Scale Direction Effects in Agree/Disagree and Item-Specific Questions: A Comparison of Question Formats

Jan Karem Höhne
University of Göttingen (Germany)

Dagmar Krebs
University of Gießen (Germany)

Abstract

The effect of the response scale direction on response behavior is a well-known phenomenon in survey research. While there are several approaches to explaining how such response order effects occur, the literature reports mixed evidence. Furthermore, different question formats seem to vary in their susceptibility to these effects. We therefore investigate the occurrence of response order effects in Agree/Disagree (A/D) and Item-Specific (IS) questions. We conducted an experiment among *n* = 930 students in which we varied the scale direction (decremental vs. incremental) within A/D and IS questions and asked respondents to evaluate the questionnaires. The results reveal response order effects within the A/D but not within the IS question format. Furthermore, respondent's evaluations suggest that completion of the IS questionnaires requires more consideration than the completion of the A/D questionnaires. Altogether, our findings indicate that IS questions are more robust against response order effects than A/D questions.

Keywords: asking manner, data quality, survey experiment, latent mean analysis, response behavior, response order effects

Introduction

In empirical social research, the use of "closed-ended" question formats to measure the attitudes and opinions of respondents is highly popular. Closed-ended implies that the response categories of a survey question are entirely specified (Lavrakas, 2008). However, during the development of such questions, researchers must make several decisions: first of all, they must set the question format, e.g., employing agree/disagree (A/D) or item-specific (IS) questions. Then, they must decide whether the response scale contains a midpoint (using an even or uneven number of categories) and determine the scale length (specifying the number of scale points). Afterwards, they must decide to what extent the response scale is numerically and/or verbally labeled (employing fully or partially labeled scales). Finally, they must specify 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).

In this study, the first and last point are of interest since methodological research only recently deals with comparisons between A/D and IS questions (see Höhne, Schlosser, & Krebs,

This document is a preprint and thus it may differ from the final version: 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. DOI: 10.1080/13645579.2017.1325566.

forthcoming; Kuru & Pasek, 2016; Lelkes & Weiss, 2015; Liu, Lee, & Conrad, 2015; Saris et al., 2010). While the A/D question format (e.g., "Before an election, I inform myself thoroughly about the suitability of the candidates" – "agree strongly" to "disagree strongly") is characterized by an indirect statement, which is evaluated on an agreement/disagreement response scale, the IS question format (e.g., "How thoroughly do you inform yourself about the suitability of the candidates before an election?" – "very thoroughly" to "not at all thoroughly") is characterized by a direct question, which is evaluated on a tailored response scale. Although there are several studies comparing extreme response styles (Liu, Lee, & Conrad, 2015) as well as acquiescence response bias (Kuru & Pasek, 2016; Lelkes & Weiss, 2015; Saris et al., 2010) between A/D and IS questions, studies comparing response order effects between the two question formats are missing.

The pertinent survey literature on scale direction effects indicates a somewhat different situation. Since Mathews (1929), many studies have shown that the direction of the response scale affects the response behavior (Höhne & Lenzner, 2015; Krebs, 2012; Krebs & Hoffmeyer-Zlotnik, 2010; Rammstedt & Krebs, 2007; Bishop & Smith, 2001; Malhotra, 2008; Rugg & Cantrill, 1942; Yan & Keusch, 2015). This type of response bias is called response order effect either occurring as primacy or recency effect. If categories at the beginning of a scale are selected more often, one speaks of primacy effects. If categories at the end of a scale are selected more often, one speaks of recency effects. The evidence on response order effects is rather mixed. Consultation of handbooks on methodology (Bradburn, Sudman, & Wansink, 2004; Fowler, 1995; Robinson, Shaver, & Wrightsman, 1999; Sudman, Bradburn, & Schwarz, 1996) reveals that there is no scientific consensus regarding the direction of response scales. Hence, it is more or less up to the researchers to decide between a decremental or incremental response order.

There are several explanations for the occurrence of scale direction effects. Some scientists argue that they are the result of the difficulty of a question itself (Bishop & Smith, 2001). Schuman and Presser (1996), in contrast, show that response order effects occur in short and simple survey questions as well. Therefore, question difficulty per se does not appear to be exhaustive as an explanation. According to the *opinion crystallization hypothesis* (Rugg & Cantrill, 1942), response order effects are the consequence of weak attitudes or opinions. Bishop (1990), however, found that effects of response order are only slightly related to the issue involvement. Additionally, response order effects seem to be associated with the survey mode. While primacy effects are mainly observed in visual presentation modes (e.g., web surveys), recency effects are mainly observed in auditory presentation modes (e.g., telephone surveys) (Krosnick & Alwin, 1987). The size of response order effects is additionally conditioned by the scale format itself because empirical evidence generally shows smaller effects for rating than categorical scales (Sudman et al., 1996). Furthermore, Höhne et al. (forthcoming) suggest that the occurrence of response bias such as response order effects might also depend on the question format used.

For this reason, we conducted an experimental study to compare the occurrence of response order effects within A/D and IS questions. In addition, we measured respondents' evaluations of the manageability and motivating potential of the A/D and IS questionnaires using pairs of opposite adjectives. The goal of gathering these evaluations is to investigate whether respondent’s perceptions are in line with Fowlers (1995) suggestion that IS questions are simpler and more direct than A/D questions.

We first discuss several (theoretical) approaches explaining the occurrence and possible causes of response order effects. Based on these approaches, we outline our research hypotheses. Afterwards, we describe the survey instruments, the data collection process (including study design), the underlying sample, and the data analysis strategy. Then we present the results of our study, and finally, we discuss the implications of our findings and suggest perspectives for further research.

Background

Since the early contributions of Mathews (1929) as well as Rugg and Cantrill (1942), the phenomenon of response order effects has been the object of studies in survey research. Generally, this response bias is caused by the order of the response categories (decremental or incremental) and can be divided up into primacy and recency effects. While primacy effects refer to higher endorsements of response categories presented early in the list, recency effects refer to higher endorsements of response categories presented late in the list (Sudman et al., 1996). The reasons for the occurrence of response order effects in surveys have not, as of yet, been conclusively elucidated. There are several (divergent) theoretical approaches that attempt to explain their origins. These include *memory limitations*, *cognitive elaboration*, and *satisficing*.

Following the *memory limitation hypothesis* suggested by Smyth et al. (1994), response order effects are simply a consequence of memory limitations. This, in turn, suggests that respondents are not capable of recalling or remembering all of the response categories. However, this explanation seems only reasonable when respondents must process a considerable amount of information presented in a fast sequence without visual aids (e.g., CATI mode), and only accounts for the occurrence of recency effects. Hence, the *memory limitation hypothesis* provides no explanation for the emergence of primacy effects and thus is of little avail to explaining response order effects in surveys (Sudman et al., 1996).

The *cognitive elaboration model* postulated by Schwarz et al. (1992) argues that response order effects are specifically caused by an interaction between the serial position of a response category (i.e., whether it is at the beginning, the middle, or the end) and the presentation mode (i.e., visual or auditory). Based on these conditions, the *cognitive elaboration model* postulates the following assumptions (Sudman et al., 1996): first, if response categories are displayed visually, primacy effects are predicted, since visual presentation makes the elaboration of response categories at the beginning of a list easier. Second, if response categories are provided auditorily, then recency effects are predicted, since auditory presentation makes the elaboration of response categories at the end easier. However, the *cognitive elaboration model* does not take memory limitations, complexity of response categories, and respondents’ (cognitive) abilities and motivation into account (Bishop & Smith, 2001).

Finally, the *satisficing theory* proposed by Krosnick (1991) primarily makes a distinction between optimizing and satisficing. While optimizing implies conscientious responding (i.e., considerate accomplishment of cognitive processing), satisficing implies a superficial one (i.e., perfunctory cognitive processing). The likelihood of the occurrence of *satisficing* is based on task difficulty, respondents' (cognitive) abilities, and motivation. In line with the *cognitive elaboration model*, the *satisficing theory* postulates that visual presentation of response categories fosters primacy effects and auditory presentation of response categories fosters recency effects. Regarding primacy effects, Krosnick (1991) suggests that respondents either select the first acceptable response category – not bothering to consider the later ones – or they are not capable of processing all the response categories equally, which leads to preferential selection of the initial ones. The reasons for recency effects are more intricate because they are based on an interplay between the (mental) processing of response categories and memory limitations. Due to a (rapid) auditory presentation of information, respondents pay more attention to later response categories and are thus more likely to choose these.

As regards rating scales, respondents do not need to process all of the underlying response categories to find the proper one, as is the case with categorical scales. For instance, rating scales running from a positive to a negative end or vice versa represent an ordered and closed response continuum. For this reason, response categories of rating scales require less processing effort than those of categorical scales because respondents can (mentally) extrapolate the response continuum (Höhne & Lenzner, 2015). Furthermore, it is well-documented in the literature (Höhne & Lenzner, 2015; Krosnick, 1991; Sudman et al., 1996; Yan & Keusch, 2015) that rating scales tempt respondents to choose response categories appearing at the beginning (i.e., on the left or top half of the scale).

Research Hypotheses

As suggested by recent research (Kuru & Pasek, 2016; Lelkes & Weiss, 2015; Liu et al., 2015; Saris et al., 2010), 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 (i.e., they provide statements for which a placement on an agreement/disagreement continuum is required) so that respondents must repeat the same answering task for all questions. Höhne et al. (forthcoming) argue that due to this unchanging "asking manner" A/D questions foster boredom and weariness and therefore a perfunctory (cognitive) response process. In addition, responses to A/D statements do not refer directly to the underlying dimension of interest (e.g., "thoroughness of informing"). This implies an indirect manner of asking, 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 change the asking manner from question to question, because response categories match the underlying content dimension. They require that respondents continuously reconsider the underlying dimension of interest, inciting 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 asking manner due to an interrogation, which refers directly to the dimension of interest and thus might support quality of responses. The direct asking manner presumably deters respondents from automatic responding so that the IS question format seems to be more complex.

In line with this reasoning, we postulate the following 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*).

Method

Survey Instruments

The questions used were adapted from the *Cross Cultural Survey for Work and Gender Attitudes* (2010). Taking questions from established social surveys offers the advantage of using repeatedly tested questions. For our 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 survey, we developed an IS counterpart that preserved the original question content as well as possible.[1] All questions were presented with a 5-point, fully-verbalized response scale and no numeric values (see Appendix for the questions used).

Extrinsic job motivation refers to the importance of anticipated job characteristics (e.g., income and career), as they are (generally) 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", 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 over a couple of years (see, for instance, 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. At the end, respondents were also asked to evaluate the manageability and motivating potential of the questionnaire using opposite adjectives pairs.[2]

Data Collection

The research was conducted at the *author's institution* in the winter terms of 2014 and 2015. Both years, the experimental study was conducted in the first week 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 as well as informed that they would be participating in a study dealing

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

[2] These pairs followed the principle of a semantic differential (see Osgood, Succi, & Tannenbaum, 1957) and thus were measured on seven-point response scales.

with different survey research topics and that their data would be treated confidentially. According to the split versions (A/D vs. IS), paper questionnaires were sorted systematically before their distribution to ensure random assignment. Completing the questionnaire took approximately 10 minutes. 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 data 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 1 outlines our study design.

Table 1. Study design defined by question format and scale direction

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

Notes. A/D: Agree/Disagree; IS: Item-Specific. The decremental scales run from positive to negative and the incremental scales vice versa.

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 as well as position of all questions were identical. Only the response scale direction was changed.

Sample

Altogether, n = 976 students participated in our study (n = 435 in 2014 and n = 541 in 2015). However, we excluded all respondents older than 30 years[3] and students participating in 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 and 85% in 2015, respectively).

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 ($\chi^2(1) = .75$, $p = .39$). With respect to these characteristics, the sample composition across the four experimental groups did not reveal any statistically significant differences. Furthermore, the two scale direction groups did not differ with respect to these demographic characteristics within either question format (A/D and IS). Thus, the two groups can be considered comparable.

[3] The threshold of 30 years is based on an outlier definition using the mean plus/minus the standard deviation multiplied by 4.

Data Analysis

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.[4] Due the fact that the indicators of the latent variables were measured on 5-point response scales, we assumed continuous scale level (see Rhemtulla et al., 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)[5] and unpaired t-tests using SPSS version 24.

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 in order to test whether the IS questionnaires are evaluated differently from the A/D questionnaires.

Descriptive Statistics

In order to ensure appropriate statistical testing, all questions were recoded to identical values from 1 "positive" to 5 "negative". To investigate the occurrence of scale direction effects in the A/D as well as IS question format, we calculated means, standard deviations, skewness, and kurtosis for all questions. Considering the results in table 2, 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 as well as 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.

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 as well as kurtosis are quite small, indicating a relatively normal distribution of respondents' answers to A/D and IS questions.

[4] Please see Appendix B for the Mplus commands for testing the invariance of latent means.

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

Table 2. Means, standard deviations, skewness, and kurtosis for decremental and incremental scale directions within the AD and IS question format

	A/D Format							
	Decremental Order				*Incremental Order*			
Questions	Mean	SD	Skewness	Kurtosis	Mean	SD	Skewness	Kurtosis
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	1.00	.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 Format							
Competition	3.01	.95	.08	-.08	3.10	.90	.23	-.40
Achievement	1.96	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

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

Measurement Equivalence

In a first step, we conducted confirmatory factor analyses (CFA) within the two question formats (A/D and IS) and formulated separate yet identical baseline models for each scale direction (decremental and incremental). The CFA model contained three correlated latent variables for 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. Using multi-group confirmatory factor analysis (MGCFA), we tested configural invariance (M_0) by analyzing the baseline model simultaneously for the two scale directions within each question format; see table 3 for the statistical results. Given CFI-values higher than .95 and RMSEA-values lower than .05, configural invariance was accepted for response scale direction within the A/D as well as 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.

Table 3. 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

Notes. An error covariance between two achievement motivation questions (“competition” and “making an effort”) was admitted. The results are based on MLR estimation. Scale correction factors (to calculate $\Delta\chi^2$ difference tests) for model comparison are in parentheses.

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 3. 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.

Differences in Latent Means

In testing the differences in latent means between response scale directions (decremental vs. incremental), we used the group with an incremental order of response categories as reference group. 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 4 shows the results of the comparison of latent means for the A/D question format (χ^2(118) = 163.91 (1.07); CFI = .974; RMSEA = .040). 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. We return to this point in the discussion section.

Table 4. Latent mean differences between decremental and incremental response order in the A/D and IS question format (unstandardized results).

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

Notes. Response scales were recoded to identical values from 1 “positive” to 5 “negative”. Reference group is the incremental response order.

In contrast, for the IS question format (see also table 4), the model testing for invariance of latent means ($\chi^2(118) = 156.51$ (1.07); CFI = .975; RMSEA = .038) 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 our expectation that the A/D question format is more susceptible to response order effects than the IS question format, as suggested by *hypothesis 1*.

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 assess these evaluations, we used five pairs of opposite adjectives and a seven-point response scale. 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, which resulted in a two-factor solution: the first factor describes the task-oriented manageability (i.e., demanding and complex) and the second factor describes the 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, $\lambda > .80$, and r = .30; incremental: S = .75, $\lambda > .70$, and r = .12). Irrespective of response scale direction, the evaluations of the questionnaires are almost identical. Table 5 displays the overall average ratings of the A/D and IS questionnaires. As an indicator of the effect sizes, 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 *hypotheses 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 5. Means and standard deviations (in parentheses) of respondents' evaluations of the A/D and IS questionnaires

Adjective Pairs	A/D	IS	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

Notes. 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.

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 as well as motivational potential, under the assumption that the IS ones would be evaluated as being simpler and more motivating than the A/D ones. 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 and thus their impression is not in line with the common notion of "simplicity". 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 questionnaire design (more precisely, by 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 in the IS question format had already been demonstrated by Krebs & Hoffmeyer-Zlotnik (2010), hence, the results of the present study add to their finding revealing that this kind of motivation is not affected by response order in the A/D question format either. This result seems to be related to the specific content of this motivational dimension since 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 Höhne et al. (forthcoming) found no differences in response quality (e.g., speeding and dropouts) 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 asking manner of the IS question format, which might elicit more attention and thus prevent the occurrence of response order effects. 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 to responding to IS than to A/D questions and, most importantly, they seem to perceive this fact. According to the concept of asking manner, 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 question 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, our 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 are in line with results of former studies 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, our 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 previous research (see Saris et al., 2010) as well as our results, 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.

References

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

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

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 and Sons.

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

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling, 9,* 233–255.

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

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

Höhne, J. K., Schlosser, S., & Krebs, D. (forthcoming). Investigating Cognitive Effort and Response Quality of Question Formats in Web Surveys Using Paradata. *Field Methods.*

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

Krebs, D., 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.

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

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

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

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.

Malhotra, N. (2008). Completion Time and Response Order Effects in Web Surveys. *Public Opinion Quarterly*, *72*, 914–934.

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

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

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

Robinson, J.P., Shaver, P.R., & Wrightsman, L.S. (1999). *Measures of Political Attitudes.* San Diego, CA: Academic Press.

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.

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

Rugg, D., & Cantril, H. (1941). The Wording of Questions in Public Opinion Polls. *Public Opinion Quarterly, 5,* 52–78.

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 Methods*, 4, 61–79.

Schuman, H., & Presser, S. (1996). *Questions and Answers in Attitude Surveys. Experiments on Question Form, Wording, and Context.* Thousand Oaks, CA: Sage.

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

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

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.

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.

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.

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

Appendix A

English translation of the original German A/D and IS questions (decremental scale direction).

AGREE/DISAGREE QUESTIONS
I like being in competition with other people. (Achievement)
It is satisfying when I achieve better results than other people. (Achievement)
I endeavor to improve my performance. (Achievement)
I try harder when I am in competition with other people. (Achievement)
It is important to me to be the best at a task. (Achievement)
A job that I can work autonomously on is important to me. (Intrinsic)
A job that allows to make use of my skills and talents is important to me. (Intrinsic)
A job where I have responsibilities for specific tasks is important to me. (Intrinsic)
A job that allows me to implement my own ideas is important to me. (Intrinsic)
A job with a high income is important to me. (Extrinsic)
A job with good promotion prospects is important to me. (Extrinsic)
A job with clear career perspectives is important to me. (Extrinsic)

agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, disagree strongly.[6]

ITEM-SPECIFIC QUESTIONS
To what extent do you enjoy competing with other people? (Achievement)
Very much, fairly, somewhat, hardly, not at all
How satisfying it is to you to achieve better results than other people? (Achievement)
Very satisfying, fairly satisfying, somewhat satisfying, hardly satisfying, not at all satisfying
How important is it to you to endeavor to improve your performance? (Achievement)
very important, fairly important, somewhat important, hardly important, not at all important
How much harder do you try when you compete with other people? (Achievement)
Very much harder, fairly harder, somewhat harder, hardly harder, not at all harder
How important is it to you to be the best at a task? (Achievement)
very important, fairly important, somewhat important, hardly important, not at all important

How important is a job that you can work autonomously on? (Intrinsic)
How important is a job that allows you to make use of your skills and talents? (Intrinsic)
How important is a job where you have responsibilities for specific tasks? (Intrinsic)
How important is a job that allows to implement your own ideas? (Intrinsic)
How important is a job with a high income? (Extrinsic)
How important is a job with good promotion prospects? (Extrinsic)
How important is a job with clear career perspectives? (Extrinsic)
very important, fairly important, somewhat important, hardly important, not at all important

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

Appendix B

Mplus commands to track the analyzes of measurement equivalence and latent means.

```
VARIABLE:
   NAMES ARE v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 version;
   USEVARIABLES ARE v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 version;
   GROUPING IS version (1 = AD15np 3 = AD15pn);

ANALYSIS:
   ESTIMATOR IS MLR;

MODEL AD15np:
   F1 by v1 v2 v3 v4 v5;
   F2 by v6 v7 v8 v9;
   F3 by v10 v11 v12;
   v1 WITH v4;
   [F1-F3@0];

MODEL AD15pn:
  F1-F3*;
  [F1-F3];
```

The variables v1 to v5 refer to achievement motivation, v6 to v9 refer to intrinsic job motivation, and v10 to v12 refer to extrinsic job motivation (see question order in Appendix A).