linked to survey questions (e.g., questions about sensitive topics, such as drug abuse) and to interview settings (e.g., interviews involving the presence of third parties). In comparison with the facets of interview length and respondent effort, respondent stress is a relatively vague facet and might vary widely across respondents, because it heavily depends on individual perceptions (Bradburn, 1978; Haraldsen, 2004).

The fourth facet, interview frequency, refers to the number of interviews or survey completions (Bradburn, 1978; Haraldsen, 2004; Hoogendoorn & Sikkel, 1998). The frequency of interviews is a particularly important issue in longitudinal studies (i.e., panel studies), where respondents are repeatedly interviewed.[26] However, the frequency of independent interviews (i.e., non-panel studies) in which respondents participate is also an important aspect of the assessment of respondent burden. That is because in the last two decades, the number of surveys conducted has steadily increased, which means that respondents today are likely to participate in far more surveys than respondents decades ago.[27]

[25] The ongoing methodological debate on how to design mobile web-surveys (e.g., the presentation of grid questions or placement of the "Next" button on the web-survey page) can be ascribed to the concept of respondent effort (see Revilla & Couper, 2018a).
[26] The time intervals between the interviews might also affect respondent burden (Hoogendoorn & Sikkel, 1998).
[27] For instance, the ADM Group (Arbeitskreis Deutscher Markt- und Sozialforschungsinstute) shows that the number of surveys in Germany has more than quadrupled from 1990 to 2016 (ADM Group, 2017). Furthermore, there is research dealing with the effects of survey experience on survey participation and responding (see Toepoel, Das, & van Soest, 2008).

Task difficulty: In the course of the CASM movement, Tourangeau (1984) proposed a general model of the cognitive processes underlying survey responses.[28] The model includes four components: (1) Question comprehension, (2) information retrieval, (3) judgment formation, (4) response reporting.[29] Thus, task difficulty can be described as the difficulty posed by a survey question to perform each step of the response process (Krosnick, 1991; Krosnick, Narayan, & Smith, 1996; Tourangeau, 1984).[30] The concept of task difficulty fits best with Bradburn's (1978) facet of respondent effort.

It is well-known that the capacity of the working memory is limited (see Just & Carpenter, 1992; MacDonald & Christiansen, 2002). Hence, having a high number of words per question might negatively affect question comprehension (Krosnick, 1991).[31] Furthermore, question comprehension can be complicated by certain linguistic features. For instance, Graesser et al. (2006) found that several text features, including complex syntactical structures or low frequency words, impede comprehension. Additionally, a survey question with an unclear pragmatic meaning impedes the comprehension of a question because respondents need to decipher the dimension of interest. This is frequently the case with the A/D question format (see Carpenter & Just, 1975; Fowler, 1995; Fowler & Cosenza, 2008; Krosnick & Presser, 2010; Saris & Gallhofer, 2014; Saris et al., 2010).

[28] In all likelihood, Cannell, Miller, and Oksenberg (1981) developed the first cognitive response process model, which distinguishes between two different paths of response formation. While one path implies a careful response process, the other path implies a superficial one.

[29] It might appear that the cognitive response process occurs in a sequence but it is more reasonable to assume that the components partially overlap (e.g., the judgment formation already starts during the information retrieval) or that respondents sometimes move back and forth between the components (Tourangeau & Bradburn, 2010).

[30] According to Krosnick (1991), it must be mentioned that task difficulty also depends on respondent's characteristics, such as motivation and/or interest.

[31] Long questions can also enhance comprehension if they contain clarifications, such as elaborations on the meanings of terms (see Conrad et al., 2006; Krosnick, 1991; Metzler, Kunz, & Fuchs, 2015). For this reason, question length is not the best indicator of task difficulty.

The difficulty of information retrieval is strongly related to the type of information that a respondent is asked to retrieve. For instance, in a recall question on cinema attendances, it is easier for respondents to answer questions that require them to recall the attendances during the last month than during the last year. In attitude questions, the difficulty of information retrieval depends on whether respondents can draw on a pre-existing attitude or whether they need to collect information to form an attitude on the spot (Strack & Martin, 1987).[32] The number of objects addressed by a question affects retrieval difficulty as well (Krosnick, 1991; Tourangeau, Rips, & Rasinski, 2000).

Task difficulty can also be affected during the stage judgment formation. The amount of information that a question requires a respondent to retrieve can have a substantial impact on the difficulty of building a judgment. The more information a respondent must retrieve, the more difficult it is to come up with a judgment. However, due to time pressure during the survey and/or motivational aspects, it is unlikely that respondents retrieve all potentially relevant information (Sudman, Bradburn, & Schwarz, 1996).[33] It is more likely that respondents stop the information collection process once they have enough information to form a judgment (see Bodenhausen & Wyer, 1987).[34]

Finally, task difficulty can be affected by characteristics of the response scale or format, such as the length of the response scale (see Krosnick & Presser, 2010), labeling of the response categories (see Menold et al., 2014), arrangement of the response categories (see Höhne & Lenzner, 2015), or terms of the response categories (see Krosnick,

[32] In social psychology response times are frequently used to assess the accessibility and strength of attitudes (Mayerl & Urban, 2008).
[33] There is a large body of research on judgmental heuristics. For a comprehensive examination of judgmental heuristics, see Schwarz et al. (1991) and Tversky and Kahneman (1973, 1974).
[34] Although Bodenhausen and Wyer (1987) provide a comprehensive discussion of respondents' cognitive information acquisition, they do not specify a "threshold" that indicates a sufficient amount of information.

1991). Questions that directly address the underlying dimension of interest in the response categories, such as IS questions, can reduce task difficulty.

Response effort: In contrast to the previously described concepts, response effort explicitly refers to specific characteristics of a task (i.e., responding to a survey question) without considering general interview and/or survey sources (e.g., interview length and questionnaire layout) or respondent characteristics (e.g., age or need for cognition).[35] A task here is defined as responding to a survey question that is posed in a specific format – in other words, responding to A/D or IS questions.[36] The specific characteristics referred to are the directness and continuity of the manner of asking survey questions (see Höhne & Krebs, 2018; Höhne & Lenzner, 2017; Höhne, Revilla, & Lenzner, 2018; Höhne, Schlosser, & Krebs, 2017). While directness refers to the actual form of the request for an answer, continuity refers to the (in)variation of the response scales. Note that the concept of response effort differentiates between two types of question complexity: The first type is the theoretically expected complexity of question formats (i.e., the required cognitive response processes, such as question comprehension). The second type is the actual effort respondents expend in responding (i.e., the effort invested in performing the required cognitive response processes, such as a thoughtful or superficial question comprehension).

[35] External sources might decrease or increase the effort required by survey tasks but they do not change the relation between these tasks in terms of effort. For instance, if performing a task (e.g., answering a question either in the A/D or IS format) on a PC is more effortful than another task, this would similarly apply to smartphones (see Study IV).

[36] There are other question formats that can be employed to measure respondents' attitudes, such as the semantic differential (Osgood, Suci, & Tannenbaum, 1957). However, this thesis focuses on the A/D and IS question formats only.

According to Saris and Gallhofer (2014), requests for an answer can be generally divided into statements (declarative form) and questions (interrogative form).[37] As mentioned in chapter 2.1 "Questions Format", the A/D question format poses statements that indirectly refer to the content dimension and, thus, represent an indirect manner of asking. This situation may reduce response quality and impede question comprehension in terms of the pragmatic meaning. In theory, respondents may deduce the pragmatic meaning by using additional information, such as the context of a question and/or the response categories (Sudman, Bradburn, & Schwarz, 1996). This, of course, would increase response effort. In contrast, the IS question format represents a direct manner of asking because it directly addresses the underlying dimension of interest in the question stem. For this reason, it is unlikely that IS questions impede question comprehension due to an unclear pragmatic meaning.[38]

The second characteristic addressed by the concept of response effort is the continuity of the manner of asking. In terms of continuity, question formats can be divided into formats with tailored and formats with non-tailored response scales. Similar to "yes/no" and "true/false" questions, A/D questions are based on an invariant manner of asking since they apply identical response scales (e.g., from "strongly agree" to "strongly disagree") for all questions, which forces respondents to repeat the same answering task over and over again (Höhne and Krebs 2017; Höhne & Lenzner, 2017; Höhne, Revilla, & Lenzner, 2018; Höhne, Schlosser, & Krebs, 2017). The A/D response categories also do not match the dimension of interest, which makes the mapping of the answers more

[37] The authors also distinguish between orders and instructions (imperative form) but, in the context of this thesis, only the declarative and interrogative forms are of interest.
[38] As stated in chapter 2.1 "Question Format", A/D statements are often not clearly located at one end of the content dimension, which makes it difficult to interpret responses unambiguously. However, this problem can be avoided by a careful question stem design.

difficult. For instance, if respondents are asked to agree or disagree with the statement *"I am interested in politics."*, they must first decide how interested they are in politics, then translate their choice into the A/D response scale (see Krosnick & Presser, 2010). As mentioned earlier, the low discriminatory power of A/D response categories (e.g., what exactly is the difference between "strongly agree" and "somewhat agree"?) might also impede responding (Fowler, 1995).[39] To sum up, the mapping process and the low discriminatory power involved in A/D response categories may increase response effort. IS questions, on the contrary, directly express the content dimension in the response categories, which generally simplifies the mapping process and may enhance the discriminatory power of the response categories.[40]

To carefully respond to A/D questions, respondents are required to perform the cognitive tasks stated in chapter 2.1 "Question Format" – comprehending the meaning of the question (tasks 1 and 2), self-placement on the content dimension (task 3), determining the stem position in relation to the own placement (tasks 4 and 5), and mapping (task 6). However, when responding to A/D questions, respondents read the response categories once because the response categories can be extrapolated from one question to another (Höhne & Lenzner, 2015). This also implies that respondents do not have to consider each A/D question carefully and in the light of its specific content. Consequently, the A/D question format induces a kind of routine responding. In line with Simon's (1957) considerations on choice behavior, Krosnick (1991) supposes that respondents often only

[39] This might apply to IS questions as well, because they also are based on rating scales that employ vague adverbial intensifiers (e.g., fairly and hardly).
[40] The response scale length (e.g., 5-points or 7-points) is an additional characteristic that can influence the difficulty of the mapping (see Krosnick & Presser, 2010). This is addressed by Study IV.

expend the minimum effort required to provide an acceptable response to a survey question.[41] Hence, it is likely that the specific characteristics of the A/D question format (i.e., the indirect and invariant manner of asking) incite respondents to satisfice (see Krosnick, 1991; Krosnick & Alwin, 1987). In other words, the A/D question format incites respondents to perform the response tasks in a superficial manner (e.g., mapping the answer to the first acceptable response category) or partially skip response tasks (e.g., tasks 4 and 5). Such response behavior implies less response effort since it requires less cognitive processing.

From a psychological perspective, the IS question format is generally simpler to process than the A/D question format since it requires less cognitive tasks. However, in contrast to A/D questions, IS questions usually demand a continuous reconsideration of the dimension of interest. This implies that respondents have to continuously re-read the question stems, including response categories, which implies higher cognitive processing and, thus, greater effort in responding (see Höhne, Schlosser, & Krebs, 2017). The specific characteristics of the IS question format (i.e., the direct and variant manner of asking) incite respondents to optimize (see Krosnick, 1991; Krosnick & Alwin, 1987). In other words, the IS question format incites respondents to perform the response tasks thoroughly and in a well-considered manner without resorting to any cognitive shortcuts.

The previous theoretical reasoning partially contradicts the prevalent scientific view associated with the A/D and IS question formats. Such reasoning also implies that response effort does not only depend on the cognitive processes that are required, but also on the way of responding, namely the degree of conscientiousness and elaborateness involved in carrying out these processes. To put it differently, even if a task (e.g., answering

[41] For a discussion of the satisficing theory (Krosnick, 1991; Krosnick & Alwin, 1987), see chapter 2.3 "Response Quality".

a question in the A/D format) is theoretically more complex compared to another task (e.g., answering a question in the IS format), the actual response effort involved in the former can be lower than the latter due to the way of responding. As a result, it is to expect that the IS question format is associated with a higher level of response effort than the A/D question format (Höhne & Lenzner, 2017; Höhne, Revilla, & Lenzner, 2018; Höhne, Schlosser, & Krebs, 2017).

2.3 Response Quality

As shown in the previous chapters, the A/D and IS question formats differ with respect to the cognitive processing they require and, thus, it is to presume that they also differ in terms of response quality. Since there are several (and often divergent) concepts of quality in survey research, it is necessary to provide a definition of response quality. Deming (1944) seems to have proposed the first typology of the sources of error that negatively affect response quality – e.g., bias arising from questionnaire design or nonresponse.[42] This contribution can be seen as the springboard of the concept of the "total survey error" (Groves & Lyberg, 2010).[43] "Total survey error refers to the accumulation of all errors that may arise in the design, collection, processing, and analysis of survey data" (Biemer, 2010, p. 817). Since Deming (1944) first discussed it, the total survey error concept has been heavily refined by survey researchers (see Biemer, 2010; Biemer & Lyberg, 2003; Callegaro, Lozar Manfreda, & Vehovar, 2015; Fuchs, 2008; Groves, 2004; Groves et al., 2004; Weisberg, 2005). In particular, the efforts to refine the concept culminated in the identification of a wider range of error sources and the development of error typologies.

Biemer and Lyberg (2003), for instance, proposed one of the most influential typologies; one that distinguished between sampling error (i.e., errors that arise due to selecting a sample instead of the population) and non-sampling error (i.e., errors that arise due to flawed data collection and processing procedures). In this typology, non-sampling error is further divided into specification error, frame error, nonresponse error, measurement error, and processing error. Weisberg (2005), in contrast, suggested a more elaborate error typology, which distinguished between respondent selection issues, response accuracy

[42] The entire list consists of 13 potential error sources, including several sub-items (see Deming, 1944). However, not all of them are related to this thesis.

[43] Errors can be classified into systematic error – i.e., measures are consistently different from the true value in one direction – and random error – i.e., measures vary inconsistently from the true value (Miller, 2008).

issues, and survey administration issues. In accordance with this typology, Groves et al. (2004) introduced the concept of the survey lifecycle from a quality perspective. The survey life cycle consists of two parallel paths, each of which is based on several successive components. The first path is the measurement path, which addresses the question "which data are to be collected about the units in the sample?" The second path is the representation path, which addresses the question "which populations are described by the survey?" Groves et al. (2004) also identify several survey quality concepts between the individual components with each quality concept indicating a mismatch between the pairs of successive components. The measurement path is of particular interest in the attempt to achieve an appropriate description and definition of response quality because, in contrast to the representation path, the measurement path explicitly addresses the process of gathering survey responses.[44] Figure 1 illustrates the survey lifecycle from a quality perspective.

As stated in chapter 2.1 "Question Format", attitudes represent constructs (first component in the measurement path), which are measured by means of survey questions (second component in the measurement path). Validity (first quality concept) can be seen as the strength of the relationship between the construct and the measurement (Saris & Gallhofer, 2014).[45] The challenge of achieving high validity lies in designing questions that most accurately reflect the constructs being measured (Groves et al., 2004). Response is the third component in the measurement path. In quantitative social research, a response generally represents a respondent's answer to a particular survey question. However, sometimes respondents fail to provide accurate responses, which results in systematic

[44] For a detailed discussion of the representation path, see Groves et al. (2004).
[45] Most frequently, the survey literature refers to three different kinds of validity (Knapp, 2008a): content validity, criterion validity, and construct validity.

measurement error (second quality concept).[46] The fourth and final component in the measurement path is edited response, which is accompanied by processing error (third quality concept), such as poor data entering and/or coding (Biemer & Lyberg, 2003). Other sources of error at this stage include data adaptations undertaken due to (seemingly) inconsistent responses and comparatively low or high (response) values.[47] Processing error only occurs after data collection and prior to the computation of survey statistics.

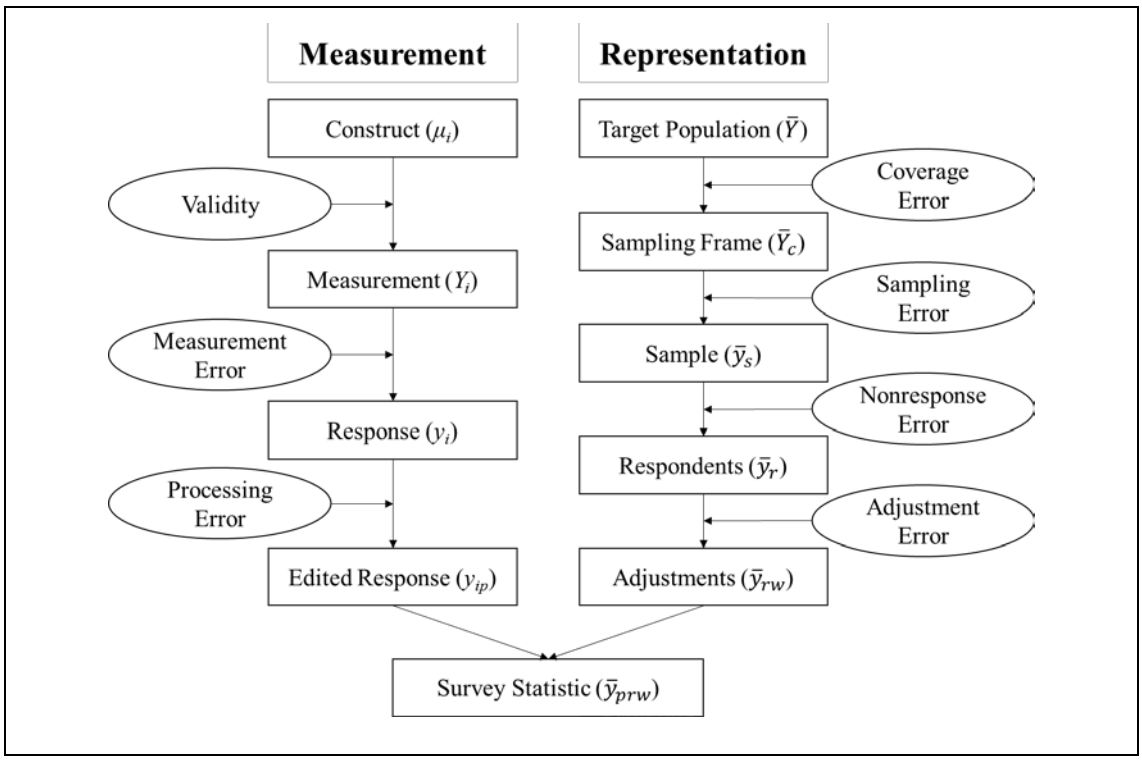

Figure 1. Survey lifecycle from a quality perspective (Groves et al., 2004)

Note. Components are in rectangles and quality concepts are in ovals. For the notations, see Groves et al. (2004).

Measurement error is one of the most serious source of error in surveys and is, therefore, frequently discussed by survey researchers (see Alwin, 2007; Billiet & McClendon, 1998,

[46] Measurement error must be distinguished from response bias. Response bias refers to a systematic distortion of the cognitive response process, due to the design of survey instruments and/or interview settings (Groves et al. 2004). This, in turn, results in systematic measurement error.

[47] One example are outlier definitions in response time analyses (see Höhne & Schlosser, 2017).

2000; DeCastellarnau, 2017; Groves, 2004; Hox, 2008; Revilla & Ochoa, 2015; Revilla & Saris, 2015; Saris & Gallhofer, 2014; Saris et al., 2010). However, measurement error is a relatively broad concept, which subsumes all kinds of errors that can occur due to the respondent, the interviewer, and the questionnaire. For instance, respondents may intentionally or unintentionally provide incorrect answers, interviewers may inappropriately influence respondents, or questionnaires may contain confusing instructions or poorly worded questions (Biemer & Lyberg, 2003). Due to the methodological complexity of the term measurement error, it is important to clearly define it to understand how it is employed in empirical studies.

In this thesis, response bias is considered as an indicator of systematic measurement error or its complement response quality (i.e., the accuracy of respondents' answers).[48] DeCastellarnau (2017), for instance, uses a very similar definition of measurement error. One explanation for the occurrence of response bias in survey responding was proposed by Krosnick (1991) and his satisficing theory, which primarily differentiates between optimizing and satisficing.[49] There are two types of satisficing: weak and strong. "It is useful to think of optimizing and strong satisficing as the two ends of a continuum indicating

[48] Classical test theory assumes that a measure or response Y can be decomposed as follows: $Y = \mu + \varepsilon$. Accordingly, a respondent's answer is based on the sum of μ (true value) and ε (error). As stated, response bias is an indicator of systematic measurement error. Meanwhile, ε only includes random error. Systematic error is ascribed to μ and, thus, cannot easily be separated from it (Bühner, 2011). Therefore, several authors proposed conceptual adaptations to the formula. For instance, Lischewski (2015) suggests extending the formula as follows: $Y = \mu + \varepsilon_s + \varepsilon_r$. In this case, ε is divided into systematic (ε_s) and random (ε_r) measurement error. However, this definition is not in line with the rationale of the classical test theory since it represents a deviation from its origin. Regarding response bias, this thesis only investigates the following question: In which question format is the occurrence of response bias more distinct? Thus, no quantification of the systematic measurement error can be provided. A detailed discussion of the classical test theory and its assumptions is beyond the scope of this chapter. So is a discussion of the existence of a true value, particularly with respect to attitudes. Therefore, interested readers are referred to Lord and Novick (1968) for a comprehensive discussion of the classical test theory and Bradburn, Sudman, and Wansink (2004) as well as Wilson, LaFleur, and Anderson (1996) for a comprehensive discussion of the notion of the true value.

[49] The terms optimizing and satisficing were actually established by Simon (1957). He used the terms to describe economic decision-making and choice behavior (Krosnick, 1991, p. 215).

the degrees of thoroughness with which the cognitive response processes are performed" (Krosnick & Presser, 2010, p. 266). The optimizing end implies an effortful and thorough response process that produces an accurate response. The (strong) satisficing end, by contrast, implies a superficial or incomplete response process that causes response bias. Hence, satisficing response behavior increases systematic measurement error and, thus, reduces response quality.[50]

In the original theory, Krosnick (1991) proposed two forms of weak satisficing and four forms of strong satisficing. The weak forms of satisficing are (1) selecting the first response alternative that seems reasonable (see Study IV and V) and (2) agreeing with assertions. The strong forms of satisficing are (1) endorsing the status quo, (2) non-differentiation in using rating scales (see Study III and IV), (3) saying don't know, and (4) and mental coin flipping. The survey literature addresses a wide range of forms of response bias that can be subsumed under the notion of satisficing. The forms of response bias related to this thesis – see Study III and IV – are as follows: Item nonresponse (Galesic & Bosnjak, 2009; Höhne, Revilla, & Lenzner, 2018; Lenzner, Kaczmirek, & Lenzner, 2010; Revilla & Couper, 2018a), dropouts (Chang & Krosnick, 2009; Höhne, Schlosser, & Krebs, 2017), and speeding (Conrad et al., 2017; Galesic & Bosnjak, 2009; Höhne, Revilla, & Lenzner, 2018; Höhne, Schlosser, & Krebs, 2017; Lenzner, Kaczmirek, & Lenzner, 2010).[51] All these forms of response bias decrease response quality.

[50] The likelihood of satisficing depends on task difficulty, respondent ability, and respondent motivation. A detailed discussion of the satisficing theory and these three factors is beyond the scope of this chapter. Interested readers are referred to Anand (2008) or Krosnick (1991).

[51] A detailed discussion of each form of response bias is beyond the scope of this chapter. Therefore, see the corresponding studies or the literature cited in the text.

As suggested by Converse and Presser (1986) and Fowler (1995), A/D questions are susceptible to response bias. As argued in chapter 2.2 “Response Effort”, A/D questions foster a satisficing response behavior because they support a less engaged response effort. In particular, A/D questions’ indirect and invariant manner of asking can be seen as responsible factor (Höhne & Krebs, 2018; Höhne & Lenzner, 2017; Höhne, Revilla, & Lenzner, 2018; Höhne, Schlosser, & Krebs, 2017). IS questions, in contrast, are based on a direct and normally alternating manner of asking. It is to assume that they foster an optimizing response behavior because they support a more engaged response effort. As a result, the IS question format is more robust against response bias than the A/D question format. In other words, the occurrence of response bias should be more pronounced in A/D questions than in IS questions.

3. General Research Questions

As suggested in the previous chapter and its three subchapters, the A/D and IS question format do not only differ regarding their formal design but also regarding response effort and response quality. This chapter outlines the current gap of knowledge in the survey literature as well as the research questions addressed in this thesis. It then describes and discusses the conceptual specifications, research designs, and characteristics of the individual studies (Study I to V).

Most of the pertinent survey literature on A/D and IS questions revolves two thematic areas: Response effort (see Carpenter & Just, 1975; Fowler, 1995; Fowler & Cosenza, 2008; Krosnick & Presser, 2010; Pasek & Krosnick, 2010; Saris & Gallhofer, 2014; Saris et al., 2010) and response quality. Response quality can be subdivided into (1) reliability and validity of responses (see Hanson, 2015; Kuru & Pasek, 2016; Lelkes & Weiss, 2015; Revilla & Ochoa, 2015; Saris & Gallhofer, 2014; Saris et al., 2010) and (2) susceptibility to response bias (see Kuru & Pasek, 2016; Liu, Lee, & Conrad, 2015).[52] Some of these contributions address several thematic areas simultaneously. Table 1 contains a systematic categorization of these contributions.

It is frequently argued that A/D questions are more cognitively demanding than IS questions because the former force respondents to accomplish a complex and intricate response process (see Carpenter & Just, 1975; Fowler, 1995; Fowler & Cosenza, 2008; Krosnick & Presser, 2010; Pasek & Krosnick, 2010; Saris & Gallhofer, 2014; Saris et al., 2010). However, taking a closer look at the contributions dealing with the response effort involved in answering A/D and IS questions, it is observable that these studies are usually

[52] At this juncture, it must be stated that the literature search does not claim completeness. Hence, it might be possible that some relevant contributions were not included. Furthermore, some studies were not taken into account because they did not directly compare A/D and IS questions.

based on theoretical considerations and fall short of empirically testing the claims they make on response effort. For instance, Fowler (1995) and Fowler and Cosenza (2008) only discuss and describe the built-in cognitive complexity of the A/D question format – compared to the IS question format – without empirically testing such complexity. In addition, the existing contributions do consider the possibility that there might be a difference between the theoretically expected and actually expended response effort, a distinction that was suggested in chapter 2.2 "Response Effort". Hence, the current state of research is characterized by a lack of empirical studies that systematically investigate response effort.

The majority of the limited number of existing empirical studies investigated the reliability and validity associated with A/D and IS questions. For instance, using a "multitrait-multimethod (MTMM)" approach, Revilla and Ochoa (2015), Saris and Gallhofer (2014), and Saris et al. (2010) found that the IS question format is associated with higher reliability and validity than the A/D question format.[53] In accordance with these findings, Kuru and Pasek (2016) reported higher criterion validity and Hanson (2015) reported higher test-retest reliability for IS questions. However, Lelkes and Weiss (2015) found no significant differences between both question formats regarding test-retest reliability and criterion validity. Overall, there is evidence indicating the superiority of the IS question format over the A/D question format in terms of reliability and validity.

A limited number of studies investigated the response quality of A/D and IS questions in terms of response bias. For instance, investigating the occurrence of acquiescence response bias in A/D question compared to IS questions, Kuru and Pasek (2016) found

[53] Saris and colleagues contributed to a systematic examination of A/D and IS questions in terms of reliability and validity. For instance, the open-source tool "Survey Quality Predictor (SQP)" that they developed provides an extensive database of quality estimates for a wide range of survey questions concerning a diversity of topics in many different forms and languages (see Survey Quality Predictor, 2017).

strong evidence for acquiescence in the A/D questions.[54] Liu, Lee, and Conrad (2015) compared both question formats with respect to extreme response style – i.e., the tendency to select the highest or lowest response category of a rating scale (van Vaerenbergh & Thomas, 2013). The authors found that this form of response bias exists in both A/D and IS questions and only slightly differs between them. In general, the findings reported in the survey literature indicate that both question formats are prone to response bias.

Table 1. Categorization of scientific contributions on A/D and IS questions

Contributions	Response Effort*	Reliability	Validity	Response Bias
Carpenter & Just (1975)	✓			
Fowler (1995)	✓			
Fowler & Cosenza (2008)	✓			
Hanson (2015)		✓		
Krosnick & Presser (2010)	✓			
Kuru & Pasek (2016)			✓	✓
Lelkes & Weiss (2015)		✓	✓	
Liu, Lee, & Conrad (2015)				✓
Pasek & Krosnick (2010)	✓			
Revilla & Ochoa (2015)		✓	✓	
Saris & Gallhofer (2014)	✓	✓	✓	
Saris et al. (2010)	✓	✓	✓	

Note. *These contributions address response effort in a theoretical way without testing it empirically.

Table 1 also reveals that there is a comparatively large number of studies that investigate the reliability and validity of both question formats. Five out of six studies reported evidence that the IS question format has higher reliability and validity than the A/D question

[54] Acquiescence can influence all questions containing confirming utterances in the response categories, such as strongly agree (Holbrook, 2008, p. 3). In general, the use of IS questions seems to be a promising way to avoid acquiescence because the response categories express the content dimension and, thus, usually do not contain confirming utterances.

format (see Hanson, 2015; Kuru & Pasek, 2016; Revilla & Ochoa, 2015; Saris & Gallhofer, 2014; Saris et al., 2010). By contrast, only two studies investigated the susceptibility to response bias and they reported inconclusive results (see Kuru & Pasek, 2016; Liu, Lee, & Conrad, 2015). There is also a comparatively large number of contributions dealing with cognitive processing and response effort but none of the existing seven contributions investigated these concepts empirically (see Carpenter & Just, 1975; Fowler, 1995; Fowler & Cosenza, 2008; Krosnick & Presser, 2010; Pasek & Krosnick, 2010; Saris & Gallhofer, 2014; Saris et al., 2010).

In conclusion, the current state of research is unbalanced. On the one hand, the existing studies are characterized by a systematic examination of the reliability and validity of A/D and IS question formats. This feat can be attributed, in large part, to the work done by Saris and colleagues. On the other hand, the existing literature suffers a shortage of empirical studies in two specific research areas. The first area is response effort. The second area is susceptibility to response bias. The studies by Kuru and Pasek (2016) and Liu, Lee, and Conrad (2015) pave the way for further studies in this area. However, further systematic research on different types of response bias is necessary because response bias remains an under-researched area. And since there are many forms of response bias, it is an area rife with research potential.[55]

Based on the earlier characterization of the manner of asking and its theoretical conjecture – A/D questions are characterized by a superficial response process and IS questions by an active and thorough one – the following two research questions are addressed in this thesis:

[55] This thesis does not answer all open questions regarding A/D and IS questions. The main goal is to decrease the current research gap and to provide prospects for future research.

- Does the IS question format require a higher response effort than the A/D question format?
- Is the IS question format more robust against response bias than the A/D question format?

4. Methodological Considerations

Having provided the theoretical background of A/D and IS questions and stated the general research questions, it is time to outline the conceptual specification of response effort and response quality and discuss appropriate research designs for the empirical studies. At the end of this chapter, a discussion of the characteristics of the studies conducted takes place.

4.1 Conceptual Specifications

Response effort: In order to investigate the respective response effort associated with the A/D and IS question formats, different methods and techniques can be employed. One way to assess response effort is to simply ask respondents to evaluate how they perceived the responding. The semantic differential (Osgood, Suci, & Tannenbaum, 1957) represents a possible method. Based on opposite adjective pairs, the semantic differential enables the measurement of a respondent's unique perception of an object (Stoutenborough, 2008, p. 810) – e.g., of a question either presented in the A/D or the IS format. In this thesis, five pairs of opposite adjectives are employed: Two pairs measure task-oriented manageability and three pairs measure motivational potential. It is assumed that the higher the perceived task-oriented manageability and motivational potential are, the higher the respondent's evaluations of these dimensions must be (see Study V).

The technology-driven advancements in web-survey research allowed – in addition to survey responses – the collection of so-called paradata that describe the way respondents process survey questions. Couper (2000, p. 393) defines paradata as data that arise during web-based surveys and can be used to observe and investigate response behavior.[56]

[56] Paradata are classified into two types (Heerwegh, 2003, 2011): (1) Server-side paradata refer to information collected on the server hosting the web survey (e.g., device information). (2) Client-side paradata refer to information collected on the respondent's device (e.g., response times). "To gather information on

Thus, the collection of different kinds of paradata, such as response times and mouse clicks[57], allows studying response effort. Response times, for instance, have a long tradition in psychology and survey research because they allow researchers to draw conclusions about cognitive information processing (see Olson & Parkhurst, 2013). Bassili and Scott (1996), for instance, state that the processing time of a survey question corresponds to the response effort that is needed to answer it. In other words, it is to assume that the longer a respondent needs to answer a question, the higher the response effort must be (see Study III and IV).

Mouse clicks – another type of paradata – can be used to determine the number of answer changes (i.e., selecting a new response category after one has already been selected). According to Heerwegh (2011) and Stern (2008), answer changes occur due to uncertainty in selecting a response category, which indicates a intense response process. As with response times, it is assumed that the higher the number of answer changes is, the higher the response effort must be (see Study III and IV).

In addition to paradata, eye-tracking methodology enables researchers to unobtrusively investigate response behavior and to draw conclusions about response effort. While reading questionnaire instructions, question stems, and response categories, respondents' eye movements are captured by infrared cameras. Among other things, these cameras record the exact duration of fixations and the number of fixations and re-fixations (see Galesic et al., 2008; Galesic & Yan, 2011; Höhne & Lenzner, 2015). According to Geise (2011), these three eye-tracking parameters reflect the intensity of information processing

respondent behavior at the level of a specific survey question, one needs to turn to client-side paradata" (Heerwegh, 2003, p. 361). Hence, this thesis uses client-side paradata.

[57] The term "mouse clicks" is somewhat confusing with respect to smartphones and other mobile devices since these devices usually do not use a mouse as a navigation tool. The same applies to laptop PCs, which usually contain touchpads. For the sake of convenience, however, all forms of clicks, regardless of the navigation tool and device, are called "mouse clicks".

and inform about the allocation of attention. Thus, it is assumed that the longer the fixation time and the more fixations and re-fixations occur, the higher the response effort must be (see Study I and II).

Response quality: As described in chapter 2.3 “Response Quality”, response bias is an indicator of systematic measurement error or its complement response quality. In order to measure response quality, several indicators of satisficing response behavior can be used. In this thesis, five indicators are employed: Response order effects (i.e., higher endorsement of response categories appearing at the beginning of the response scale), nondifferentiation (i.e., rating several survey questions identically), item nonresponse (i.e., survey questions with missing data), dropouts (i.e., premature break off from the survey), and speeding (i.e., extremely fast responding). It is assumed that the more pronounced the occurrence of these forms of response bias is, the lower the response quality (see Study III, IV, V).

4.2 General Research Designs

A research design can be seen as the plan of a study, which means that research designs largely depend on the research question and hypotheses at hand. Research designs can be generally divided into quantitative, qualitative, and mixed-method research designs.[58] Quantitative research designs can be further divided into non-experimental, quasi-experimental, and experimental research designs (Shadish, Cook, & Campbell, 2002). These three types of quantitative research designs differ with respect to their ability to determine causality (Kalaian, 2008; Lavrakas, 2008b; Nunes Silva, 2008; Shadish, Cook, & Campbell, 2002), which is of particular importance for this thesis.[59] While non-experimental research designs prohibit researchers from making internally valid conclusions about cause-effect relationships, experimental research designs principally enable researchers to make internally valid conclusions about cause-effect relationships.

Shadish, Cook, and Campbell (2002, p. 18), for instance, state that non-experimental designs "refer to situations in which a presumed cause and effect are identified and measured but in which other structural features of experiments are missing". This category of research designs subsumes the following types of studies (Kalaian, 2008): (1) Correlational studies (i.e., determining the relationship between two or more variables to make predictions based on these relationships). (2) Causal-comparative studies (i.e., exploring existing differences among groups to determine possible causes or reasons for the differences). (3) Meta-analyses (i.e., quantitatively and systematically summarizing empirical findings of primary studies that address identical research questions). Unlike experimental designs, non-experimental designs do not explicitly control for an independent

[58] In the context of this thesis, only quantitative research designs are of interest. For a comprehensive discussion of qualitative and mixed-method research designs, see Creswell and Plano Clark (2007).
[59] For a comprehensive introduction to causality, see Russo (2009) or Shadish, Cook, and Campbell (2002).

variable to investigate its impact on a dependent variable by means of random assignment.[60]

Similar to non-experimental research designs, quasi-experimental research designs lack random assignment (Shadish, Cook, & Campbell, 2002). The assignment of cases or participants to different treatments is carried out through self-selection (i.e., respondents select the experimental condition for themselves) or researchers-driven selection (i.e., researchers match respondents in terms of specific characteristics, such as age and gender). Longitudinal research designs with repeated measurements can be considered a type of quasi-experimental research designs (see Kalaian, 2008). Typically, researchers analyze the measurements before an event (at a time point called t_0) and after the occurrence of an event (at a time point called t_1). Quasi-experimental research designs are usually based on probabilistically nonequivalent groups or repeated measurements and, thus, differences between groups or time points cannot be attributed with certainty to the treatment.

In an experimental research design, the researcher intentionally manipulates at least one independent variable and randomly assigns participants to different treatment conditions (i.e., cases or participants are divided into experimental groups). Random assignment generates groups of units that (on average) do not differ probabilistically from one another, except with respect to the treatment, so that outcome differences between those groups are likely to be due to the treatment (Shadish, Cook, & Campbell, 2002, p. 13). In other words, environmental or extraneous variables are controlled through random as-

[60] Random assignment refers to the circumstance that respondents are allocated to different conditions (e.g., treatment and control group) in such a way that each respondent has an identical chance to be allocated to one of the conditions (Knapp, 2008b, p. 674).

signment, which enables researchers to measure cause-effect relationships between variables (Kalaian, 2008; Knapp, 2008b; Nunes Silva, 2008; Shadish, Cook, & Campbell, 2002; Ziniel, 2008).[61]

It is also important to distinguish between field experiments (i.e., experiments where cases or participants stay in their familiar surroundings, such as at home or in their offices) and lab experiments (i.e., experiments where cases or participants are invited to a laboratory). Both types are employed in this thesis.[62] While field experiments are characterized by a high level of external validity (i.e., the extent to which the results are generalizable), lab experiments are characterized by a high level of internal validity (i.e., the extent to which the research design provides evidence of the existence of a causal relationship).[63]

Because they allow for the establishments of causality, experiments are seen the "gold standard" for investigating causal relationships, including survey research. For instance, experiments can be used to investigate the impact of different questionnaire design strategies (e.g., using a specific question format) on response behavior. Because experiments enable researchers to determine treatment effects, all empirical studies conducted in this thesis are based on experimental research designs. While Study I and II are based on lab experiments, Study III and IV are based on field experiments. Study V is based on

[61] There are many ways to employ experimental research designs, which mostly differ with respect to their complexity. The simplest form is the split-half experiment (i.e., randomized two-group posttest-only design).

[62] The term field experiment was established to have a conceptual contrast with the term lab experiment. In both cases, participants (usually) do not know that they take part in an experiment, which is called a blind test. Taking a closer look at the literature, it is observable that field experiments are the predominant type of experiments in survey research.

[63] A detailed discussion of external and internal validity is beyond the scope of this chapter. Therefore, interested readers are referred to further literature (see Kalaian & Kasim, 2008; Lavrakas, 2008b).

a field experiment but has a quasi-experimental design, where the question format is randomized but the response scale direction is not (for a detailed description of the research design, see Study V).

4.3 Characteristics of the Studies

All of the empirical studies conducted in this thesis are based on non-probability samples that mainly used a web-based, self-administered survey mode relying on PCs and/or smartphones. The study that deviates from this arrangement, Study V, is based on a self-administered paper-pencil survey mode. Thus, all studies employed visual communication channels.[64] While Study I and II are lab experiments that include incentives, Study III and V are field experiments with no incentives. Study IV is a field experiment that includes incentives. All studies were conducted in German, except Study IV, which was conducted in Spanish. The topics of the questions vary across the studies, except for Study III and V, both of which address the same topic. While all studies investigate response effort (using different methods and techniques), only Study III, IV, and V investigate response quality. Table 2 summarizes the characteristics of the studies.

The response scales employed in the studies differ with respect to their presentation (i.e., grid and single presentation mode), labeling (i.e., fully and end labeled), numbering (i.e., with and without numeric values), and the layout of single questions (i.e., horizontal and vertical arrangement of the response categories). However, the studies employ the same scale directions (decremental) and the same numbers of scale points (5-points). Study V, which investigates response order effects, employs decremental and incremental scale directions and Study IV, which compares the two question formats across PCs and smartphones, tested 5-point and 7-point response scales.

[64] Web-based surveys or interviews can include not only text (visual) but also tones (audio) and videos (visual or audio-visual). Thus, the communication channels can vary within modes (Couper, 2005, 2011). For a more detailed discussion of communication channels, see Jans (2008).

Table 2. Methodological characteristics of the studies (alphabetical order)

Characteristics	Study I	Study II	Study III	Study IV	Study V
Administra-tion	Self-administered	Self-administered	Self-administered	Self-administered	Self-administered
Device	PC	PC	PC	PC & mobile	No device
Experiment	Lab	Lab	Field	Field	Field
Incentives	Yes	Yes	No	Yes	No
Language	German	German	German	Spanish	German
Mode	Web	Web	Web	Web	Paper-pencil
Question content	Health issues	Political issues	Motivational aspects	Personality traits	Motivational aspects
Response					
Effort	Eye tracking	Eye tracking	Paradata	Paradata	Evaluations
Quality	No	No	Satisficing	Satisficing	Satisficing
Sample	Non-probability	Non-probability	Non-probability	Non-probability	Non-probability
Scale					
Direction	Decremental	Decremental	Decremental	Decremental	Decremental & incremental
Labeling	Fully labeled	Fully labeled	End labeled	End labeled	Fully labeled
Numbering	No	No	No	Yes	No
Points	Five	Five	Five	Five & seven	Five
Presentation	Grid (A/D) & single (IS)	Single only	Grid & single	Single only	Grid & single
Layout single questions	Horizontal	Vertical	Horizontal	Vertical	Horizontal

The five empirical studies have some design and layout differences. First, they were conducted at different times. While Study I and II were already conducted in 2012, Study IV was conducted in 2016. Second, the study design in general and the question and response scale design in particular were partially dependent on research collaborations. For instance, the data collection of Study IV was sponsored by the access panel Netquest. Because Netquest planned to take practical decisions based on the empirical results of the study, they chose to have questions revolve around their topic of interest, namely personality traits. Another example is Study V, which had to be based on paper-pencil because the viable methods of data collection were limited. However, such methodological differences between the empirical studies do not represent an imperfection per se. If the studies

arrive at similar results, this would indicate that the differences between A/D and IS questions do not depend on specific design decisions, such as question content and/or scale labeling.

5. Empirical Studies

This part of the thesis contains five empirical studies that were conducted to investigate the respective response effort and response quality associated with the A/D and IS question formats. Study I and II are an in-depth investigation of response effort using eye-tracking methodology. Study III and IV address both response effort and response quality by means of paradata across different device types. Finally, Study V investigates the response quality associated with the two question formats in terms of their susceptibility to response order effects.

5.1 Study I[65]

This eye-tracking study is a web-based lab experiment conducted with two conditions. The study investigates the cognitive processing of A/D and IS questions. By recording respondents' eye movements, it is possible to evaluate how respondents process information in surveys, including question stems and response categories.

5.1.1 Research Hypotheses

As argued in the theoretical background chapters, the A/D question format seems to promote superficial rather than optimal responding, due to the indirect and unchanging manner of asking. This should reflect itself in respondents' gaze behavior and, thus, in the eye-tracking data; in particular, in a comparatively low fixation number (i.e., the total count of fixations on a region of interest) and a comparatively short fixation time (i.e., the total duration of fixations on a region of interest). The IS question format, by contrast, seems to promote active and intensive rather than superficial responding, due to the direct

[65] This chapter is based on:
Höhne, J.K. (2018). Eye-tracking methodology: Exploring the processing of question formats in web surveys. *International Journal of Social Research Methodology*. DOI: 10.1080/13645579.2018.1515533

and changing manner of asking. This should also reflect itself in the eye-tracking data; in particular, in a comparatively high fixation number and a comparatively long fixation time. Previous research has demonstrated that these two eye-tracking parameters are good indicators of effort in survey responding (Galesic et al., 2008; Galesic & Yan, 2011; Höhne & Lenzner, 2015, 2017; Kamoen et al., 2011; Lenzner, Kaczmirek, & Galesic, 2011, 2014; Menold et al., 2014).

This argumentation is based on two conjectures about the relationship between eye-tracking data and mental processes (Just & Carpenter, 1980, p. 330): (1) The immediacy assumption states that the processing of objects that are fixated is not deferred because it occurs as soon as possible. (2) The eye-mind assumption states that a considerable delay between the fixation of an object and its processing does not occur. Hence, it can be assumed that the fixation number and time for an object are similar to the fixation number and time required for processing it.

Two research hypotheses are postulated: On the one hand, it is hypothesized that respondents fixate more frequently and longer on the question stems when responding to IS questions than when responding to A/D questions (Hypothesis 1). On the other hand, it is hypothesized that respondents fixate more frequently and longer on the response categories when responding to IS questions than when responding to A/D questions (Hypothesis 2).

5.1.2 Method

5.1.2.1 Research Design

An eye-tracking experiment was conducted to investigate the processing of A/D and IS questions. Participants were randomly assigned to one of two groups: The first group (n = 43) received three A/D questions in a grid (agree/disagree condition). The second group

(n = 41) received three individual IS questions presented on the same page (item-specific condition).

5.1.2.2 Survey Questions

The three survey questions were adapted from existing social surveys. In each case, both an A/D and IS counterpart that preserved question content as much as possible were developed (see Appendix A for the questions used).[66] The questions were designed in German, which was the mother tongue of 93% of the participants. All questions were presented with 5-point, completely verbalized response scales. The A/D questions were presented in a grid with horizontally arranged response categories alongside each question, which is the predominant way of employing them (see Couper et al., 2013; Saris et al., 2010). The IS questions, by contrast, were presented individually with horizontally arranged response categories below each question. All questions were displayed on the same page with black text on a white background (see Figure 2).

5.1.2.3 Sample

84 participants took part in the experiment. Due to technical difficulties, the eye movements of two participants could not be recorded accurately. The recorded eye fixations of six other participants were not satisfactory because there was a systematic shift in the eye-tracking recordings. Consequently, these participants were excluded from the data, leaving 76 in the statistical analyses. Participants were between 17 and 76 years old with a mean age of 35.9 (SD = 14.5). 54% of the participants were female. 22% had graduated from a lower secondary school, 11% from an intermediate secondary school, and 67%

[66] Strictly speaking, the A/D response scales used are based on unipolar scales, which is the most common way to ask A/D questions in German. In his pioneering study, Rohrmann (1978) can additionally show that both types of the German A/D scales (unipolar and bipolar) do not differ regarding equidistance.

from a college preparatory secondary school or university. The great majority used a computer and the Internet every day or almost every day (88% and 87%, respectively). 80% had participated in at least one web survey prior to this study.

To evaluate the sample composition between the two experimental groups, several chi-square tests were conducted. There are no significant differences regarding age [$\chi^2(2) = 2.74$, $p = .25$], gender [$\chi^2(1) = .00$, $p = .99$], education [$\chi^2(2) = 1.60$, $p = .45$], computer usage [$\chi^2(1) = .96$, $p = .33$], Internet usage [$\chi^2(1) = 1.61$, $p = .21$], and survey experience [$\chi^2(1) = .56$, $p = .45$].

5.1.2.4 Eye-Tracking Equipment

Participants' eye movements were recorded by a Tobii T120 Eye Tracker and the data were analyzed with the Tobii Studio 3.2.1 software. The Tobii T120 is a remote eye tracker embedded in a 17'' TFT monitor (resolution 1280 × 1024), with two binocular infrared cameras located underneath the computer screen. The system is accurate within 0.5° with less than 0.3° drift over time and permits head movements within a range of 30 × 22 × 30 cm. Eye movements were recorded at a sampling rate of 120 Hz. The online questionnaire was programmed with a font size of 18 and 16 pixels and double-spaced text with a line height of 40 and 32 pixels for the question text and response categories, respectively. Before analyzing the eye-tracking data, we applied Tobii Studio's I-VT fixation filter in the default setting (gap fill-in: enabled, 75 ms; eye selection: average; noise reduction: disabled; velocity calculator window length: 20 ms; I-VT classifier: 30°/s;

merge adjacent fixations: enabled, max time between fixations: 75 ms, max. angle between fixations: 0.5°; discard short fixations: enabled, minimum fixation duration: 60 ms) to identify "true" fixations in the raw data.[67]

5.1.2.5 Procedures

This experiment was conducted at GESIS – Leibniz-Institute for the Social Sciences in 2012 and was part of a larger study including cognitive interviewing and several independently randomized eye-tracking experiments (see Höhne & Lenzner, 2015, 2017; Lenzner et al., 2014; Menold et al., 2014). Participants completed this experiment after taking part in a cognitive interview.

Participants were seated in front of the eye-tracking system and completed a standardized calibration procedure (i.e., following a moving dot on the screen with the eyes). After a successful calibration, they began the web survey. The entire experiment was executed and supervised by an experimenter who also observed respondents' eye movements on a computer screen in real-time. Participants were asked to read at a normal pace and to try to understand the questions as well as possible. At the very beginning of the web survey, two questions were asked to determine the individual fixation and reading rate of respondents (see Appendix A for the questions used).[68] Both parameters served as covariates in the subsequent analyses. Completing the web survey lasted approximately 12 min and participants received a compensation of €30 for taking part in the entire study.

[67] The Tobii I-VT filter is an update of older fixation filters and allows for more sophisticated data cleaning. The default values were selected to provide the best possible fixation classification across recordings with different levels of noise. Detailed descriptions of the general principles behind the I-VT fixation filter can be found in Tobii Technology (2012).

[68] Whereas fixation rate is defined as the mean number of fixations on these questions, reading rate is defined as the mean time of fixation on these questions.

5.1.3 Results

Question stems and response categories of the A/D and IS questions differed in the number of words. This could only have been avoided by developing artificial questions. To take length differences into account, fixation count and time of all question stems and response categories, respectively, were corrected for length differences by dividing them by the number of characters (see Ferreira & Clifton, 1986).

In order to evaluate the cognitive information processing of the A/D and IS question format, general linear models for the question stems and response categories were calculated. In addition, fixation rate and reading rate were employed as covariates to control for inter-individual differences.

The analyses were conducted for the three aggregated questions. This strategy was adopted because the A/D grid questions could not simply be separated. It also makes it possible to reduce the statistical tests and to efficiently summarize the results.

5.1.3.1 Fixation Count and Fixation Time

In accordance with the research hypotheses, the analysis focused on whether the three IS questions cause more and longer fixations than the three A/D questions. Table 3 reveals that this is only partially supported by the eye-tracking data, because there are differences with respect to the question stems and response categories. For the question stems, the fixation number [$F(1,73) = 0.81$, $p < 0.37$, partial $\eta^2 = 0.01$] and fixation time [$F(1,73) = 0.13$, $p < 0.72$, partial $\eta^2 = 0.00$] do not significantly differ between the A/D and IS question formats. This indicates that both stems are processed similarly. For the response categories, by contrast, the fixation number [$F(1,73) = 16.66$, $p < 0.001$, partial $\eta^2 = 0.19$] and fixation time [$F(1,73) = 7.05$, $p < 0.01$, partial $\eta^2 = 0.09$] differ significantly between

the two question formats. These results indicate that the IS response categories are processed more intensively than their A/D counterparts.

Table 3. Study I: Means and standard errors (in parentheses) of fixation number and time per character for question stems and response categories of A/D and IS questions

Eye Tracking Parameter	Question Part	Agree/Disagree (A/D)	Item-Specific (IS)
Fixation Count	Question Stems	.18	.20
		(.01)	(.01)
	Response Categories	.10	.15
		(.01)	(.01)
Fixation Time (sec)	Question Stems	.03	.03
		(.00)	(.00)
	Response Categories	.03	.04
		(.00)	(.00)

Note. The table reports estimated marginal means after controlling for the covariates fixation rate and reading rate, respectively. To control for length differences of question stems and response categories between the two question formats, the two eye-tracking parameters were divided by the number of characters (see Ferreira & Clifton, 1986).

In addition, the differences in fixation number and time between the question stems and response categories within the A/D and IS group were calculated, respectively, to analyze differences in the attention allocated to both question parts. On the basis of these differences in means, analyses of variance for fixation number and time between the group with A/D questions and the group with IS questions were conducted. The statistical results reveal marginally significant differences for fixation number [$F(1,74) = 3.43$, $p < .10$, partial $\eta^2 = .04$] and significant differences for fixation time [$F(1,74) = 5.00$, $p < .05$, partial $\eta^2 = .06$]. Hence, respondents seem to invest more effort when processing the response categories of IS questions than of A/D questions, compared to the respective question stems.

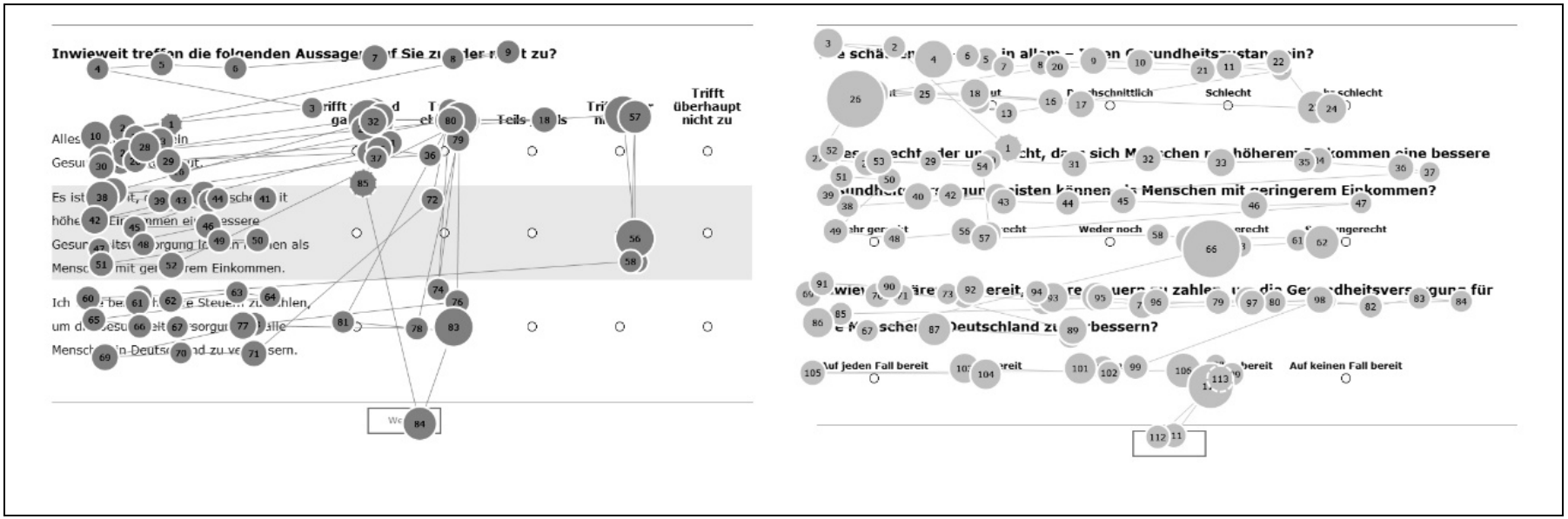

Figure 2. Study I: Gaze plots of two respondents for the three A/D and IS questions

Note. The gaze plot on the right side corresponds to the first experimental group (agree/disagree condition) and the gaze plot on the left side corresponds to the second experimental group (item-specific condition). The circles represent fixations and the lines between the circles represent saccades. The numbers in the center of the circles represent the sequence of the fixations and the size of the circles is proportional to the fixation time.

To explore the processing of A/D and IS questions, the scan paths of respondents were additionally inspected by means of gaze plots. Figure 2 contains two exemplary gaze plots from two respondents, one of whom answered the three A/D questions presented in a grid, whilst the other respondent answered the three IS questions presented on the same page. Gaze plots depict the order of respondents' eye movements across objects, such as question stems and response categories. The circles indicate fixations and the size of the circles is proportional to the fixation time. The lines indicate saccades (i.e., quick eye movements that initiate fixations).

Closer inspection of Figure 2 reveals that especially the categories at the beginning and/or the middle of the A/D and IS response scales are fixated most intensively. It is also apparent that both respondents do not fixate all response categories, irrespective of the question format. In fact, the response categories at the end of the scales are mostly overlooked. Whereas the respondent in the IS group processed the questions relatively sequentially (i.e., from the question stem to the response categories), the respondent in the A/D group showed relatively many re-fixations between the question stems and response categories.

5.1.4 Discussion and Conclusion

The main goal of this eye-tracking experiment was to explore how respondents process A/D and IS questions and to draw conclusions about the respective response effort associated with both question formats. In line with the study by Höhne and Lenzner (2017), the empirical findings indicate that there are no substantial differences with respect to the processing of the A/D and IS question stems. This can thus be considered as an indicator for equality in terms of question stem processing. A closer look at the questions used in this study reveals that they do not differ substantially in terms of semantic and/or syntactic

issues (see Appendix A for details about the questions). In fact, they only differ with respect to the form of wording: Whilst the A/D questions are worded in a declarative form (i.e., as indirect statements), the IS questions are worded in an interrogative form (i.e., as direct questions).

In contrast to the question stems, the A/D and IS response categories differ significantly in terms of fixation number and time. In particular, this indicates that the response categories of the IS questions are more intensively processed than the response categories of the A/D questions. This finding supports the notion of the manner of asking (see Höhne, Schlosser, & Krebs, 2017). Technically, the A/D question format demands complex and elaborated processing. However, its indirect and repetitive manner of asking seem to promote perfunctory responding, which manifests itself in a lower fixation number and time. The IS question format, by contrast, is characterized by a direct and varied manner of asking, which seems to promote intensive processing. This manifests itself in a higher fixation number and time.

Interestingly, the gaze plots show that respondents do not fixate all response categories and, thus, do not read all of them. In fact, they do not consider the response categories at the end of the scale, irrespective of the question format. It seems that respondents usually fixate the categories at the beginning and the center of horizontally arranged response scales (see Höhne & Lenzner, 2015). Consequently, this indicates that respondents seem to be able to mentally extrapolate the response continuum of A/D and IS questions. One explanation for this phenomenon might be that both question formats are based on rating scales. In other words, they follow an ordered and closed response continuum that, in principle, allows an extrapolation of subsequent response categories.

A further interesting point is that the gaze plots reveal that the IS questions apparently cause more re-fixations between the response categories than the A/D questions. This would additionally indicate a more intensive processing of the response categories (see Study II).

There are some limitations to this study. First of all, the eye-tracking experiment only explores the processing of A/D and IS questions without considering the quality of responses. Therefore, further research that investigates the respective response effort and the response quality associated with the A/D and IS question formats would be desirable. Second, this experimental study compared A/D questions employed in grid presentation mode with IS questions employed in single presentation mode. However, previous research has shown that questions employed in grids are frequently accompanied by several undesirable outcomes, such as superficial responding and low response quality (Couper et al., 2013). Furthermore, this study only used IS questions that change the manner of asking (i.e., addressing different content dimensions). Therefore, it would be interesting if future studies could also employ IS questions without changing the manner of asking (i.e., addressing the same content dimension, such as intensity or importance).

To conclude: This eye-tracking study provides empirical evidence for the notion of the manner of asking survey questions (see Höhne, Schlosser, & Krebs, 2017). It seems that an indirect and invariant manner of asking, as is the case with the A/D question format, promotes relatively superficial responding. In general, this suggests a difference between the theoretically presumed complexity of question formats and the actually effort expended in responding. The empirical findings also indicate that the IS question format seems to encourage respondents to engage in more thoughtful and deliberate responding to each question in the light of its specific content.

5.2 Study II[69]

This study is based on a lab experiment with two conditions. The study investigates the cognitive processing associated with A/D and IS questions in web surveys using eye-tracking methodology. On the basis of recordings of respondents' eye movements, it is possible to draw conclusions on how respondents process survey questions.

5.2.1 Research Hypotheses

If IS questions do indeed promote a more conscientious responding than A/D questions due to a more active and more intensive cognitive response process (see chapter 2.2 "Response Effort"), then this should manifest itself in the eye-tracking data in the form of higher fixation counts, longer fixation times, and more re-fixations. Fixation count is defined as the total number of fixations on a specific area of interest (e.g., the question stem) including re-readings. Fixation time is defined as the total duration of fixations on a specific area of interest (e.g., the question stem), again including re-readings. Re-fixations of response categories are defined as the total number of response categories that respondents re-fixate (i.e., fixate again after reading at least one other response category). These eye-tracking parameters – fixation count, fixation time, and re-fixations – have been proven as good indicators of effort in responding to survey questions (Galesic et al., 2008; Galesic & Yan, 2011; Höhne & Lenzner, 2015; Kamoen et al., 2011; Lenzner, Kaczmirek, & Galesic, 2011, 2014; Menold et al., 2014). Furthermore, Lenzner (2012) was able to show that several linguistic text features impeding question understanding, which were detected by means of two of the eye-tracking parameters mentioned above (fixation count and fixation time; Lenzner, Kaczmirek, & Galesic, 2011), negatively affect response

[69] This chapter is based on:
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.

quality (e.g., amount of non-substantive responses). Hence, these parameters seem to be good predictors of the quality of survey responses.

Our reasoning is based on two basic assumptions about the relationship between eye fixations and cognitive processing (Just & Carpenter, 1980, p. 330): (1) The immediacy assumption posits that words or objects that are fixated by the eyes are processed directly; their interpretation is not deferred. (2) The eye-mind assumption posits that the eyes remain fixated on a word or an object as long as it is being processed. Taken together, these assumptions postulate that there is a close connection between fixation behavior and mental processing: The fixation count and time spent on a word or an object is (more or less) equal to the count and time required for processing it. Adopting these two assumptions, we investigate whether answering IS questions is indeed characterized by higher fixation counts, longer fixation times, and more re-fixations, indicating a more deliberate response process.

Under optimal conditions, we would expect respondents to thoroughly carry out all six of the mental tasks described in chapter 2.1 “Question Format” when answering A/D questions. This, in turn, would show up in the eye-tracking data in the form of higher fixation counts, longer fixation times, and more re-fixations for the A/D format, because it is cognitively more complex than the IS format. However, assuming that this kind of optimal responding only occurs rarely and that the indirect and constant manner of asking associated with A/D questions promotes a superficial response process, we expect more fixations and re-fixations as well as longer fixations in the IS question format instead.

With respect to our reasoning, we postulate the following three hypotheses: Respondents fixate more frequently and longer on the question stems and the response categories when answering IS questions than when answering A/D questions (Hypothesis

1). Respondents read more response categories in the IS question format than in the A/D question format (Hypothesis 2). Respondents show more re-fixations between the response categories when answering IS questions compared to A/D questions, that is, they re-read response categories they have previously read (Hypothesis 3).

5.2.2 Method

5.2.2.1 Research Design

We conducted an eye-tracking experiment to investigate the processing of A/D and IS questions during web-survey completion. Respondents were randomly assigned to one of two experimental groups. The first group (n = 44) received three individual A/D questions with a 5-point response scale (agree/disagree condition). The second group (n = 40) received three individual IS questions with a 5-point response scale (item-specific condition).

5.2.2.2 Survey Questions

The three questions used were adapted from the European Social Survey (2008) and the International Social Survey Program (2004) and dealt with various political issues, such as political interest. For each IS question adapted from these surveys, we developed an A/D counterpart that preserved the question's content as much as possible. The questions were designed in German, which was the mother tongue of 93% of the participants. Respondents answered both A/D and IS questions on 5-point, fully-labeled response scales with a vertical arrangement of the response categories and no numerical labels (see Appendix B for details about the questions used).

5.2.2.3 Sample

In total, 84 participants took part in the experiment. Due to technical difficulties, the eye movements of two respondents could not be recorded accurately and the recorded eye fixations of seven were not satisfactory, because there was a systematic shift to the line below or above the one that was fixated. These participants were excluded from the data, leaving 75 in the analyses. The respondents were between 17 and 76 years old with a mean age of 35.7 (SD = 14.6), and 53% of them were female. 20% had graduated from a lower secondary school, 12% from an intermediate secondary school, and 68% from a college preparatory secondary school or university. The great majority used a computer and the Internet every day or almost every day (89% and 88%, respectively), and 81% had participated in at least one web survey prior to this study.

To evaluate the effectiveness of random assignment, we additionally calculated chi-square tests. The results showed no statistically significant differences between the two experimental groups with respect to the following socio-demographic characteristics: Age [$\chi^2(2) = 2.23$, $p = .33$], gender [$\chi^2(1) = 2.26$, $p = .13$], education [$\chi^2(2) = .76$, $p = .68$], computer usage [$\chi^2(1) = .26$, $p = .61$], Internet usage [$\chi^2(1) = .64$, $p = .42$], and survey experience [$\chi^2(1) = .59$, $p = .44$].

5.2.2.4 Eye-Tracking Equipment

Participants' eye movements were recorded by a Tobii T120 Eye Tracker and the data were analyzed with the Tobii Studio 3.2.1 software. The Tobii T120 is a remote eye tracker embedded in a 17'' TFT monitor (resolution 1280 × 1024) with two binocular infrared cameras placed underneath the computer screen. The system is accurate within 0.5° with less than 0.3° drift over time and permits head movements within a range of 30 × 22 × 30 cm. Eye movements were recorded at a sampling rate of 120 Hz. The online

questionnaire was programmed with a font size of 18 and 16 pixels and double-spaced text with a line height of 40 and 32 pixels for the question text and response categories, respectively. Before analyzing the eye-tracking data, we applied Tobii Studio's I-VT fixation filter in the default setting (gap fill-in: enabled, 75 ms; eye selection: average; noise reduction: disabled; velocity calculator window length: 20 ms; I-VT classifier: 30°/s; merge adjacent fixations: enabled, max time between fixations: 75 ms, max. angle between fixations: 0.5°; discard short fixations: enabled, minimum fixation duration: 60 ms) to identify "true" fixations in the raw data.[70] As a sensitivity check, we repeated the analyses of the fixation counts and times on the question stems and response categories using Tobii's ClearView fixation filter set to include only fixations that lasted at least 100 milliseconds and encompassed 20 pixels. The results were similar to the ones we obtained by applying the I-VT filter in the default setting and all of our conclusions remained unchanged.

5.2.2.5 Procedures

The study was conducted at GESIS – Leibniz Institute for the Social Sciences in Mannheim (Germany) in October and November of 2012 and was part of a larger study with several unrelated experiments (see Höhne, under review; Höhne & Lenzner, 2015; Lenzner, Kaczmirek, & Galesic, 2014). All experiments were independently randomized to reduce the possibility of any systematic carryover effects. The entire study lasted ap-

[70] The Tobii I-VT filter is an update of older fixation filters and allows for more sophisticated data cleaning. The default values were selected to provide the best fixation classification possible across recordings with different levels of noise. Detailed descriptions of the general principles behind the I-VT fixation filter can be found in Tobii Technology (2012).

proximately 90 min, 30 min of which were devoted to eye tracking and 60 min to cognitive interviewing. The present experiment was embedded in a web survey that participants completed after taking part in a cognitive interview during the second half of the study.

After participating in the cognitive interview, participants were seated in front of the eye tracker so that their eyes were approximately 60 cm from the computer screen. After completing a standardized calibration procedure (during which they were asked to follow a moving red dot on the screen with their eyes), they started the web survey. The calibration procedure was carried out by an experimenter who also oversaw the experiment from a separate observer room next to the laboratory. The experimenter monitored respondents' eye movements on a computer screen in real time. Respondents were instructed to read at a normal pace while trying to understand the questions as well as they could. Only one question at a time was displayed on the screen and the questions were written in black text against a white background. At the beginning, all participants answered the same two questions, which were used to calculate their individual fixation rate, reading rate, and re-fixation rate (these parameters were used as covariates in the statistical analyses to control for inter-individual differences).[71] The whole web survey took about 12 min to complete. For their participation in the whole study (including the cognitive interview), respondents received a compensation of €30.

[71] Fixation rate refers to the average number of fixations on these questions, reading rate refers to the average fixation time on these questions, and re-fixation rate refers to the average number of response categories that were re-fixated in these questions.

5.2.2.6 Analytical Strategies

The results of this experimental study will be reported as follows: We first look separately at the fixation counts and times for the question stems and response categories to investigate whether these two question parts affect the process of responding in different ways. Afterwards, we investigate the number of response categories read and re-fixated. Since there are no substantial differences on the question-level and to reduce the number of statistical tests and efficiently summarize the results, we conducted the analyses on the means of the three A/D and IS questions over all respondents.

Due to technical limitations the number of response categories read and the number of re-fixations could not be detected by the Tobii eye-tracking system so that the questions had to be coded by two coders, each of which coded the eye movements of one half of the respondents (n = 41). In addition, the eye movements of a randomly selected subset of 10% of the respondents (n = 8) were coded by both coders for the purpose of estimating reliability. Interrater agreement was excellent (Fleiss, Levin, & Paik, 2003), with an Intraclass Correlation Coefficient (ICC) of .95. Discrepancies between the two ratings were examined and discussed with the second author until consensus was reached.

The question stems and response categories of the A/D and IS questions differed in the number of words, since this was unavoidable without formulating artificial-sounding survey questions. According to Ferreira and Clifton (1986), we corrected for length differences of question stems and response categories between the two question formats (A/D vs. IS) by dividing fixation count, fixation time, and re-fixations by the number of characters. Hence, fixation count and time for the question stems and response categories as well as re-fixations for the response categories per character are reported in the results.

5.2.3 Results

In order to determine the response effort associated with A/D and IS questions, we calculated general linear models for the question stems and response categories and used fixation rate, reading rate, and re-fixation rate as covariates to control for inter-individual differences in respondents' reading and response behavior.

5.2.3.1 Fixation Count and Fixation Time

In line with Hypothesis 1, Table 4 shows that IS questions indeed cause (on average) a higher fixation number and longer fixation times than their A/D counterparts. However, there are substantial differences between question stems and response categories: While for question stems, we cannot find any mean differences between the two question formats, for response categories we observe large mean differences with respect to fixation count and time. For the question stems, there are no significant differences between the A/D and IS questions for fixation count [$F(1,72) = .03$, $p = .87$, partial $\eta^2 = .00$] and fixation time [$F(1,72) = .08$, $p = .78$, partial $\eta^2 = .00$]. In contrast, for the response categories, there are significant differences between the two question formats for fixation count [$F(1,72) = 21.49$, $p < .001$, partial $\eta^2 = .23$] and fixation time [$F(1,72) = 13.97$, $p < .001$, partial $\eta^2 = .16$].

We additionally calculated the differences of fixation count and fixation time between the question stems and response categories within A/D and IS questions, respectively, to investigate whether there are differences in the allocation of attention. Based on these mean differences, we subsequently calculated analyses of variance for fixation count and fixation time between the group that received A/D questions and the group that received IS questions. The results show significant differences for fixation count [$F(1,73) = 15.93$, $p < .001$, partial $\eta^2 = .18$] and fixation time [$F(1,73) = 11.50$, $p < .01$, partial η^2

= .14] between the two experimental groups for both question formats. Accordingly, respondents expend much more effort when processing the response categories of IS questions than of A/D questions compared to the respective question stems. This is also suggested by the estimated marginal means in Table 4.

Table 4. Study II: Means and standard errors (in parentheses) of fixation count and time per character for question stems and response categories of A/D and IS questions

Eye Tracking Parameter	Question Part	Agree/Disagree (A/D)	Item-Specific (IS)
Fixation Count	Question Stems	.22	.22
		(.01)	(.02)
	Response Categories	.16	.29
		(.02)	(.02)
Fixation Time (sec)	Question Stems	.04	.04
		(.00)	(.00)
	Response Categories	.05	.10
		(.01)	(.01)

Note. The table reports estimated marginal means after controlling for the covariates fixation rate and reading rate, respectively. To control for length differences of question stems and response categories between the two question formats, we divided the two eye-tracking parameters by the number of characters (Ferreira & Clifton, 1986).

5.2.3.2 Response Categories Read and Re-Fixated

With respect to Hypothesis 2, we compared the average number of response categories read in the IS and in the A/D question format. Table 5 shows that, contrary to our expectations, respondents do not read more response categories in the IS than in the A/D question format. An analysis of variance of the means of the three questions revealed no significant differences in the number of response categories read between the two experimental groups [$F(1,73) = .04$, $p = .85$, partial $\eta^2 = .00$].

Table 5. Study II: Means and standard errors (in parentheses) of response categories read and re-fixated (per character) in the A/D and IS question format

Eye Tracking Parameter	Agree/Disagree (A/D)	Item-Specific (IS)
No. of response categories read	3.10	3.06
	(.12)	(.13)
No. of response categories re-fixated	.02	.07
	(.01)	(.01)

Note. The table reports the response categories read (on average) and the estimated marginal means for the number of response categories re-fixated after controlling for the covariate re-fixation rate. To control for length differences of response categories between the two question formats, we divided the number of response categories re-fixated by the number of characters (Ferreira & Clifton, 1986).

Figure 3 includes six exemplary gaze plots of different respondents for the three questions on political issues for both experimental groups (A/D vs. IS). Gaze plots display the scan path (i.e., sequence of the eye movements) across visual stimuli – e.g., question stems and response categories. While the circles denote fixations, the lines between them denote saccades. The size of the circles is proportionally related to the duration of the fixation itself. The gaze plots show that in both groups, the center of the response scales was fixated most intensively. This finding is in line with the response distributions revealing that respondents most frequently selected the middle response category. This indicates a relation between the intensity of looking at an area of the response scale and selecting a category from this area (see Höhne & Lenzner, 2015). Furthermore, the respondents did not fixate on the last categories at the bottom of the scales and, thus, did not read all response categories, irrespective of the question format. Respondents selected these categories less frequently. In addition, it can be observed that especially the IS categories were fixated more intensively than their A/D counterparts. More precisely, it is to see that respondents cause more and longer fixations as well as more re-fixations.

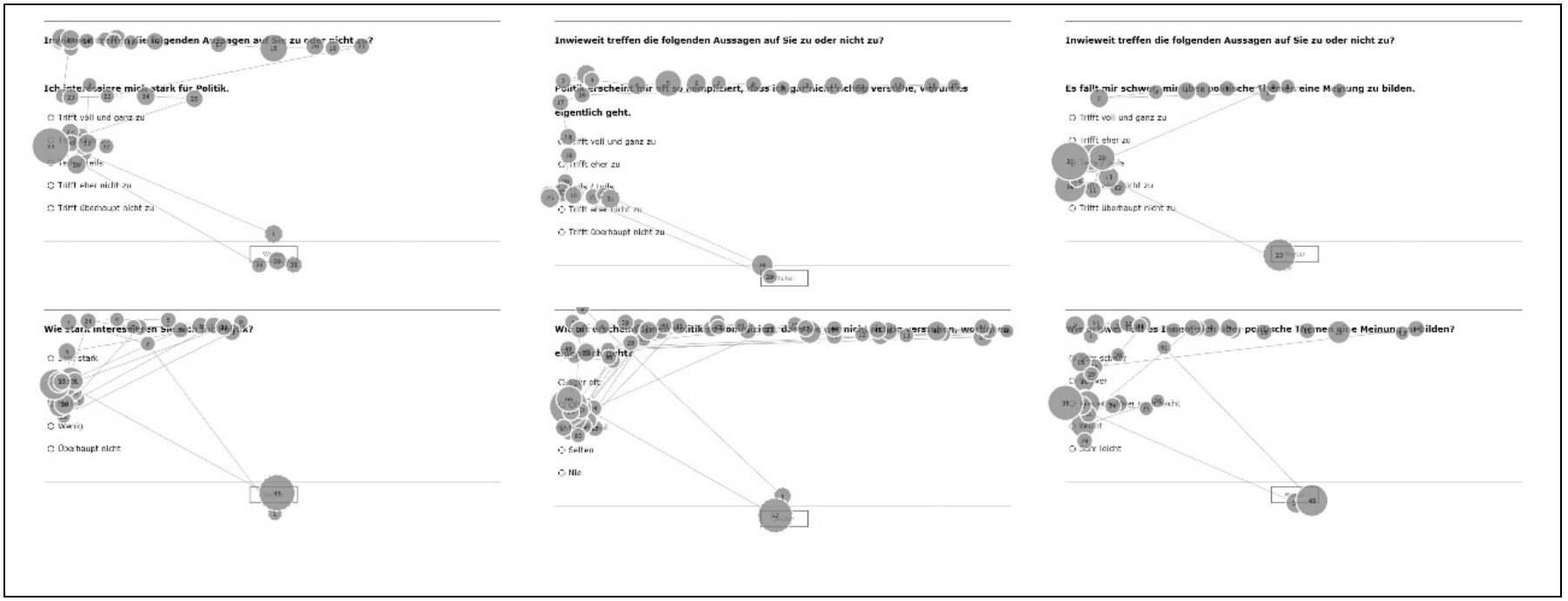

Figure 3. Study II: Gaze plots of different respondents for the three A/D and IS questions

Note. The three gaze plots above correspond to the first experimental group (agree/disagree condition) and the three gaze plots below correspond to the second experimental group (item-specific condition). Each gaze plot displays the eye movements of one respondent. The circles indicate fixations and the lines between the circles indicate saccades. The numbers within the circles indicate the order of the fixations and the size of the circles is proportional to the fixation time.

Hypothesis 3 postulated that respondents re-fixate the response categories more often when answering IS questions compared to A/D questions, indicating more conscientious processing of the response categories. In line with our expectation, Table 5 shows considerable differences in the mean number of re-fixations between the A/D and IS question format. The results of the statistical analyses show highly significant differences between the two experimental groups [$F(1,72) = 11.43$, $p < .001$, partial $\eta^2 = .14$]. Hence, it again appears that respondents process the IS response categories much more intensively than the A/D response categories.

5.2.4 Discussion and Conclusion

The aim of this eye-tracking study was to investigate the processing of A/D and IS questions to evaluate the response effort associated with these two question formats. The results show no differences between the A/D and IS questions with respect to respondents' fixation on the question stems. Considering the characteristics of A/D and IS questions in general and the questions tested in particular (see Appendix B for details about the questions), it can be seen that there is no substantial semantic and/or syntactic difference between them (for a general overview of question comprehension, see Graesser et al., 2006). The main difference between these two question formats is that the stems are either formulated as statements (declarative form) or formulated as "real" questions (interrogative form).

Regarding the processing of the response categories, however, the results show that the responding to IS questions is characterized by more conscientious processing than the responding to A/D questions. This is indicated by the fact that the response categories of the IS questions are fixated more frequently and longer than those of the A/D questions. In addition, the response categories of IS questions are more frequently re-fixated than

their A/D counterparts. These findings support the notion of the manner of asking survey questions. According to this reasoning, A/D questions – though theoretically requiring intricate and sophisticated cognitive processing – promote a state of boredom and weariness and, thus, superficial cognitive processing. IS questions, in contrast, change the manner of asking permanently. It is to presume that they demand and encourage a more conscientious consideration of the response categories and, thus, require more effort in responding.

Contrary to our expectations, we did not find any significant differences in the number of response categories read between A/D and IS questions. The gaze plots presented in Figure 3 reveal that respondents in both question formats do not fixate (and therefore do not read) all response categories, but rather frequently skip the last response categories. Similar to the findings of Höhne and Lenzner (2015) in their study of A/D questions, respondents also seem to be able to extrapolate the response continuum of IS questions after reading the initial response categories. This circumstance might be attributed to the fact that the response categories in both A/D and IS questions form rating scales, which follow an ordered and closed response continuum. This implies that respondents do not have to fixate all response categories. Nonetheless, as shown in this study, the IS response categories seem to trigger a more intensive processing than the A/D ones.

All in all, there are two limitations to this study. First, the experiment does not investigate the response quality obtained by A/D and IS questions. While we found evidence that IS questions trigger a more considerate response process than A/D questions, it remains open whether this has also a positive effect on the quality of respondents' answers. Saris et al. (2010), for instance, found that responses to IS questions are of a higher quality than responses to comparable A/D questions. Nevertheless, further research is

necessary in order to systematically evaluate the connection between the cognitive processing of both question formats and their response quality. Second, the experimental study only investigated IS questions changing the manner of asking (i.e., employing different response categories). However, IS questions do not necessarily have to change the manner of asking if all questions deal with the same dimension of interest (see chapter 2.1 “Question Format” as well as Study III and V). Hence, it is yet unclear whether the crucial difference between A/D and IS questions is the repetition of the response categories and/or the directness of the question format. In an attempt to explore this issue further – and as suggested by one of the anonymous reviewers – we additionally compared the fixations on A/D and IS questions for each of the three experimental questions. If it is correct that the crucial difference between the two question formats is the repetition of response categories, then for the very first question the differences in fixation counts and times should be smaller than for the subsequent questions or even non-existent. The analysis did not reveal substantial differences between the three experimental questions with respect to means and effect sizes. This might be an indicator that not only the continuity (i.e., repetitiveness of response categories) but also the directness (i.e., addressing the content dimension) seems to affect the processing of the questions. This, however, is only an attempted explanation lacking empirical evidence. It is also possible – again, as suggested by the anonymous reviewer – that we did not find these differences between our experimental questions because the participants were skilled survey respondents (i.e., 81% had participated previously in at least one web survey) who might have been familiar with the A/D question format and recognized it from their past experience. Therefore, future research is needed to investigate whether the directness and/or continuity of the question format is responsible for the differences in processing A/D and IS questions.

The empirical findings of this study have theoretical and practical implications for social science research. From a theoretical point of view, we found supporting evidence for the notion of the manner of asking survey questions affecting the process of responding to A/D and IS questions. More precisely, this implies that an indirect and continuous manner of asking survey questions – as is the case with the A/D question format – negatively affects diligence and thoughtfulness in responding. Furthermore, we were able to show that there are substantial differences between the presumed cognitive complexity of question formats and the response effort expended in responding. Therefore, we argue that the manner of asking should be taken into consideration in future studies to get a better understanding of how respondents process A/D and IS questions. From a practical perspective, the data indicate that IS questions are characterized by a more active and intensive response process than A/D questions.

5.3 Study III[72]

This study is based on a field experiment with four conditions. The study investigates the respective response effort and response quality associated with the A/D and IS question format using paradata. Response effort is assessed by using response times and answer changes, the letter of which is operationalized by means of mouse clicks. Response quality is measured by using different forms of satisficing.

5.3.1 Research Hypotheses

Before presenting the research hypotheses, we introduce the notion of "presentation mode" of survey questions. In this study, we presented A/D and IS questions in single and grid presentation modes. For single questions, response categories are arranged below each question, but for grids they are arranged alongside each question. As mentioned in chapter 2.1 "Question Format", IS questions do not have to necessarily change the manner of asking if the questions address the same dimension of interest (e.g., importance). Then, IS questions can be employed in grid presentation mode. While differences between A/D and IS single questions are discussed in the survey literature, they are (mostly) neglected within grid presentation mode.

Measuring response times enjoys a long tradition in social psychology and survey research (Couper & Kreuter, 2013; Yan & Tourangeau, 2008) and has been proven as a useful strategy to investigate cognitive information processing (Bassili, 1996; Bassili & Fletscher, 1991; Bassili & Scott, 1996; Fazio, 1990; Yan & Olson 2013). It is generally assumed that the time of processing corresponds (directly) to the response effort required to answer a question. This, in turn, suggests that the longer respondents need to respond,

[72] This chapter is based on:
Höhne, J.K., Schlosser, S., & Krebs, D. (2017). Investigating cognitive effort and response quality of question formats in web surveys using paradata. *Field Methods, 29*, 365–382.

the higher the response effort must be. As discussed in chapter 2.2 “Response Effort”, the IS question format supports a more active and intensive response process than the A/D question format due to a direct and commonly alternating manner of asking. For this reason, we hypothesize that IS questions cause significantly higher response times than their A/D counterparts, irrespective of the presentation mode (Hypothesis 1).

As suggested by Heerwegh (2011) and Stern (2008), answer changes can be caused by different aspects of response formats, such as low discriminatory power of response categories. In general, answer changes can result from the uncertainty of respondents to decide between response categories, especially if an intensive response process is supposed. Hence, the number of answer changes can be used as an (additional) indicator of response effort. Due to the direct and variant manner of asking, this theoretical linkage seems to apply to IS questions. We therefore hypothesize that IS questions produce significantly more answer changes than their A/D counterparts, irrespective of the presentation mode (Hypothesis 2).

Response quality depends largely on the respondent’s motivation (Krosnick, 1991) implying that the lower the motivation, the higher the probability of low response quality. As indicators of low response quality, we use speeding – i.e., extremely fast responding without the probability of careful processing – dropouts – i.e., break off from the survey – and non-differentiation – i.e., rating several issues identically (see Chang & Krosnick, 2009; Galesic & Bosnjak, 2009; Lenzner, Kaczmirek, & Lenzner, 2010; Roßmann, Gummer, & Silber, 2017).[73] Because A/D questions incite a perfunctory cognitive processing due to an indirect and unchanging manner of asking, we hypothesize to observe a lower

[73] Although speeding – operationalized by response times – can also be seen as an indicator of response effort, in this thesis (Study III and IV) speeding is used as an indicator of response quality (see Conrad et al., 2017).

response quality for A/D than for IS questions, again this is irrespective of the presentation mode (Hypothesis 3).

5.3.2 Method

5.3.2.1 Research Design

To control for carryover effects from grid to single questions and vice versa, we varied the presentation sequence within each question format. Table 6 describes our 2-by-2 research design.

Table 6. Study III: Research design defined by question format and presentation sequence

Experimental Group	Question Format	Presentation Sequence	Group Size
1	A/D	GS	255
2	IS	GS	237
3	A/D	SG	278
4	IS	SG	235

Note. A/D = agree/disagree and IS = item-specific; GS = grid/single questions and SG = single/grid questions.

5.3.2.2 Survey Questions

The single questions used were adapted from the Cross Cultural Survey for Work and Gender Attitudes (2010). The grid questions, in contrast, were partially adapted from the German General Social Survey (2006). Taking questions from established social surveys includes the advantage of using implicitly valid questions for our study.

The web survey contained 24 experimental questions: 8 single questions that dealt with work and competition (achievement motivation) and 16 grid questions that dealt with the importance of job motivation. For each question adapted from the surveys, we developed an A/D and IS counterpart preserving question content as well as possible. All

questions were attitudinal questions with a 5-point end-labeled response scale (see Appendix C for all single and grid questions in both question formats).

One key requirement to investigate the response effort associated with different question formats using response times is to keep the questions identical in the number of syllables because these influence the processing time (Baddeley, 1992). Therefore, we developed all single and grid questions including response categories, so that they did not differ in more than two syllables from one another.

5.3.2.3 Procedure

This study was conducted at two German Universities in May 2015.[74] Participants were invited by email (in total, we sent out 58,829 email invitations). The invitation included an introduction to the topics of the survey and a URL link directing respondents to the survey. Once there, an introductory page informed them about the procedure of the survey and instructed them to read the questions carefully and in the given order. Respondents were additionally informed that they were participating in an experiment and that different types of paradata (e.g., response times) would be collected during the survey. While each single question appeared on an extra screen, grid questions were presented on two screens, with eight questions per grid, respectively.

We used "Embedded Client Side Paradata (ECSP)" (Schlosser, 2016), a JavaScript-based system to observe respondents' activities and behavior during the web survey. Recorded and stored paradata include response times in milliseconds (i.e., the time elapsing between question presentation on the screen and the time the page was submitted), mouse clicks, and information about the activity of the web-survey page (i.e., detecting whether

[74] We tested whether the results differ between the Universities but there are no substantial differences.

the window that hosts the web survey is also the active or processed one; see Callegaro, 2013, p. 269).

5.3.2.4 Sample

A total of 2,884 students participated in the web survey, which corresponds to a response rate of 4.9%. Due to technical difficulties, we excluded all participants with mobile devices, such as smartphones and tablets (n = 709), versions of Microsoft Internet Explorer earlier than 11 (n = 24), and those who had deactivated JavaScript (n = 28). Moreover, some participants were ineligible because they only visited the title page (n = 122), they dropped out of the web survey before being asked any experimental questions (n = 357), German was not their mother tongue (n = 107), and two experimental conditions being subject of an additional study on adverbial intensifiers (n = 532). Altogether, 1,005 participants remained for statistical analyses.

To deal with response time outliers, we used, among others, “SurveyFocus (SF),” a new outlier definition procedure (see Höhne & Schlosser, 2017). Accordingly, we first exclude as outliers all respondents who left the web-survey page for a certain time. For the remaining respondents, we applied an outlier definition based on the response time distributions (Hoaglin, Mosteller, & Tukey, 2000): Excluding as outliers all respondents with response times below or above the median plus/minus the upper and lower quartile range multiplied by three. The amount of outliers is evenly distributed over the experimental groups.

Participants were between 17 and 52 years old with a mean age of 24.8 (SD = 4.2). A total of 52% of the participants were female. All participants were graduates of a college preparatory school or university and 93% had participated in a web survey once be-

fore. Sample composition over the four experimental groups did not reveal any statistically significant differences regarding age [$\chi^2(3) = 2.44$, $p = .49$], gender [$\chi^2(3) = 2.35$, $p = .50$], and survey experience [$\chi^2(3) = .60$, $p = .89$].

5.3.3 Results

Depending on the specific research question, the analysis of web-survey paradata can be classified into different levels of aggregation (Heerwegh, 2011; Yan & Olson, 2013). To investigate the response effort associated with A/D and IS questions, we analyze paradata on a survey level (i.e., paradata are aggregated across respondents and variables for each experimental group).

5.3.3.1 Response Times

Response times as an indicator of response effort are the primary dependent variable of the following analyses. Due to the fact that response times typically are right skewed (Fazio, 1990; Ratcliff, 1993), we applied a log transformation to decrease the skewness of response time distributions. Moreover, the following statistical analyses are based on unadjusted response time measurements without checking for baseline reading speed (see Couper & Kreuter, 2013).

In order to investigate whether the question format (A/D vs. IS) and/or the presentation sequence (grid/single vs. single/grid questions) have an effect on response times, we calculated two-way analyses of variance for single and grid questions separately. While question format has a highly significant effect on response times for single questions [$F(1,915) = 37.64$, $p < .001$], there is no main effect of the presentation sequence and no interaction effect between the two factors. For grid questions, however, neither significant main effects nor an interaction effect are observable.

Next, to investigate whether IS questions produce significantly higher response times than A/D questions, we calculated two one-way analyses of variance with the factor question format (A/D vs. IS) for single and grid questions separately and Cohen's d as a measure of effect size (see Cohen, 1969). Table 7 displays the statistical results for single questions and reveals significant differences in average response times between the four experimental groups. In line with the expectations, single IS questions require consistently higher response times than their A/D counterparts. This result is independent of the presentation sequence. Moreover, there are no significant differences between the two groups (1 and 3) receiving A/D questions and the two groups (2 and 4) receiving IS questions. This result is confirmed by Cohen's d because the effect sizes of those groups getting the same question format are very small (d = .12 and d = .06). Hence, presentation sequence does not matter within question format, thereby strengthening the result of the two-way analysis of variance for single questions.

For grid questions, the empirical significance level of p = .197 corroborate that there are no differences in average response times between the four experimental groups (see Table 8). Accordingly, Cohen's d shows consistently very small effect sizes (d < .2). Hence, the response effort associated with A/D and IS questions seems to be the same when they employed in grids. To annotate the observed small tendency of longer response times for the "grid/single" than the "single/grid" sequence in Table 8, we additionally analyzed the time to respond to the first grid question. The results reveal that the grid/single sequence causes significantly longer response times than the single/grid sequence for the first grid question.

5.3.3.2 Answer Changes

We used answer changes measured by mouse clicks as a second indicator of response effort (Heerwegh, 2011; Stern, 2008). Again, we analyzed the amount of answer changes separately for single and grid questions. Regarding single questions, there are no significant differences between the four experimental groups [$F(3,915) = .15$, $p = .928$]. This implies that answer changes in single questions are neither affected by question format nor by presentation sequence. For IS grid questions, however, there is a (slightly) significant difference in average answer changes between the groups 2 and 4 [$F(3,918) = 2.72$, $p = .044$]. This reveals a higher number of answer changes for the single/grid sequence. Thus, presentation sequence seems to have an impact on answer changes within IS grid questions.

5.3.3.3 Response Quality

We used the following indicators to measure response quality: speeding, dropouts, and non-differentiation. Since the presentation sequence for these analyses is irrelevant, groups 1 and 3 (having received A/D questions) as well as groups 2 and 4 (having received IS questions) were pooled. The analyses were still conducted separately for single and grid questions.

As an indicator of speeding, we considered the lower 10th percentile of all response times to be extremely fast in responding. For single questions, 6.6% (n = 61) of respondents producing the fastest response times did so in response to an A/D question and 2.8% (n = 26) in response to an IS question [$\chi^2(1) = 11.45$, $p < .001$]. This indicates a significant difference in speeding between the two question formats. For grid questions, 6.3% (n = 58) of the respondents with the fastest response times produced them in response to an A/D question and 4.8% (n = 44) in response to an IS question [$\chi^2(1) = .97$, $p = .33$]. The

response quality of grid questions in terms of extremely fast responding does not significantly differ between the two question formats. Again, the observed differences occur only between single A/D and IS questions.

The second response quality indicator are dropout rates, implying a premature break off from the survey. For A/D single questions 2.4% (n = 28) and for IS single questions 2.7% (n = 31) of the respondents dropped prematurely from the web survey [$\chi^2(1) = .77$, $p = .38$]. Similarly, for A/D grid questions 1.8% (n = 20) and for IS grid questions 1.6% (n = 18) of the respondents have left the web survey sooner [$\chi^2(1) = .01$, $p = .92$]. Apparently, the two question formats do not differ in terms of dropouts, irrespective of the presentation mode.

Finally, identical ratings of different issues or objects are an indicator of non-differentiation. To investigate this response quality indicator, the proportion of respondents who answered several survey questions equally was considered. Comparing the number of these respondents for single questions results in marginal differences; 9.6% (n = 88) for the A/D question format and 8.9% (n = 82) for the IS question format [$\chi^2(1) = .11$, $p = .75$]. For grid questions, a similar pattern is observable since selection of the same response category over all questions occurred for 15.5% (n = 143) of the A/D questions and for 13.6% (n = 125) of the IS questions [$\chi^2(1) = .18$, $p = .67$]. Accordingly, A/D and IS single and grid questions yield a comparable response quality in terms of non-differentiation.

Table 7. Study III: Mean differences of response times between the four experimental groups for the eight aggregated single questions

	Differences of means for log-transformed response time data					
Presentation Form	Agree/Disagree (GS) Group 1	Item-Specific (GS) Group 2	Agree/Disagree (SG) Group 3	F value $df_1 = 3$	df_2	p value
Item-Specific	–.108**			13.75	915	p < .001
(GS) Group 2	(.44)					
Agree/Disagree	–.030	*.078**				
(SG) Group 3	(.12)	*(.33)*				
Item-Specific	*–.123***	–.015	–.093**			
(SG) Group 4	*(.50)*	(.06)	(.40)			

Note. *p < .01; **p < .001. Coefficients in italics represent differing experimental conditions and cannot be compared directly. Mean differences: means of row conditions minus means of column conditions. Cohen's d in parentheses states the effect sizes. GS = grid/single sequence and SG = single/grid sequence.

Table 8. Study III: Mean differences of response times between the four experimental groups for the sixteen aggregated grid questions

	Differences of means for log-transformed response time data					
Presentation Form	Agree/Disagree (GS) Group 1	Item-Specific (GS) Group 2	Agree/Disagree (SG) Group 3	F value $df_1 = 3$	df_2	p value
Item-Specific	.018			1.56	918	p = .197
(GS) Group 2	(.07)					
Agree/Disagree	.049	*.032*				
(SG) Group 3	(.19)	*(.13)*				
Item-Specific	*.020*	.003	–.029			
(SG) Group 4	*(.08)*	(.01)	(.12)			

Note. Coefficients in italics represent differing experimental conditions and cannot be compared directly. Mean differences: means of row conditions minus means of column conditions. Cohen's d in parentheses states the effect sizes. GS = grid/single sequence and SG = single/grid sequence.

5.3.4 Discussion and Conclusion

The aim of this study was to investigate the response effort and response quality associated with A/D and IS questions by means of paradata. Overall, we found that IS single questions require a higher response effort – in terms of response times – than A/D single questions. Regarding the theoretical complexity of cognitive processing and response effort, we conclude – for single presentation mode – that while cognitive processing of A/D questions is theoretically seen as more demanding than that of IS questions, the actually expended response effort is lower.

This discrepancy between the complexity of cognitive processing and response effort can be explained using the notion of the manner of asking. For A/D single questions, the response continuum runs continuously from agree to disagree or vice versa, so that respondents have to read response categories only once and repeat the same answering task over and over again, requiring little response effort and, thus, resulting in shorter response times than IS questions. In contrast, IS single questions usually vary the manner of asking from question to question since the underlying response categories address the dimension of interest directly. Therefore, reading IS questions, including the IS response categories, requires not only more time but challenges respondents to engage in an active and intensive response process for each question in the light of its specific content.

Besides question format, the presentation mode (single or grid) seems to have an impact on response effort. Although we expected that A/D and IS questions employed in grid presentation mode reveal a similar outcome as A/D and IS questions employed in single presentation mode (see Hypothesis 1), no differences in response times could be observed and, thus, no differences with respect to response effort. At a first glance, it seems that the directness of the manner of asking IS questions is overshadowed by the

presentation mode so that IS questions organized in grids adjust to A/D questions. However, this is only an attempted explanation, which lacks empirical evidence. As similarly suggested by Study I and II, it would be interesting if future studies continue the research on the presentation of both question formats.

Next, we investigated the number of answer changes in A/D and IS questions presented in single and grid mode as further indicator of response effort. Contrary to recent research indicating systematic differences in answer changes between response formats (Heerwegh, 2011; Stern, 2008), we could not find major differences between A/D and IS questions, irrespective of presentation mode. Hence, the difficulty of mapping answers into response categories seems to be equal for both question formats in the two presentation modes. However, it must be noted that we tested end instead of fully labeled response scales, eventually rendering the offered response categories virtually equivalent.

Regarding low response quality, we found only minor differences between the two question formats, except for speeding. Contrary to the previous expectation (see Hypothesis 3), we observed that speeding only occurs more often for single A/D questions than for their IS counterparts. Again, we consider the grid presentation mode as a responsible factor for this finding.

Using a student sample might be seen as a limitation to this study. However, the use of student samples does not limit the external validity of empirical results per se. Moreover, since the participants of the study are university students presumably with above average (cognitive) abilities participating voluntarily without any compensation, we tested our research question under harsh conditions. Because of self-selection, however, it is possible that mainly participants with high motivation or interest started the survey (see Couper et al., 2004). Since these factors are important regarding response

quality, it would be desirable for researchers in the future to control for the influence of motivation and/or interest.

Due to the fact that the research is related to PC devices, we suggest that further research compares the response effort and response quality associated with A/D and IS questions across different device types (see Study IV). Moreover, the influence of presentation sequence on grids might be interesting because the results indicate that the time respondents need to answer the first grid question is longer for the grid/single than for the single/grid sequence. Therefore, attention should be given to the processing time for the first and subsequent grid questions.

In sum, the findings have both theoretical and practical implications. From a theoretical perspective, we can show that A/D questions do not require a higher response effort than IS questions. Indeed, theoretically, A/D questions imply a quite difficult and sophisticated response process, as suggested by Carpenter and Just (1975). In the context of the empirical findings, it seems more likely that they tempt respondents to engage in a perfunctory cognitive response process. We consider the indirect and invariant manner of asking a major reason for this. This objection is especially applicable to single questions. The empirical findings and the theoretical reasoning illustrate apparent differences between the presumed complexity of cognitive processing and the expended response effort. In other words, demanding cognitive processing might be attenuated by an indirect and unchanging manner of asking decreasing diligence in question processing.

5.4 Study IV[75]

Similar to the previous Study III, this study is based on a field experiment with four experimental conditions. The study investigates the response effort and response quality associated with A/D and IS questions. However, unlike Study III, where the comparison of the question formats includes only PCs, this study includes both PCs and mobile devices (i.e., smartphones). Response effort is measured by using response times and response quality is measured by using different forms of satisficing.

5.4.1 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 response effort and response quality before providing a summary of them in Table 9.

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 response effort must be (Conrad et al., 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

[75] This chapter is based on:
Höhne, J.K., Revilla, M., & Lenzner, T. (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*. DOI: 10.1027/1614-2241/a000151.

the manner of asking concept (Höhne, Schlosser, & Krebs, 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, 2016). 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).

Table 9. Study IV: Research hypotheses on response times and response quality

Response Times	IS questions produce longer response times than A/D questions. (Hypothesis 1)
	The differences between A/D and IS questions are higher in PCs than in smartphones. (Hypothesis 1a)
	The differences between A/D and IS questions are higher in 7-point than in 5-point response scales. (Hypothesis 1b)
Response Quality	IS questions produce better response quality than A/D questions. (Hypothesis 2)
	The differences between A/D and IS questions are higher in PCs than in smartphones. (Hypothesis 2a)
	The differences between A/D and IS questions are higher in 7-point than in 5-point response scales. (Hypothesis 2b)

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, Schlosser, & Krebs, 2017; Kunz, 2017;

Malhotra, 2008). 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 et al., 2009; Conrad et al., 2017; Höhne, Schlosser, & Krebs, 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 manner of asking, 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 (Revilla, Saris, & Krosnick, 2014).

5.4.2 Method

5.4.2.1 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). 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).

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 by clicking "Next") and SurveyFocus (Höhne & Schlosser, 2018; Höhne, Schlosser, & Krebs, 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).

5.4.2.2 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.[76] 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.

[76] 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.

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.

5.4.2.3 Research Design

To test our research hypotheses on response 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 10.

Table 10. Study IV: Research design defined by device type and question format

Experimental Group	Device Type	Question Format	Group Size
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).

5.4.2.4 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.

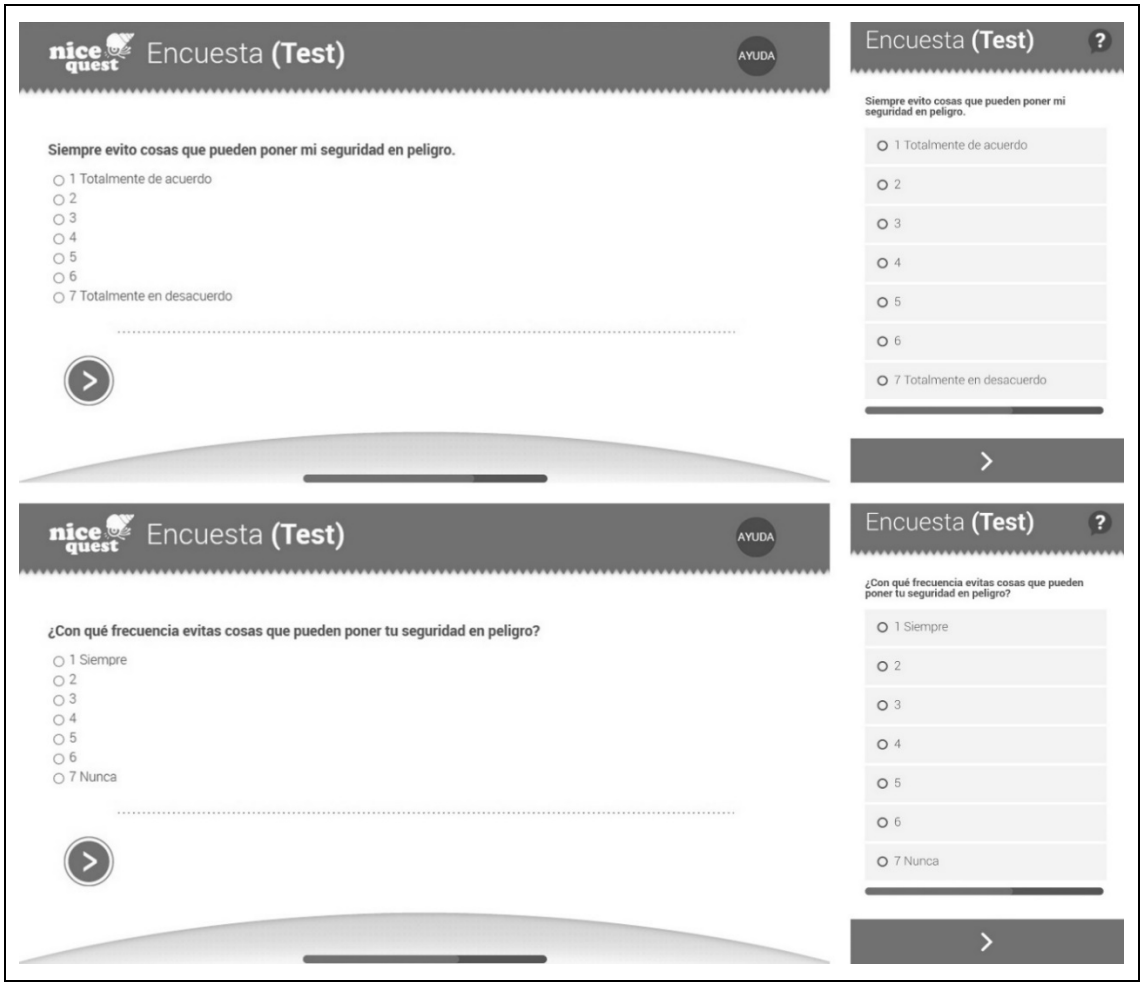

Figure 4. Study IV: Screenshots of one A/D question and one IS question with a 7-point response scale for PCs and smartphones

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). See Appendix D for English translations of all questions including response categories.

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 Appendix D for the questions used and Figure 4 for an example of a question with a 7-point response scale). 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.

5.4.2.5 Analytical Strategies

Response effort: To deal with response time outliers, we employed a two-step procedure using the paradata "SurveyFocus (SF)"[77] (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.[78] For the remaining respondents, we applied a

[77] 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).
[78] At the page-level, discontinuously processing respondents are about 2%.

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.

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 response 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 15th percentile of all response times – after outlier definition[79] – and compared the proportion of speeders between the experimental groups. Hence, we considered speeding as a relative phenomenon (Conrad et al.,

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

2017; Höhne et al., 2017; Malhotra, 2008). As a robustness check, we tested other thresholds, such as the 5th, 10th, 20th, and 25th percentile. Again, the main conclusions did not change. Regarding primacy effects, we used the number of responses to the first response category.[80] 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-square

[80] 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.

tests for speeding and unpaired t-tests for primacy effects. To test the second order hypotheses on response times and response quality (Hypotheses 1a and b as well as Hypotheses 2a and b, respectively), we conducted significance tests across the conditions of interest for means and proportions using the Z-statistic.

5.4.3 Results

5.4.3.1 Response Times

For the following statistical analyses, response times – as indicator of response effort – are the primary dependent variable. Table 11 provides the results of the one-way analyses of variance (ANOVAs). In line with our previous expectation (see Hypothesis 1), 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 Hypothesis 1a), 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.

A comparison of the findings for the questions with 5-point and 7-point response scales (see Table 11) did not result in higher response time differences for the 7-point scale questions (see Hypothesis 1b). 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).

5.4.3.2 Response Quality

To evaluate the response quality of A/D and IS questions, we used speeding and primacy effects. Table 12 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.

As previously hypothesized (see Hypothesis 2), 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 Hypothesis 2a), 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 Hypothesis 2b), the differences are somewhat more pronounced for questions with 5-point compared to 7-point scales for PCs (Z = .13) and smartphones (Z = .23).

Table 11. Study IV: Mean differences of response times per syllable for the six aggregated A/D and IS questions with 5- and 7-point response scales for PCs and smartphones

Experimental Comparison	5-Point Scales			7-Point Scales		
	Mean Differences	Effect Size	p value	Mean Differences	Effect Size	p value
A/D (PC) vs. IS (PC)	–2.01*	.22	p < .001	–1.87*	.23	p < .001
A/D (Smartphone) vs. IS (Smartphone)	–2.10**	.24		–1.43*	.20	
A/D (PC) vs. A/D (Smartphone)	–1.03	.12		–1.67*	.22	
IS (PC) vs. IS (Smartphone)	–1.11	.12		–1.22	.15	

Note. *p < .05; **p < .01. Mean differences: first group minus second group. The p values are based on F-tests. We also calculated Cohen's d as measure of effect size.

Table 12. Study IV: 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 < .001.

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 12 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 Hypothesis 2). 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 Hypothesis 2a), 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 Hypothesis 2b), 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.

5.4.4 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 response effort – assessed by response times – and response quality – assessed by speeding and primacy effects. Thus, it is an extension of Study III. 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 the manner of asking.

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 (2018), 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 (see Study V). 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 response 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 manner of asking, which provokes a superficial response process. Thus, we argue that the manner of asking 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 the former studies (see Study I, II, and III) and the empirical findings, it is recommendable to employ IS instead of A/D questions since they evoke a more thoughtful responding.

5.5 Study V[81]

This study is based on a quasi-experimental design with four conditions. The study primarily investigates the response quality associated with A/D and IS questions in terms of their susceptibility to response order effects. Furthermore, respondents are asked to evaluate the A/D and IS questionnaires in terms of task-oriented manageability and motivational potential.

5.5.1 Research Hypotheses

As suggested by recent research (Kuru & Pasek, 2016; Liu, Lee, & Conrad, 2015), there seem to be differences in the occurrence of response bias in A/D and IS questions. In this study, we define response bias as susceptibility of a given question format (A/D and IS) to scale direction effects (decremental vs. incremental). A/D questions apply identical response scales to all questions so that respondents must repeat the same answering task for all questions. Due to this unchanging manner of asking, it is to assume that A/D questions foster boredom and weariness and a perfunctory response process. In addition, responses to A/D statements do not refer directly to the underlying dimension of interest, which additionally impedes responding to A/D questions. For these reasons, the A/D question format seems to dismay respondents and discourage them from expending much effort when responding. By way of comparison, IS questions normally change the manner of asking. Thus, they might incite respondents to engage in an active and comparatively more attentive response process for each question. In addition, the IS question format does not suffer from an indirect manner of asking, which might support response quality.

[81] This chapter is based on:
Höhne, J.K., & Krebs, D. (2018). Scale direction effects in agree/disagree and item-specific questions: A comparison of question formats. *International Journal of Social Research Methodology*, *21*, 91–103.

In line with this reasoning, we postulate the following research hypotheses: First, we expect significant response order effects within the A/D question format but no response order effects within the IS question format (Hypothesis 1). Second, we expect respondents to perceive the IS questionnaires as more demanding and complex than the A/D questionnaires (Hypothesis 2a). Finally, we expect respondents to perceive the IS questionnaires as more interesting, inspiring, and diversified than the A/D questionnaires (Hypothesis 2b).

5.5.2 Method

5.5.2.1 Survey Instruments

The questions used were adapted from the Cross Cultural Survey for Work and Gender Attitudes (2010) and the German General Social Survey (2006). Taking questions from established social surveys offers the advantage of using repeatedly tested questions. For the study, we used 12 questions – 5 dealt with achievement motivation, 4 with intrinsic job motivation, and 3 with extrinsic job motivation. For each question adapted from the surveys, we developed an A/D and IS counterpart preserving question content as well as possible.[82] All questions were presented with a 5-point, fully-verbalized response scale and no numeric values (see Appendix E for the questions used).

Extrinsic job motivation refers to the importance of anticipated job characteristics (e.g., income and career), as they are generally but not solely under an individual's control. Intrinsic job motivation, in contrast, refers to job commitment (e.g., autonomy and responsibility). Achievement motivation, on the other hand, refers to competitiveness,

[82] While the A/D questions were presented in grids alone, the IS questions were presented both in grids (job motivation) and in single presentation mode (achievement motivation).

implying an appreciation of interpersonal challenges. While extrinsic job motivation describes expectations with respect to job characteristics, intrinsic job and achievement motivation describe attitudes toward a job or other people (see Krebs, Berger, & Ferligoj, 2000; Spence & Helmreich, 1983).

The decision to use motivational questions for this study is based on the author's experience of nearly identical results across several student-based surveys (see Krebs & Hoffmeyer-Zlotnik, 2010). Achievement as well as intrinsic and extrinsic job motivation have been proven to be stable across different student cohorts and over time.

Following the 12 motivational questions, the questionnaires contained several questions on political and societal issues, which were also either presented as A/D or IS questions. At the end, respondents were asked to evaluate the manageability and motivational potential of the questionnaire using opposite adjective pairs.[83]

5.5.2.2 Data Collection

The research was conducted at the University of Göttingen (Germany) in the winter terms of 2014 and 2015. Both years, the experimental study was conducted in the beginning of the winter term among students participating in an introductory lecture on methodology; the aim was to ensure the recruitment of "freshmen" on the topic of social science and survey methodology. All students taking part in the lecture (meaning all students present in the lecture hall) were invited to participate in the study and informed that they would be participating in a study dealing with different survey research topics and that their data would be treated confidentially. According to the split versions (A/D vs. IS), paper ques-

[83] These pairs followed the principle of a semantic differential (see Osgood, Suci, & Tannenbaum, 1957) and, thus, were measured on 7-point response scales (see Appendix E for the questions used).

tionnaires were sorted systematically before their distribution to ensure random assignment. Completing the questionnaire took approximately 10 min. At the end of each semester, the results of the study were presented to the students and they received a debriefing, during which they were told that they had participated in an experiment on response quality. While response scale direction in 2014 followed a decremental order (i.e., running from positive to negative), in 2015 the response scale direction was changed to an incremental order (i.e., running from negative to positive). Table 13 outlines the research design.

Table 13. Study V: Research design defined by question format and response scale direction

Experimental Group	Question Format	Scale Direction	Group Size	Field Time
1	A/D	Decremental	209	2014
2	IS	Decremental	202	
3	A/D	Incremental	268	2015
4	IS	Incremental	251	

The study is based on the same surveys conducted in two successive years. To ensure comparability, the questions employed did not differ with respect to content and format. Furthermore, the order and position of all questions were identical. Only the response scale direction was changed.

5.5.2.3 Sample

Altogether, n = 976 students participated in the study (n = 435 in 2014; n = 541 in 2015). However, we excluded all respondents older than 30 years[84] and students participating in

[84] The threshold of 30 years is based on an outlier definition using the mean plus/minus the standard deviation multiplied by 4. We conducted all statistical analyses with and without these respondents but there are no differences.

both years, leaving for the analyses n = 411 in 2014 (with a mean age of 21 and SD of 2.1) and n = 519 in 2015 (with a mean age of 21 and SD of 2.2). 56% of these students were female, 81% (2014) and 83% (2015) were in their first semester, and the bulk of the students were enrolled in a social science program (83% in 2014; 85% in 2015).

In order to evaluate the sample composition across the two groups from 2014 and 2015, we compared them with respect to the following characteristics: Age [$\chi^2(1) = .00$, $p = .96$], gender [$\chi^2(1) = .00$, $p = .98$], subject of study [$\chi^2(1) = .42$, $p = .52$], and number of semesters [$\chi^2(1) = .75$, $p = .39$]. The sample composition across the four experimental groups did not reveal any statistically significant differences with respect to these characteristics. Furthermore, the two scale direction groups did not differ with respect to these characteristics within either question format. Thus, the two groups can be considered comparable.

5.5.2.4 Analytical Strategies

First, descriptive statistics (e.g., mean and standard deviation) for all 12 questions were computed. Second, a first order confirmatory factor analysis (CFA) model containing three latent variables (achievement as well as intrinsic and extrinsic job motivation) was tested with respect to measurement equivalence between decremental and incremental response order within the AD and IS question format. Third, comparisons of latent means were conducted. For these analyses we used Mplus version 6.12. Due to the fact that the indicators of the latent variables were measured on 5-point response scales, we assumed metric scale level (see Rhemtulla, Brosseau-Liard, & Savalei, 2012) and, thus, used the MLR (instead of MLM) estimator, which provides robust standard errors and takes non-normality into account. Finally, to investigate the dimensionality of the opposite adjective

pairs and respondents' evaluations of the A/D and IS questionnaires, we conducted an explorative factor analysis (EFA)[85] and unpaired t-tests using SPSS version 24.

5.5.3 Results

To investigate the occurrence of response order effects with respect to A/D and IS questions, we firstly take a look at the empirical distributions. Afterwards, we control for measurement equivalence and compare latent means of achievement as well as intrinsic and extrinsic job motivation across the scale direction groups within each question format. Finally, we analyze respondents' evaluations to test whether the IS questionnaires are evaluated differently from the A/D questionnaires.

5.5.3.1 Descriptive Statistics

In order to ensure appropriate statistical testing, all questions were recoded to identical values from 1 "positive" to 5 "negative". We calculated means, standard deviations, skewness, and kurtosis for all questions. Considering the results in Table 14, it can be observed that responses to the decremental A/D questions are more positive than those to the incremental A/D questions – i.e., lower means for the decremental than for the incremental scale direction. In particular, this is observable for achievement and intrinsic job motivation. The results for extrinsic job motivation, however, break ranks since only a slight tendency toward the postulated direction can be observed. The IS questions also show a lower mean with respect to the decremental scale direction, irrespective of the motivational dimension. However, there are no substantial mean differences between the two scale directions within the IS question format. All in all, this is a first evidence that IS questions are more robust against effects of the scale direction than A/D questions.

[85] We conducted a Principal Axes Factor Analysis (PAF) using direct Oblimin rotation in the default setting ($\Delta = 0$).

Table 14. Study V: Means, standard deviations, skewness, and kurtosis for decremental and incremental scale directions within the AD and IS question format

	A/D Question Format							
	Decremental Order				Incremental Order			
Questions	Mean	SD	Skew-ness	Kur-tosis	Mean	SD	Skew-ness	Kur-tosis
Competition	2.92	.98	–.09	–.50	3.15	1.03	–.07	–.52
Achievement	2.78	1.07	.17	–.77	2.90	1.15	.19	–.79
Improvement	3.19	1.07	–.37	–.53	3.36	1.09	–.23	–.62
Making an effort	2.60	1.07	.32	–.61	2.79	1.09	.29	–.65
Being the best	2.73	1.07	.34	–.41	2.94	1.09	.24	–.50
Autonomy	1.95	.76	.47	.15	2.08	.83	.62	.41
Applying skills	1.64	.69	.87	.57	1.86	.86	.99	.87
Responsibility	2.04	.77	.55	.45	2.22	.84	.53	.35
Realizing ideas	1.91	.78	.70	.59	2.10	.87	.51	–.06
Income	2.40	.85	.25	–.13	2.43	.84	.41	.38
Prospects	2.26	.85	.48	–.06	2.36	.90	.82	.72
Career	2.44	.92	.27	–.32	2.49	.98	.46	–.13
	IS Question Format							
Competition	3.01	.95	.08	–.08	3.10	.90	.23	–.40
Achievement	2.96	1.05	.08	–.46	3.04	.94	.08	–.49
Improvement	3.27	.99	–.22	–.30	3.31	.89	.03	–.25
Making an effort	2.49	.88	.55	.32	2.51	.86	.48	.00
Being the best	2.54	.97	.33	–.23	2.64	.82	.09	–.02
Autonomy	1.98	.79	.35	–.58	2.05	.81	.45	–.04
Applying skills	1.65	.68	.67	–.31	1.72	.70	.66	–.02
Responsibility	2.09	.73	.41	.12	2.14	.79	.28	–.13
Realizing ideas	1.92	.82	.42	–.74	1.95	.81	.67	.56
Income	2.41	.82	.43	.37	2.51	.79	.43	.46
Prospects	2.29	.92	.33	–.35	2.40	.91	.68	.40
Career	2.37	.94	.39	–.30	2.48	.92	.44	–.02

Note. The first 5 questions refer to achievement motivation, the next 4 questions to intrinsic job motivation, and the last 3 questions to extrinsic job motivation, respectively. Responses were recoded to identical values from 1 “positive” to 5 “negative”.

Standard deviations do not differ substantially between the two scale directions (decremental and incremental), regardless of the question format. In addition, the values of the skewness and kurtosis are quite small, indicating a relatively normal distribution of respondents’ answers to A/D and IS questions.

5.5.3.2 Measurement Equivalence

In a first step, we conducted confirmatory factor analyses (CFA) within the two question formats (A/D and IS) and formulated separate but yet identical baseline models for each scale direction (decremental and incremental). The CFA model contained three correlated latent variables: Achievement motivation (5 indicators), intrinsic job motivation (4 indicators), and extrinsic job motivation (3 indicators), respectively. In each model, we admitted one error covariance between two questions on achievement motivation. All baseline models revealed good fit statistics. Table 15 displays the statistical results.

Using multi-group confirmatory factor analysis (MGCFA), we tested configural invariance (M_0) by analyzing the baseline model simultaneously for both scale directions within each question format. Given CFI-values higher than .95 and RMSEA-values lower than .05, configural invariance was accepted for response scale direction within the A/D and IS question format. Next, in order to test the metric invariance, factor loadings were constrained to equality between the decremental and incremental order of the response categories within the A/D and IS question format. The model fit statistics were sufficient to accept metric invariance as well. Finally, to compare latent means, scalar invariance must also be tested, which is accomplished by additionally imposing equality constraints on the intercepts. Again, scalar invariance holds for both response scale directions in both question formats.

Criteria for comparability between models with increasing equality constraints are, firstly, non-significant differences between (mean-adjusted) chi-square values (Byrne, 2012) and, secondly, differences between CFI's and RMSEA's lower than .01 (Cheung & Rensvold, 2002). These two criteria hold for all model differences in Table 15. Despite the equality constraints imposed on factor loadings (M_1) and additionally on intercepts

across the two response scale directions, the model (M_2) fits the data quite well. Thus, the estimates associated with this solution seem to be trustworthy and are interpreted accordingly.

Table 15. Study V: Test of measurement equivalence of decremental and incremental response order within the A/D and IS question format

A/D Format	χ^2	df	CFI	RMSEA
Configural (M_0)	156.93 (1.05)	100	.97	.049
Metric (M_1)	163.34 (1.07)	112	.97	.044
Scalar (M_2)	182.77 (1.07)	124	.97	.044
IS Format				
Configural (M_0)	147.37 (1.07)	100	.97	.045
Metric (M_1)	160.87 (1.07)	112	.97	.044
Scalar (M_2)	166.94 (1.07)	124	.97	.039

Note. An error covariance between two achievement motivation questions ("competition" and "making an effort") was admitted. The results are based on MLR estimation. We report scale correction factors (in parentheses) to calculate likelihood ratio (LR) tests (see Reinecke, 2014, p. 119-121).

5.5.3.3 Differences in Latent Means

In testing the differences in latent means between response scale directions (decremental vs. incremental), we used the groups with an incremental order of response categories as reference groups. Since respondents' answers for scale directions were recoded from 1 "positive" to 5 "negative", negative signs of estimates indicate that responses to questions with decremental order tend to the positive scale point. Table 16 shows the results of the comparison of latent means for the A/D question format [$\chi^2(121) = 168.34$ (1.07); CFI = .973; RMSEA = .040] and the IS question format [$\chi^2(121) = 163.54$ (1.07); CFI = .972; RMSEA = .03]. Although the absolute numbers are relatively small, they show significant differences in latent means of the response scale direction for two of the three motivation

dimensions, namely, achievement and intrinsic job motivation. However, for extrinsic job motivation, latent means do not significantly differ between the two scale directions.

Table 16. Study V: Latent mean differences between decremental and incremental response order in the A/D and IS question format

A/D Format	Est.	S.E.	C.R.	p value
Achievement motivation	–.129	.060	–2.079	.038
Job motivation (intrinsic)	–.155	.047	–3.267	.001
Job motivation (extrinsic)	–.057	.059	–.963	.336
IS Format				
Achievement motivation	–.036	.042	–.857	.392
Job motivation (intrinsic)	–.063	.056	–1.129	.259
Job motivation (extrinsic)	–.094	.060	–1.575	.115

Note. Response scales were recoded to identical values from 1 "positive" to 5 "negative". Reference group is the incremental response order. C.R. = critical ratio (indicating z-values). Unstandardized results.

In contrast, for the IS question format Table 16 reveals only minor differences. Again, the negative signs of estimates suggest that respondents' answers to questions with a decremental order are also shifted to the positive scale point, although not significantly. While response order affects latent means in the A/D question format, in the IS question format differences are noticeably smaller. Hence, this result supports the expectation that the A/D question format is more susceptible to response order effects than the IS question format, as suggested by Hypothesis 1.

5.5.3.4 Evaluation of Questionnaires

As regards respondents' ratings of the questionnaires, we hypothesized that the IS ones are not only perceived as more demanding and complex than the A/D ones, but also as more interesting, inspiring, and diversified (Hypotheses 2a and 2b). To investigate these evaluations, we used five pairs of opposite adjectives and 7-point response scales. According to the decremental or incremental response order in the questionnaires and to be

consistent, the evaluation scales started with an adjective carrying either a positive or a negative connotation. All responses were recoded to identical values from 1 “positive” to 7 “negative”.

For both scale directions, we conducted an exploratory factor analysis for the five opposite adjective pairs, which resulted in a two-factor solution. The first factor describes task-oriented manageability (i.e., demanding and complex) and the second factor describes motivational potential (i.e., interesting, inspiring, and diversified) of the questionnaires. The explained variance (S), factor loadings (λ), and factor correlations (r) are similar for both scale directions; decremental [S = .79, λ > .80, r = .30] and incremental [S = .75, λ > .70, r = .12]. Irrespective of response scale direction, the evaluations of the questionnaires are almost identical.

Table 17 displays the overall average ratings of the A/D and IS questionnaires. As an indicator of the effect size, we additionally calculated Cohen’s d coefficient. In line with Hypothesis 2a, respondents assessed the IS questionnaires as significantly more demanding and complex than the A/D questionnaires – Cohen’s d exhibits values equal to or higher than .2. With respect to Hypothesis 2b, which refers to interest, inspiration, and diversification (and their counterparts), no significant differences are observable between the two questionnaires. Hence, Cohen’s d reveals only very small effect sizes (d < .1). Altogether, it appears that respondents perceive the completion of the IS questionnaires as more effortful than the completion of the A/D questionnaires without affecting motivation.

Table 17. Study V: Means and standard deviations (in parentheses) of respondents' evaluations of the A/D and IS questionnaires

Adjective Pairs	A/D Format	IS Format	Effect Size	p value
interesting/boring	3.60 (1.45)	3.69 (1.47)	.06	.368
undemanding/demanding	2.56 (1.40)	2.95 (1.58)	.26	.001
inspiring/tedious	3.96 (1.36)	3.98 (1.39)	.02	.791
simple/complex	2.39 (1.42)	2.69 (1.72)	.20	.004
diversified/monotonous	3.49 (1.56)	3.54 (1.58)	.03	.584

Note. Responses were recoded to identical values from 1 "positive" to 7 "negative". We calculated Cohen's d to determine the effect sizes between the ratings of the two question formats. The significance levels, however, are based on the results of unpaired t-tests.

5.5.4 Discussion and Conclusion

Although the effects of response order as well as A/D and IS questions on response behavior have already been the subject of separate investigations, these two methodological issues have not yet been investigated simultaneously. In addition, we investigated respondents' perception of the A/D and IS questionnaires with respect to manageability and motivational potential. Taken together, the results of this study revealed first and foremost the existence of response order effects within the A/D but not within the IS question format. Second, respondents evaluated the IS questionnaires as more challenging than the A/D questionnaires. Third, IS questionnaires are not rated as being more inspiring, interesting, or diversified than A/D questionnaires since both only received moderate evaluations.

The postulated response order effects are only observable for achievement and intrinsic job motivation, which refer to individual self-descriptions. Although motivation is generally seen as a relatively stable personality trait, its measurement can be affected by different circumstances, such as question format and response scale direction. In contrast, for extrinsic job motivation, there were no observable response order effects. However,

the fact that extrinsic job motivation is not affected by the response order had already been demonstrated by Krebs & Hoffmeyer-Zlotnik (2010). Hence, the results of the present study add to their finding. Furthermore, this result seems to be related to the specific content of this motivational dimension; the indicators address commonly desirable job characteristics, such as income and career. Altogether, it seems that respondents follow a "hierarchy of importance" with question content over scale direction and question format (see Toepoel & Dillman, 2011). This implies that a question's content might not be susceptible to response order effects, irrespective of the question format used. However, this is only an attempted explanation lacking empirical evidence. To get more information about the relation between question content and question format and/or scale direction, we recommend that future research investigates a hierarchical order between question content and different question design strategies.

A further important point is that Study III found no differences in response quality (i.e., speeding, dropouts, and non-differentiation) between A/D and IS questions presented in grids. In this study, however, we are able to show that scale direction effects occur in A/D but not in IS grid questions. This suggests that IS questions seem to be more robust against effects of response order than A/D questions, even if they are employed in grid presentation mode. In our opinion, this is attributable to the direct manner of asking, which might elicit more attention and, thus, prevent the occurrence of response order effects in IS questions. Again, further research is necessary to improve the current state of research.

With respect to respondents' evaluations of the A/D and IS questionnaires, only Hypothesis 2a, which refers to task-oriented aspects, was supported by the data. Obviously, respondents must devote more effort in responding to IS than to A/D questions

and, most importantly, they seem to perceive this fact. We also expected that the IS questionnaires would be perceived as more interesting, inspiring and diversified than the A/D ones. The empirical results, however, do not corroborate Hypothesis 2b. Therefore, it would be interesting to investigate the influence of different question contents and questionnaire lengths on respondents' evaluations. Furthermore, it would be worthwhile to combine these standardized evaluations with more objective criteria (e.g., response times) in upcoming studies.

All in all, there are two limitations to this study. First, the empirical findings of this study are based on two cross-sectional student surveys with random assignment of respondents to the A/D or IS question format within each survey. However, response order (decremental vs. incremental) was not randomized but rather assigned to the two data collection points in 2014 and 2015. Therefore, it would be desirable for future research to apply a research design with repeated measurements (i.e., a within-subject design). Accordingly, in the first wave, respondents might be randomly assigned to one of four groups (combining question format and response order); then, in the second wave, they would be assigned to the same question format but with the opposite response order. Second, as our results are based on students' responses, we have a relatively unique sample. This, however, does not fundamentally restrict the validity and generalizability of the empirical findings. Due to the fact that the respondents are university students presumably with above-average (cognitive) abilities taking part voluntarily without any incentives and/or credits, we tested our research question under harsh conditions and would expect larger differences in a general population sample. Furthermore, the results correspond to the findings of prior research on response order effects (see Krebs & Hoffmeyer-Zlotnik, 2010), which additionally corroborate the results.

To conclude: The empirical findings correspond to the results of the former studies (see Study I, II, III, and IV) and expectations about response behavior regarding A/D and IS questions. They also show that the A/D question format is indeed more susceptible to scale direction effects than the IS question format. Thus, the results allow the conjecture to the effect that refined IS questions are very promising and appropriate for coping with response bias. Furthermore, the results suggest that the question content matters insofar as questions contain internal rather than external self-descriptions. Finally, and most importantly, the results reveal that the scale direction is not only a "matter of taste" since it can have implications for the process of drawing conclusions from survey results. For instance, health surveys or surveys on political satisfaction based on question formats suffering from response order effects might affect decision-making. For this reason, further research should address response order effects within A/D and IS questions using different contents to improve the quality of and the trust in survey responses. Although there is no final recommendation regarding the correct response scale direction, we nevertheless suggest that response order be kept in mind when designing surveys. In line with the previous studies, we recommend the use of the IS instead of the A/D question format because it appears to be more robust against response order effects.

6. General Discussion and Conclusion

Fowler's (1995) well-known monograph "Improving Survey Questions: Design and Evaluation" started a methodological discussion, which continues to this day, of the adequacy of A/D and IS question formats. This thesis contributes to this discussion by investigating the respective response effort and response quality associated with these two question formats using different methods and techniques (see chapter 4.1 "Conceptual Specifications"). In addition, this thesis introduced a new theoretical framework – called "manner of asking" – that explains differences between A/D and IS questions in terms of response effort and response quality. According to this framework, the two question formats can be distinguished by two specific characteristics: The directness (i.e., the form of the request for an answer) and the variation (i.e., the continuity of the response scales) of the manner of asking survey questions. While the A/D question format poses indirect statements and identical response scales, the IS question format poses direct questions and usually changing response scales. Based on the manner of asking framework and contrary to the current state of research, it was expected that the IS question format is characterized by a higher response effort than the A/D question format (see chapter 3. "General Research Questions"). Furthermore, it was expected that IS questions produce a higher response quality than A/D questions. The effort in responding to survey questions – either posed in the A/D or IS question format – is not solely based on the mental tasks that are theoretically demanded. It is based on the thoughtfulness and adequacy with which these theoretically demanded mental tasks are accomplished.

Five experimental studies were conducted to investigate the response effort and response quality associated with the A/D and IS question format. All studies investigated response effort and three of them (Study III, IV, and V) additionally investigated response

quality. The studies were based on non-probability samples and most of them were conducted using web-based, self-administered survey modes. However, they differed in terms of characteristics that included question topics, language, and response scale design (for a list of all study characteristics, see Table 2).

The five studies employed different indicators of response effort. Study I and II used eye-tracking methodology (e.g., fixation count, fixation time, and re-fixations), Study III and IV used paradata (i.e., response times and answer changes), and Study V used respondents' evaluations (i.e., ratings across five opposite adjective pairs). In general, the results of the five studies indicate that IS single questions require more response effort than A/D single questions.[86] Only regarding answer changes no differences were observed between the question formats. This lack of differences in terms of answer changes can be attributed to response scale characteristics. Both studies (Study III and IV) that investigated answer changes employed end-labeled instead of fully labeled response scales. It is possible that the response scales offered appear virtually equivalent.

With respect to grid questions it is more difficult to draw a clear conclusion. Study III shows no differences in terms of response times or in terms of answer changes. In addition, Study V does not allow to differentiate clearly between single and grid questions because in that study respondents rated the entire survey instrument, which contained either A/D or IS questions in both presentation modes.

The use of eye-tracking methodology returned an interesting finding on response effort. Study I and II showed that A/D and IS questions differ regarding respondents' gaze behavior (i.e., fixation count, fixation time, and re-fixations). However, this phenomenon

[86] To be precise, Study I compared three A/D grid questions with three IS single questions. The results, however, indicated that the IS questions produced more and longer fixations than the A/D counterparts employed in a grid.

is only true for the response categories, it does not apply to question stems. This indicates inequality between the two question formats in terms of response category processing and equality in terms of question stem processing.

As shown in Study V, there are also no differences between the A/D and IS question formats in terms of motivational potential – i.e., interest, inspiration, and diversification (and their counterparts). Potential explanations for this finding are the content of the questions, the length of the questionnaires, and/or the presentation modes (i.e., single and grid). It seems worthwhile to employ additional methods, such as cognitive interviewing, to allow a refined and elaborated investigation of motivational potential, which is difficult to achieve by using standardized evaluations.

Like the findings on response effort, there is a clear tendency for single questions with respect to response quality. The three studies investigating response quality (Study III, IV, and V) reveal that IS single questions produce better response quality than A/D single questions in terms of speeding (i.e., extremely fast responding) and/or response order effects (i.e., higher endorsement of categories at the beginning of the scale). As described in chapter 4.3 “Characteristics of the Studies”, Study IV tested 7-point response scale questions in addition to 5-point response scale questions. Although all response quality differences tended toward the expected direction, they were only significant for the questions with 5-point response scales. Study III and IV also investigated dropouts (i.e., premature break off from of the survey), item nonresponse (i.e., questions with missing data), and non-differentiation (i.e., rating different questions identically). The findings indicate that the occurrence of these forms of response biases was either very small or did not vary substantially between the two question formats. Future studies should systematically investigate the susceptibility of A/D and IS questions to other forms of response

bias using a range of research designs. Also, it is worthwhile for survey researchers to examine how the variation of response scale length can affect the occurrence of response bias.

With respect to A/D and IS grid questions the results are somewhat ambiguous. For instance, Study III found no differences between the two question formats in terms of response effort and response quality. Study V, in contrast, revealed that A/D questions employed in grids are more susceptible to response order effects than IS questions employed in grids. However, this seems to apply only to a specific class of question contents, namely questions dealing with individual self-descriptions, such as intrinsic job motivation. This finding is in line with previous studies (see Krebs & Hoffmeyer-Zlotnik, 2010) and indicates a "hierarchy of importance" (Toepoel & Dillman, 2011). This means that some question contents might not be susceptible to response bias, such as response order effects, regardless of the question design strategy used. To build on the findings of Study V, future research should systematically investigate the relationship between different question contents, question formats, and scale directions.

While the findings for single questions are relatively consistent and in line with the expectations derived from the manner of asking framework, the findings for grid questions are mixed; this applies to response effort and response quality. Potential explanations for the mixed results are the presentation mode and/or the invariant manner of asking that they employ. As shown by previous research, survey questions organized in grids seem to incite superficial responding and produce low response quality (see Couper et al., 2013). Similarly, grid questions' invariant manner of asking may also support a superficial responding. Thus, when IS questions are used with an invariant response scale, they

adjust to A/D questions in terms of response effort and response quality. In order to investigate this issue further, it would be necessary to compare the response behavior incited by A/D questions presented individually and in grids with IS questions presented individually and in grids, keeping the content dimension constant (e.g., importance). Such a 2-by-2 research design would allow to figure out whether the response scale continuity and/or the directness of the manner of asking affects response behavior. For this purpose, researchers can make use of the sixteen A/D and IS grid questions of Study III (see Appendix C).

Taking a closer look at the methodological characteristics of the five empirical studies, it is to see that there are several methodological aspects that future studies can address. First, it is worthwhile to investigate the performance of A/D and IS questions across different respondent groups using probability samples. The respondents of all five studies are characterized by comparatively high levels of survey experience and/or education. Therefore, it would be interesting to investigate how the two question formats perform across less experienced and/or educated respondents. Second, all studies employed visual communication channels. Thus, it remains unclear how the A/D and IS question formats perform in aural or verbal channels, such as face-to-face interviews without show cards. Finally, it also remains open whether the empirical findings reported in this thesis can be generalized to other languages and/or cultural settings because the data used are only from two countries (Germany and Spain). Hence, it is necessary to extend the comparison of the two question formats to cross-national and cross-cultural survey settings.

In conclusion, the main goal of this thesis was to provide a methodological contribution to the survey literature addressing the gap in the research on A/D and IS questions.

This goal has been achieved by conducting five empirical studies. The notion of the manner of asking provides a reasonable and empirically based framework to explain the differences between A/D and IS questions regarding response effort and response quality. The utility of the manner of asking framework is particularly useful in analyses involving A/D and IS questions in single presentation mode because this mode influences respondents' attentiveness and diligence in survey responding. The manner of asking framework also illustrates a mismatch between the theoretically expected complexity of question formats and the actually expended effort in responding. It is worthwhile for researchers to make use of the notion of the manner of asking in upcoming studies.

References

Alwin, D.F. (2007). *Margins of Error: A study of Reliability in Survey Measurement*. Hoboken, NJ: John Wiley & Sons.

Anand, S. (2008). Satisficing. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – Second Volume* (p. 797–799). Thousand Oaks, CA: Sage.

Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V. (2017). Zahlen zur Marktforschung. Retrieved July 14, 2017, https://www.adm-ev.de/index.php?id=21&L=0

Aronson, E., Wilson, T., & Akert, R. (2014). *Social Psychology.* Essex, UK: Pearson.

Baddeley, A. (1992). Working memory. *Science, 255*, 556–59.

Bassili, J.N. (1996). How and why of response time latency measurement in telephone surveys. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative process in survey research* (pp. 319–346). San Francisco, CA: Jossey-Bass Publisher.

Bassili, J.N., & Fletcher, J.F. (1991). Response-time measurement in survey research: A method for CATI and a new look at nonattitudes. *Public Opinion Quarterly, 55*, 331–346.

Bassili, J.N., & Scott, B.S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly, 60*, 390–99.

Bauman, C.W. (2008). Attitude strength. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – First Volume* (p. 42). Thousand Oaks, CA: Sage.

Baumgartner, H., & Steenkamp, J.B. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143–156.

Ben-Nun, P. (2008). Respondent fatigue. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – Second Volume* (pp. 742–743). Thousand Oaks, CA: Sage.

Biemer, P.P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly, 74*, 817–848.

Biemer, P.P., & Lyberg, L.E. (2003). *Introduction to Survey Quality*. Hoboken, NJ: John Wiley & Sons.

Billiet, J.B., & McClendon, M.J. (1998). On the identification of acquiescence in balanced sets of items using structural models. *Metodološki Zvezki, 14*, 129–150.

Billiet, J.B., & McClendon, M.J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling: A Multidisciplinary Journal, 7*, 608–628.

Bishop, G.F., & Smith, A. (2001). Response-order effects and the early Gallup split-ballots. *Public Opinion Quarterly, 65*, 479–505.

Bodenhausen, G.V., & Wyer, R.S. (1987). Social cognition and social reality: Information acquisition and use in the laboratory and the real world. In H.J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social Information Processing and Survey Methodology* (pp. 6–41). New York, NY: Springer.

Bradburn, N. (1978). Respondent burden. *Paper presented at the American Statistical Association*. Alexandria, VA: 35–40.

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

Bühner, M. (2011). *Einführung in die Test- und Fragebodenkonstruktion*. München, GER: Pearson.

Byrne, B.M. (2012). *Structural Equation Modeling with Mplus. Basic Concepts, Applications, and Programming*. New York, NY: Routledge.

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., Lozar Manfreda, K., & Vehovar, V. (2015). *Web Survey Methodology*. Thousand Oaks, CA: Sage.

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.

Cannell, C.F., Miller, P.V., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological Methodology* (pp. 389–437). San Francisco, CA: Jossey-Bass.

Carpenter, P.A., & Just, M.A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review, 82*, 45–73.

Chang, L., & Krosnick, J.A. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly, 73*, 641–678.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 233–255.

Cliff, N. (1959). Adverbs as multipliers. *Psychological Review, 66*, 27–44.

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

Conrad, F.G., Couper, M.P., Tourangeau, R., & Peytchev, A. (2006). Use and non-use of clarification features in web surveys. *Journal of Official Statistics, 22*, 245–269.

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

Converse, P.E. (1964). The nature of belief systems in mass publics. In D.A. Apter (Ed.), *Ideology and Discontent* (p. 1–74). New York, NY: The Free Press of Glencoe.

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

Couper, M.P. (2000). Usability evaluation of computer-assisted survey instruments. *Social Science Computer Review, 23*, 384–396.

Couper, M.P. (2005). Technology trends in survey data collection. *Social Science Computer Review, 18*, 486–501.

Couper, M.P. (2011). The future of modes of data collection. *Public Opinion Quarterly, 75*, 889–908.

Couper, M.P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society, 176*, 271–86.

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

Couper, M.P., Tourangeau, R., Conrad, F.G. & Crawford, S.D. (2004). What they see is what we get: Response options for web surveys. *Social Science Computer Review, 22*, 111–27.

Couper, M.P., Tourangeau, R., Conrad, F.G., & Zhang, C. (2013). The design of grids in web surveys. *Social Science Computer Review, 31*, 322–345.

Creswell, J.W., & Plano Clark, V.L. (2007). *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage.

DeCastellarnau, A. (2017). A classification of response scale characteristics that affect data quality: A literature review. *Quality and Quantity*. DOI 10.1007/s11135-017-0533-4

de Leeuw, E., Hox, J.J., & Dillman, D.A. (Eds.). (2008). *International Handbook of Survey Methodology*. New York, NY: Taylor & Francis.

Deming, W.E. (1944). On errors in surveys. *American Sociological Review, 9,* 359–369.

Eaton, A.A., & Visser, P.S. (2008). Attitudes. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – First Volume* (pp. 39–42). Thousand Oaks, CA: Sage.

Eagly, A.H., & Chaiken, S. (1998). Attitude structure and function. In D.T. Gilbert, S.T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology – Volume 1* (pp. 269–322). New York, NY: Oxford University Press.

Fazio, R.H., & Olson, M.A. (2003). Attitudes: Foundations, functions, and consequences. In M.A. Hogg & J. Cooper (Eds.), *The SAGE Handbook of Social Psychology* (pp. 139–160). London: Sage.

Fazio, R.H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M.S. Clark (Eds.), *Research Methods in Personality and Social Research* (pp. 74–97). Newbury Park, CA: Sage.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25,* 348–368.

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

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

Fowler, F., & Cosenza, C. (2008). Writing effective questions. In E. de Leeuw, J.J. Hox, & D.A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 136–160). New York, NY: Taylor & Francis.

Frankel, J. (1981). The effect of interview length and question type ("effort") on perceived respondent burden. *Paper presented at the American Educational Research Association*. Los Angeles, CA: 3–17.

Fuchs, M. (2008). Total survey error. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – Second Volume* (pp. 896–902). Thousand Oaks, CA: Sage.

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics, 22,* 313–328.

Galesic, M., Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73,* 349–360.

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

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

Geise, S. (2011). Eye tracking in communication and media studies: Theory, method, and critical reflection. *Studies in Communication and Media, 2,* 149–263

Graesser, A.C., Cai, Z., Louwerse, M., & Daniel, F. (2006). Question Understanding Aid (QUAID). A web facility that tests question comprehensibility. *Public Opinion Quarterly, 70*, 3–22

Graf, I. (2008). Respondent burden. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – Second Volume* (p. 739). Thousand Oaks, CA: Sage.

Groves, R.M. (2004). Measurement error across disciplines. In P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 1–28). Hoboken, NJ: John Wiley & Sons.

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

Groves, R.M., Lyberg, L.E. (2010). Total survey errors: Past, present, and future. *Public Opinion Quarterly, 74*, 849–879.

Groves, R.M., Singer, E., Corning, A.D., & Bowers, A. (1999). A laboratory approach to measuring the effects on survey participation of interview length, incentives, differential incentives, and refusal conversion. *Journal of Official Statistics, 15*, 251–268.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150.

Hanson, T. (2015). Comparing agreement and item-specific response scales: Results from an experiment. *Social Research Practice, 1*, 17–26.

Haraldsen, G. (2004). Identifying and reducing response burdens in Internet business surveys. *Journal of Official Statistics, 20*, 393–410.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review, 21*, 360–373.

Heerwegh, D. (2011). Internet survey paradata. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet. Advances in Applied Methods and Research Strategies* (pp. 325–348). New York, NY: Routledge.

Hippler, H.J., Schwarz, N., & Sudman, S. (Eds.). (1987). *Social Information Processing and Survey Methodology*. New York, NY: Springer.

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. (2018). 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., Revilla, M., & Lenzner, T. (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*. DOI: 10.1027/1614-2241/a000151

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 in web surveys using paradata. *Field Methods, 29*, 365–382.

Holbrook, A.L. (2008). Acquiescence response bias. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – First Volume* (pp. 3–4). Thousand Oaks, CA: Sage.

Hoogendoorn, A.W. (2004). A questionnaire design for dependent interviewing that addresses the problem of cognitive satisficing. *Journal of Official Statistics, 20*, 219–232.

Hoogendoorn, A.W., Sikkel, D. (1998). Response burden and panel attrition. *Journal of Official Statistics, 14*, 189–205.

Hox, J.J. (2008). Accommodating measurement error. In E.D. de Leuw, J.J. Hox, & D.A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 387–402). New York, NY: Taylor & Francis.

Jabine, T.B., Straf, M.L., Tanur, J.M., & Tourangeau, R. (1983). *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*. Washington, D.C.: National Academy Press.

Jans, M. (2008). Mode effects. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – First Volume* (pp. 475–480). Thousand Oaks, CA: Sage.

Jäckle, A. (2008). Dependent interviewing: Effects on respondent burden and efficiency of data collection. *Journal of Official Statistics, 24*, 411–430.

Jones, J. (2012). *Response burden: Introductory overview lecture*. Retrieved July 13, 2017, from https://ww2.amstat.org/meetings/ices/2012/papers/302289.pdf.

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

Just, M.A., & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122–149.

Kalaian, S.A. (2008). Research design. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – Second Volume* (pp. 724–731). Thousand Oaks, CA: Sage.

Kalaian, S.A., & Kasim, R.M. (2008). External validity. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – First Volume* (pp. 255–256). Thousand Oaks, CA: Sage.

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

Kennedy, C. (2008). Bipolar scale. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – First Volume* (pp. 63–64). Thousand Oaks, CA: Sage.

Knapp, T.R. (2008a). Validity. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – Second Volume* (pp. 938–939). Thousand Oaks, CA: Sage.

Knapp, T.R. (2008b). Random assignment. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – Second Volume* (pp. 674–675). Thousand Oaks, CA: Sage.

Krebs, D., Berger, M., & Ferligoj, A. (2000). Approaching achievement motivation. Comparing factor analysis and cluster analysis. *New Approaches in Statistical Applications: Metodoloski Zvezki, 16*, 147–171.

Krebs, D., & Hoffmeyer-Zlotnik, J.H. (2010). Positive first or negative first? Effects of the order of answering categories on response behavior. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*, 118–127.

Krebs, D., & Schmidt, P. (Eds.). (1993). *New Directions in Attitude Measurement*. Berlin: de Gruyter.

Kreuter, F. (Ed.). (2013). *Improving Surveys with Paradata. Analytic Uses of Process Information*. Hoboken, NJ: John Wiley & Sons.

Krosnick, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213–236.

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

Krosnick, J.A., & Alwin, D.F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*, 201–219.

Krosnick, J.A., Narayan, S., & Smith, W.R. (1996). Satisficing in Surveys: Initial evidence. In M.T. Braverman & J.K. Slater (Eds.), *New Directions for Evaluation: Advances in Survey Research* (pp. 29–44). San Francisco, CA: Jossey-Bass Publisher.

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.

Kühnel, S.M. (1993). Lassen sich ordinale Daten mit linearen Strukturgleichungsmodellen analysieren? *ZA-Information, 33*, 29–51.

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.

Lavrakas, P.J. (2008a). Closed-ended question. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – First Volume* (pp. 96–97). Thousand Oaks, CA: Sage.

Lavrakas, P.J. (2008b). Internal validity. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – First Volume* (pp. 345–351). Thousand Oaks, CA: Sage.

Lelkes, Y., & Weiss, R. (2015). Much ado about acquiescence: the relative validity and reliability of construct-specific and agree-disagree questions. *Research and Politics*. DOI: 10.1177/2053168015604173

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

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology*, *24*, 1003–1020.

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

Lenzner, T., Kaczmirek, L., & Galesic, M. (2014). Left feels right: A usability study on the position of answer boxes in web surveys. *Social Science Computer Review, 32*, 743–764.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1–55.

Lischewski, J. (2015). *Soziale Erwünschtheit im Licht des Rational-Choice Ansatzes*. Retrieved August 14, 2017, https://zenodo.org/record/55169#.V1gBT-RSE4E

Liu, M., Lee, S., & Conrad, F.G. (2015). Comparing extreme response styles between agree-disagree and item-specific scales. *Public Opinion Quarterly, 79*, 952–975.

Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., & Trewin, D. (Eds.). (1997). *Survey Measurement and Process Quality*. Chichester, UK: John Wiley & Sons.

Lord, F.M., & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

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

MacDonald, M.C., & Christiansen, M.H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review, 109*, 35–54.

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

Marsden, P.V., & Wright, J.D. (Eds.). (2010). *Handbook of Survey Research*. Bingley, UK: Emerald.

Mathews, C.O. (1929). The effect of the order of printed response words on an interest questionnaire. *Journal of Educational Psychology, 30*, 128–134.

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*, 191–205.

Mavletova, A., & Couper, M.P. (2015). A meta-analysis of breakoff rates in mobile web surveys. In D. Toninelli, R. Pinter, & P. de Pedraza (Eds.), *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies* (pp. 81–98). London, UK: Ubiquity Press.

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? *Field Methods, 26*, 21–39.

Metzler, A., Kunz, T., & Fuchs, M. (2015). The use and positioning of clarification features in web surveys. *Psihologija, 48*, 379–408.

Miller, P.V. (2008). Measurement error. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – First Volume* (pp. 457–460). Thousand Oaks, CA: Sage.

Neuert, C.E., & Lenzner, T. (2016). Incorporating eye tracking into cognitive interviewing to pretest survey questions. *International Journal of Social Research Methodology, 19*, 501–519.

Noelle-Neumann, E. (1976). Die Empfindlichkeit demoskopischer Messinstrumente. In Institut für Demoskopie Allensbach (Ed.), Allensbacher Jahrbuch der Demoskopie (pp. 7–26). Wien: Verlag Fritz Molden.

Nunes Silva, C. (2008). Experimental design. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – First Volume* (pp. 253–254). Thousand Oaks, CA: Sage.

Oldendick, R.W. (2008). Response alternatives. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – Second Volume* (pp. 749–751). Thousand Oaks, CA: Sage.

Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluation. In F. Kreuter (Ed.), *Improving Surveys with Paradata. Analytic Uses of Process Information* (pp. 43–72). Hoboken, NJ: John Wiley & Sons.

Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Parducci, A. (1983). Category ratings and the relational character of judgment. In H.G. Geissler, H.F.J.M. Buffart, E.L.J. Leeuwenberg, & V. Sarris (Eds.), *Modern Issues in Perception* (pp. 262–282). Berlin: VEB Deutscher Verlag der Wissenschaften.

Pasek, J., & Krosnick, J.A. (2010). Optimizing survey questionnaire design in political science: Insights from psychology. In J.E. Leighley (Ed.), *The Oxford Handbook of American Elections and Political Behavior* (pp. 27–50). New York, NY: Oxford University Press.

Peytchev, A., Couper, M.P., McCabe, S.E., & Crawford, S.D. (2006). Web survey design: Paging vs. scrolling. *Public Opinion Quarterly, 70*, 596–607.

Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., & Singer, E. (Eds.). (2004). *Methods for Testing and Evaluating Survey Questionnaires*. New York, NY: John Wiley & Sons.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*, 510–32.

Reinecke, J. (2014). *Strukturgleichungsmodelle in den Sozialwissenschaften*. München: Oldenbourg Verlag.

Rhemtulla, M., Brosseau-Liard, P.E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*, 354–373.

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., & Ochoa, C. (2015). Quality of different scales in an online survey in Mexico and Colombia. *Journal of Politics in Latin America, 7*, 157–177.

Revilla, M., & Saris, W. (2015). Estimating and comparing the quality of different scales of an online survey using an MTMM approach. In E. Engel (Ed.), *Survey Measurement: Techniques, Data Quality, and Sources of Error* (pp. 76–96). Frankfurt: Campus Verlag.

Revilla, M., Saris, W., & Krosnick, J.A. (2014). Choosing the number of categories in agree-disagree scales. *Sociological Methods and Research, 43*, 73–97.

Roe, D.J. (2008). Question stem. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – Second Volume* (pp. 665–666). Thousand Oaks, CA: Sage.

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie, 9*, 222–245.

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health, 14*, 1101–1108.

Roßmann, J., Gummer, T., & Silber, H. (2017). Mitigating satisficing in cognitively demanding grid questions: Evidence from two web-based experiments. *Journal of Survey Statistics and Methodology*. DOI: http://dx.doi.org/10.1093/jssam/smx020.

Rugg, D. (1941). Experiments in wording questions: II. *Public Opinion Quarterly, 5*, 91–92.

Rugg, D. & Cantril, H. (1944). The wording of questions. In H. Cantril (Ed.), *Gauging Public Opinion* (pp. 23–50). Princeton, NJ: Princeton University Press.

Russo, F. (2009). *Causality and Causal Modeling in the Social Sciences.* Berlin: Springer Verlag.

Saris, W., & Gallhofer, I.N. (2014). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ: John Wiley & Sons.

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.

Schaeffer, N.C., & Charng, H.W. (1991). Two experiments in simplifying response categories: Intensity ratings and behavioral frequencies. *Sociological Perspectives, 34*, 165–182.

Schnell, R., Hill, P.B., & Esser, E. (2013). *Methoden der empirischen Sozialforschung*. München: Oldenbourg Verlag.

Schlosser, S. (2016). Embedded client side paradata (ECSP). Retrieved June 8, 2016, https://zenodo.org/record/55169#.V1gBT-RSE4E

Schuman, H., & Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Thousand Oaks, CA: Sage.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology, 21*, 277–287.

Schwarz, N., Strack, F., Bless, H., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 47*, 195–202.

Schwarz, N., & Sudman, S. (Eds.). (1996). Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. San Francisco, CA: Jossey-Bass.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and Quasi-Expeimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin Company.

Sharp, L.M., & Frankel, J. (1983). Respondent burden: A test of some common assumptions. *Public Opinion Quarterly, 47,* 36–53.

Simon, H. (1957). *Model of Man: Social and Rational.* Hoboken, NJ: John Wiley & Sons.

Sirken, M., Hermann, D., Schechter, S., Schwarz, N., Tanur, J., & Tourangeau, R. (Eds.). (1999). *Cognition and Survey Research*. New York: Wiley.

Spence, J.T., & Helmreich, R.L. (1983). Achievement-related motives and behavior. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 10–74). San Francisco, CA: Freeman.

Survey Quality Predictor (2017). SQP Users' Manual. Retrieved September 9, 2017, http://sqp.upf.edu/media/files/sqp_users_manual.pdf

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.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680

Stocké, V., & Langfeldt, B. (2004). Effects of survey experience on respondents' attitudes towards surveys. *Science, 81,* 5–32.

Stoutenborough, J.W. (2008). Semantic differential technique. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – Second Volume* (pp. 810–812). Thousand Oaks, CA: Sage.

Strack, F., & Martin, L.L. (1987). Thinking, judging, and communicating: A process account of context effects in attitude surveys. In H.J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social Information Processing and Survey Methodology* (pp. 123–148). New York, NY: Springer.

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

Thurstone, L.L. (1929). Theory of attitude measurement. *Psychological Review, 36*, 222–241.

Tobii Technology (2012). Determining the Tobii I-VT fixation filter's default values. Retrieved November 23, 2016, http://www.tobiipro.com/siteassets/tobii-pro/learn-and-support/analyze/how-do-we-classify-eye-movements/determining-the-tobii-pro-i-vt-fixation-filters-default-values.pdf.

Toepoel, V., Das, M., & van Soest, A. (2008). Effects of design in web surveys: Comparing trained and fresh respondents. *Public Opinion Quarterly, 72*, 985–1007.

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

Toninelli, D., & Revilla, M. (2016). 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*, 153–169.

Tourangeau, R. (1984). Cognitive science and survey methods. In T.B. Jabine, M.L. Straf, J.M. Tanur, & R. Tourangeau (Eds.), *Cognitive Aspects of Survey Methodology:*

Building a Bridge between Discipline (pp. 73–100). Washington D.C.: National Academy Press.

Tourangeau, R., & Bradburn, N.M. (2010). The psychology of survey response. In P.V. Marsden & J.D. Wright (Eds.), *Handbook of survey research* (pp. 315–346). Bingley, UK: Emerald.

Tourangeau, R., Rips, L.J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Trobia, A. (2008). Questionnaire. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – Second Volume* (pp. 652–655). Thousand Oaks, CA: Sage.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Vannette, D.L., & Krosnick, J.A. (Eds.). (2017). The Palgrave Handbook of Survey Research. London, UK: Palgrave MacMillan.

van Vaerenbergh, Y., & Thomas, T.D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research, 25,* 195–217.

Weisberg, H.F. (2005). *The Total Survey Error Approach. A Guide to the New Science of Survey Research.* Chicago, IL: The University of Chicago Press.

Wenemark, M., Hollman Frisman, G., Svensson, T., & Kristenson, M. (2010). Respondent satisfaction and respondent burden among differently motivated participants in a health-related survey. *Field Methods, 22,* 378–390.

Whyte, W.F. (1943). *Street Corner Society: The Social Structure of an Italian Slum*. Chicago, IL: The University of Chicago Press.

Willis, G. (2008). Cognitive Aspects of Survey Methodology (CASM). In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – First Volume* (pp. 103–106). Thousand Oaks, CA: Sage.

Wilson, T.D., LaFleur, S.J., & Anderson, D.E. (1996). The validity and consequences of verbal reports about attitudes. In N. Schwarz & S. Sudman (Eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research* (pp. 91–114). San Francisco, CA: Jossey-Bass Inc.

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

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions. The effects of age, experience, and question complexity on web survey response times. *Applied Cognitive Psychology*, *22*, 51–68.

Yan, T., Williams, D., Maitland, A., & Tourangeau, R. (2016). Use of eye-tracking to measure response burden. *Paper presented at the Conference of the American Association of Public Opinion Research*. New Orleans, LA.

Ziniel, S. (2008). Split-half. In P.J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods – Second Volume* (pp. 833–834). Thousand Oaks, CA: Sage.

Appendix A

QUESTIONS TO COMPUTE BASELINE SPEED (Covariates)

BS 1: How successful do you think the government is nowadays in dealing with threats to Germany's security?

BS 2: And how successful do you think the government is nowadays in fighting unemployment?

Response categories to BS 1 and BS 2 are "very successful, quite successful, neither successful nor unsuccessful, quite unsuccessful, very unsuccessful"

AGREE/DISAGREE QUESTIONS (Grid Questions)

A/D 1: All in all, my health is good.

A/D 2: It is fair that people with higher incomes can afford better health care than people with lower incomes.

A/D 3: I am willing to pay higher taxes in order to improve health care for all people in Germany.

Response categories to A/D 1 – A/D 3 are "agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, disagree strongly"

ITEM-SPECIFIC QUESTIONS (Single Questions)

IS 1: How would you rate your health overall?

"Very good, good, neither good nor bad, bad, very bad"

IS 2: Is it fair or unfair that people with higher incomes can afford better health care than people with lower incomes?

"Very fair, fair, neither fair nor unfair, unfair, very unfair"

IS 3: To what extent would you be willing to pay higher taxes to improve health care for all people in Germany?

“In any case willing, fairly willing, somewhat willing, hardly willing, not at all willing”

Note. The order of the questions and the response categories are consistent with the presentation order in Appendix A. The A/D questions were presented in a grid with horizontally arranged response categories alongside each question and the IS questions were presented individually with horizontally arranged response categories below each question. The original German wordings of the questions are available from the author (hoehne@uni-mannheim.de) upon request. Abbreviations: BS = baseline speed; A/D = agree/disagree; IS = item-specific.

Appendix B

QUESTIONS TO COMPUTE BASELINE SPEED (Covariates)

BS 1: How successful do you think the government is nowadays in dealing with threats to Germany's security?

BS 2: And how successful do you think the government is nowadays in fighting unemployment?

Response categories to BS 1 and BS 2 are "very successful, quite successful, neither successful nor unsuccessful, quite unsuccessful, very unsuccessful"

AGREE/DISAGREE QUESTIONS (Single Questions)

A/D 1: I am very interested in politics.

A/D 2: Politics very often seem so complicated that I can't really understand what is going on.

A/D 3: I find it very difficult to make my mind up about political issues.

Response categories to A/D 1 – A/D 3 are "agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, disagree strongly"

ITEM-SPECIFIC QUESTIONS (Single Questions)

IS 1: How interested would you say you are in politics?

"very interested, fairly interested, somewhat interested, hardly interested, not at all interested"

IS 2: How often does politics seem so complicated that you can't really understand what is going on?

"very often, often, sometimes, rarely, never"

IS 3: How difficult or easy do you find it to make your mind up about political issues?

"very difficult, difficult, neither difficult nor easy, easy, very easy"

Note. The order of the questions and response categories correspond to the presentation in Appendix B. The response categories of all questions were presented vertically below the question stem. The original German wordings of the questions are available from the author (hoehne@uni-mannheim.de) upon request. Abbreviations: BS = baseline speed; A/D = agree/disagree; IS = item-specific.

Appendix C

AGREE/DISAGREE (Single Questions)

A/D S1: It is satisfying when I do everything as well as possible.

A/D S2: It is satisfying when I learn to do something special.

A/D S3: I endeavor to improve my performance.

A/D S4: I enjoy being in competition with other people.

A/D S5: I try harder when I am in competition with other people.

A/D S6: It annoys me when other people perform better than me.

A/D S7: It is important for me to accomplish a task better than other people.

A/D S8: It is important for me to win.

Response categories to A/D S1 – A/D S8 are "1 agree strongly – 5 disagree strongly"

ITEM-SPECIFIC (Single Questions)

IS S1: How satisfying is it to do everything as well as possible?

"1 very satisfying – 5 not at all satisfying"

IS S2: How satisfying is it to learn something special?

"1 very satisfying – 5 not at all satisfying"

IS S3: How much do you endeavor to improve your performance?

"1 very much – 5 not at all"

IS S4: How much do you enjoy being in competition with other people.

"1 very much – 5 not at all"

IS S5: How much harder do you try when you compete with other people?

"1 very much harder – 5 not at all harder"

IS S6: How much does it annoy you when other people perform better than you?

"1 very much – 5 not at all"

IS S7: How important is it to accomplish a task better than other people?

"1 very important – 5 not at all important"

IS S8: How important is it to win.

"1 very important – 5 not at all important"

AGREE/DISAGREE (Grid Questions)

A/D G1: A job with a high income is important for me.

A/D G2: A job with good promotion prospects is important for me.

A/D G3: A job with clear career perspectives is important for me.

A/D G4: A job that I can work autonomously on is important for me.

A/D G5: A job that allows to make use of my skills and talents is important for me.

A/D G6: A job where I have responsibilities for specific tasks is important for me.

A/D G7: A job that allows me to implement my own ideas is important for me.

A/D G8: A job with regular working hours is important for me.

A/D G9: A job where I am guided by a supervisor is important for me.

A/D G10: A job where I receive credit by other people is important for me.

A/D G11: A job where I can help other people is important for me.

A/D G12: A job with a safe professional future is important for me.

A/D G13: A job that contributes to the society is important for me.

A/D G14: A job with flexible working hours is important for me.

A/D G15: A job with a good working atmosphere is important for me.

A/D G16: A job with a short distance to work is important for me.

Response categories to A/D G1 – A/D G16 are "1 agree strongly – 5 disagree strongly"

ITEM-SPECIFIC (Grid Questions)

IS G1: How important is a job with a high income?

IS G2: How important is a job with good promotion prospects?

IS G3: How important is a job with clear career perspectives?

IS G4: How important is a job that you can work autonomously on?

IS G5: How important is a job that allows to make use of your skills and talents?

IS G6: How important is a job where you have responsibilities for specific tasks?

IS G7: How important is a job that allows you to implement your own ideas?

IS G8: How important is a job with regular working hours?

IS G:9 How important is a job where you are guided by a supervisor?

IS G10: How important is a job where you receive credit by other people?

IS G11: How important is a job where you can help other people?

IS G12: How important is a job with a safe professional future?

IS G13: How important is a job that contributes to the society?

IS G14: How important is a job with flexible working hours?

IS G15: How important is a job with a good working atmosphere?

IS G16: How important is a job with a short distance to work?

Response categories to IS G1 – IS G16 are "1 very important – 5 not at all important"

Note. The order of the questions and response categories correspond to the presentation in Appendix C. In single questions, response categories were horizontally arranged below each question and in grid questions response categories were horizontally arranged alongside each question. The original questions, including response categories, did not differ in more than two syllables from one another. The original German wordings of the questions are available from the author (hoehne@uni-mannheim.de) upon request. Abbreviations: A/D = agree/disagree; IS = item-specific; S = single questions; G = grid questions.

Appendix D

AGREE/DISAGREE (Single Questions; 5-Point Response Scales)

A/D 1: I like sharing my private life.

A/D 2: I never leave my belongings around.

A/D 3: I am relaxed most of the time.

A/D 4: I don't pay attention to details.

A/D 5: I get suspicious easily.

A/D 6: I don't like to draw attention to myself.

Response categories to A/D 1 – A/D 6 are "1 agree strongly – 5 disagree strongly"

ITEM-SPECIFIC (Single Questions; 5-Point Response Scales)

IS 1: How much do you like sharing your private life?

"1 I like it extremely – 5 I don't like it at all"

IS 2: How often do you leave your belongings around?

"1 always – 5 never"

IS 3: How often are you relaxed?

"1 always – 5 never"

IS 4: How much do you pay attention to details?

"1 I pay extremely attention – 5 I don't pay attention at all"

IS 5: How easily do you get suspicious?

"1 extremely easily – 5 not at all easily"

IS 6: How much do you like to draw attention to yourself?

"1 I like it extremely – 5 I don't like it at all"

AGREE/DISAGREE (Single Questions; 7-Point Response Scales)

A/D 7: I am quiet around strangers.

A/D 8: I don't trust people in general.

A/D 9: I am always looking for new things to do.

A/D 10: It is important for me to live in secure surroundings.

A/D 11: I avoid anything that can endanger my safety.

A/D 12: It is important to be free and not depend on others.

Response categories to A/D 1 – A/D 6 are "1 agree strongly – 7 disagree strongly"

ITEM-SPECIFIC (Single Questions; 7-Point Response Scales)

IS 7: How quiet are you around strangers?

"1 extremely quiet – 7 not at all quiet"

IS 8: How much do you trust people in general?

"1 I completely trust – 7 I don't trust at all"

IS 9: How often are you looking for new things to do?

"1 always – 7 never"

IS 10: How important is it for you to live in secure surroundings?

"1 extremely important – 7 not at all important"

IS 11: How often do you avoid anything that can endanger your safety?

"1 always – 7 never"

IS 12: How important is it to be free and not depend on others?

"1 extremely important – 7 not at all important"

Note. The order of the questions was randomized within the 5- and 7-point response scales, respectively, to avoid question order effects. The original questions, including response categories, were formulated in Spanish. The original Spanish wordings of the questions can be retrieved from *ww2.netquest.com/respondent/glinn/mobile2016*. Abbreviations: A/D = agree/disagree; IS = item-specific.

Appendix E

AGREE/DISAGREE QUESTIONS (Grid Questions)

A/D G1: I like being in competition with other people. (A)

A/D G2: It is satisfying when I achieve better results than other people. (A)

A/D G3: I endeavor to improve my performance. (A)

A/D G4: I try harder when I am in competition with other people. (A)

A/D G5: It is important to me to be the best at a task. (A)

Response categories to A/D G1 – A/D G5 are "agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, disagree strongly"

ITEM-SPECIFIC QUESTIONS (Single Questions)

IS S1: To what extent do you enjoy competing with other people? (A)

"very much, fairly, somewhat, hardly, not at all"

IS S2: How satisfying is it to you to achieve better results than other people? (A)

"very satisfying, fairly satisfying, somewhat satisfying, hardly satisfying, not at all satisfying"

IS S3: How important is it to you to endeavor to improve your performance? (A)

"very important, fairly important, somewhat important, hardly important, not at all important"

IS S4: How much harder do you try when you compete with other people? (A)

"very much harder, fairly harder, somewhat harder, hardly harder, not at all harder"

IS S5: How important is it to you to be the best at a task? (A)

"very important, fairly important, somewhat important, hardly important, not at all important"

AGREE/DISAGREE QUESTIONS (Grid Questions)

A/D G6: A job that I can work autonomously on is important to me. (I)

A/D G7: A job that allows to make use of my skills and talents is important to me. (I)

A/D G8: A job where I have responsibilities for specific tasks is important to me. (I)

A/D G9: A job that allows me to implement my own ideas is important to me. (I)

A/D G10: A job with a high income is important to me. (E)

A/D G11: A job with good promotion prospects is important to me. (E)

A/D G12: A job with clear career perspectives is important to me. (E)

Response categories to A/D G6 – A/D G12 are "agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, disagree strongly"

ITEM-SPECIFIC (Grid Questions)

IS G1: How important is a job that you can work autonomously on? (I)

IS G2: How important is a job that allows you to make use of your skills and talents? (I)

IS G3: How important is a job where you have responsibilities for specific tasks? (I)

IS G4: How important is a job that allows to implement your own ideas? (I)

IS G5: How important is a job with a high income? (E)

IS G6: How important is a job with good promotion prospects? (E)

IS G7: How important is a job with clear career perspectives? (E)

Response categories to IS G1 – IS G7 are "very important, fairly important, somewhat important, hardly important, not at all important"

RESPONDENTS' EVALUATIONS

Please indicate, how did you perceive responding to the survey questions?

"1 interesting – 7 boring"

"1 undemanding – 7 demanding"

"1 inspiring – 7 tedious"

"1 simple – 7 complex"

"1 diversified – 7 monotonous"

Note. The order of the questions corresponds to the presentation in Appendix E (it only contains decremental response scale directions). In IS single questions, response categories were horizontally arranged below each question. In A/D and IS grid questions, response categories were horizontally arranged alongside each question. The original German wordings of the questions are available from the author (hoehne@uni-mannheim.de) upon request. Abbreviations: A/D = agree/disagree; IS = item-specific; S = single questions; G = grid questions; A = achievement motivation; I = intrinsic job motivation; E = extrinsic job motivation.