

Rating Scales in Web Surveys

A Test of New Drag-and-Drop Rating Procedures

INAUGURALDISSERTATION

zur Erlangung des Grades eines Doktors der Philosophie (Dr. phil.) im
Fachbereich Gesellschafts- und Geschichtswissenschaften an der Technischen
Universität Darmstadt

Genehmigte Dissertation von:

Diplom-Sozialwissenschaftlerin Tanja Kunz aus Eberhardzell

Referenten:

Prof. Dr. Marek Fuchs, Technische Universität Darmstadt

Prof. Dr. Mark Trappmann, Institut für Arbeitsmarkt- und Berufsforschung

Tag der Einreichung: 06.10.2014

Tag der mündlichen Prüfung: 15.12.2014

Darmstadt 2015

D17

Please cite this document as:

Kunz, T. (2015). Rating scales in Web surveys. A test of new drag-and-drop rating procedures. Doctoral Dissertation. Darmstadt University of Technology.

urn:nbn:de:tuda-tuprints-51512

<http://tuprints.ulb.tu-darmstadt.de/id/eprint/5151>

This document is provided by tuprints, E-Publishing-Service of the TU Darmstadt

<http://tuprints.ulb.tu-darmstadt.de>

tuprints@ulb.tu-darmstadt.de



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Germany License.

<http://creativecommons.org/licenses/by-nc-nd/3.0/de/deed.en>

ABSTRACT

In Web surveys, rating scales measuring the respondents' attitudes and self-descriptions by means of a series of related statements are commonly presented in grid (or matrix) questions. Despite the benefits of displaying multiple rating scale items neatly arranged and supposedly easy to complete on a single screen, respondents are often tempted to rely on cognitive shortcuts in order to reduce the extent of cognitive and navigational effort required to answer a set of rating scale items. In order to minimize this risk of cognitive shortcuts resulting in satisfying rather than optimal answers, respondents have to be motivated to spend extra time and effort on the attentive and careful processing of rating scales.

A wide range of visual and dynamic features are available in interactive Web surveys allowing for visual enhancement and greater interactivity in the presentation of survey questions. To date, however, only a few studies have systematically examined new rating scale designs using data input methods other than conventional radio buttons. In the present study, two different rating scales were designed using drag-and-drop as a more interactive data input method: Respondents have to drag the response options towards the rating scale items ('drag-response'), or in the reverse direction, the rating scale items towards the response options ('drag-item').

In both drag-and-drop rating scales, the visual highlighting of the items and response options as well as the dynamic strengthening of the link between these key components are aimed at encouraging the respondents to process a rating scale more attentively and carefully. The effectiveness of the drag-and-drop rating scales in preventing the respondents' susceptibility to cognitive shortcuts is assessed on the basis of five systematic response tendencies that are typically accompanied by rating scales, i.e., careless, nondifferentiated, acquiescent, and extreme responding as well as the respondents' systematic tendency to select one of the first response options, so called primacy effects. Moreover, item missing data, response times, and respondent evaluation are examined.

The findings of the present study revealed that although both drag-and-drop scales entail a higher level of respondent burden as indicated by an increase in item missing data and longer response times compared to conventional radio button scales, they promote the respondents' attentiveness and carefulness towards the response task which is accompanied by the respondents' reduced susceptibility to cognitive shortcuts in processing rating scales.

ZUSAMMENFASSUNG

Ratingskalen sind gewissermaßen ein fester Bestandteil standardisierter Befragungen. Zumeist werden sie eingesetzt, um Einstellungen, Meinungen, Interessen sowie Persönlichkeitsmerkmale des Befragten mittels einer mehrstufigen Skala zu erfassen. In selbst-administrierten Befragungen werden die zu bewertenden Fragen oder Aussagen (Items) einer Ratingskala bevorzugt in Form einer Matrixfrage (Grid) dargestellt: Die zu bewertenden Items werden zeilenweise aufgeführt, während sich die Antwortmöglichkeiten der mehrstufigen Skala in den Spalten befinden. Matrixfragen bieten gewisse Vorzüge hinsichtlich der übersichtlichen Darstellung und schnellen Bearbeitung mehrerer Items. Gleichzeitig sind sie jedoch auch anfälliger für systematische Antworttendenzen (z.B. Tendenz zu gleichförmigen Urteilen, Ja-Sage-Tendenz) und die Nichtbeantwortung einzelner Items (Krosnick & Alwin, 1988; Tourangeau, et al., 2004).

Online-Befragungen werden selbst-administriert durchgeführt, d.h. in der jeweiligen Befragungssituation ist kein Interviewer anwesend, der bei gegebenem Anlass Hilfestellung anbieten und den Befragten zur sorgfältigen Bearbeitung der Fragen motivieren könnte. Daher liegt es insbesondere an der Fähigkeit und Motivation eines jeden Befragten, vollständige und exakte Antworten zu geben. Die Fähigkeit und Motivation des Befragten wiederum wird maßgeblich von der Schwierigkeit der Fragen beeinflusst (Krosnick & Alwin, 1987). Aus diesem Grund wird in selbst-administrierten Befragungen der Konstruktion des Fragebogens eine besondere Bedeutung beigemessen. Insbesondere gilt, dass die Fragen gut verständlich und einfach zu beantworten sein sollen, um den kognitiven Aufwand des Befragten möglichst gering zu halten. Zusätzlich zu den verbalen Gestaltungsmerkmalen eines Fragebogens können in Online-Befragungen visuelle und interaktive Elemente eingesetzt werden, um die Frageschwierigkeit zu reduzieren (Beatty & Herrmann, 2002; Krosnick, 1991). So können beispielsweise Begriffserklärungen im Bedarfsfall dynamisch eingeblendet werden, um das Frageverständnis zu erhöhen und dem Befragten die Beantwortung der Frage zu erleichtern (Conrad, et al., 2007; Derouvray & Couper, 2002; Yan, et al., 2011). Dynamisch eingeblendete farbliche Hervorhebungen können ebenfalls gezielt eingesetzt werden, um den Befragten auf zentrale Bestandteile einer Frage aufmerksam zu machen und die Orientierung auf einer Fragebogenseite zu erleichtern (Couper, et al., 2013; Kaczmirek, 2010). Darüber hinaus können visuelle und dynamische Elemente genutzt werden, um den Fragebogen optisch aufzuwerten

und das Ausfüllen eines Fragebogens zu einem positiven und interessanten Erlebnis für den Befragten zu machen (Sikkel, et al., 2014; Stanley & Jenkins, 2007).

Für die Umfrageforschung hat selbstverständlich das Erzielen einer hohen Datenqualität auch in Online-Befragungen höchste Priorität. Dabei ist die Datenqualität maßgeblich von der Motivation des Befragten abhängig, vollständige und exakte Antworten zu geben. Diese—zunächst trivial erscheinende Prämisse—lenkt den Blick auf die verschiedenen Phasen des sog. kognitiven Frage-Antwort-Prozesses, den die Befragten bei der Beantwortung der Fragen durchlaufen. Diese Phasen umfassen (1) Verstehen und Interpretation der Frage, (2) Abruf relevanter Informationen, (3) Generieren einer Antwort, sowie (4) Editieren und Zuordnung der Antwort zu einer Antwortvorgabe (Cannell, et al., 1981; Sudman, et al., 1996, p. 56ff; Tourangeau, et al., 2000, p. 165ff). Jedoch ist der Befragte nicht immer in der Lage und/oder ausreichend motiviert, diese vier Phasen vollständig und mit der nötigen Sorgfalt zu durchlaufen. Satisficing beschreibt ein Antwortverhalten, bei dem anstatt nach einer optimalen Antwort zu suchen, der kognitive Frage-Antwort-Prozess von dem Befragten frühzeitig abgebrochen wird, sobald nach eigener Einschätzung eine zufriedenstellende Antwort erzielt wurde. Die Abkürzung der zur optimalen Antwort erforderlichen kognitiven Schritte ist in Online-Befragungen zumeist dadurch begründet, dass der Befragte den Aufwand für Kognition und Navigation möglichst gering halten möchte (Krosnick, 1991, 1999).

Matrixfragen sind besonders anfällig für derartige Abkürzungsstrategien, wobei dies häufig auf Ermüdungserscheinungen („fatigue“) oder Unzufriedenheit/Missmut („frustration“) der Befragten zurückgeführt wird (Krosnick, 1991; Wenemark, et al., 2010). Genau an diesem Punkt setzt die vorliegende Studie an. Zwei unterschiedliche Drag-and-Drop-Ratingskalen wurden mit dem Ziel konzipiert, den monotonen Charakter einer Ratingskala im Allgemeinen und einer Matrixfrage im Speziellen zu durchbrechen und den Befragten zu einem aufmerksameren und sorgfältigeren Antwortverhalten zu motivieren: In der Drag-Response-Skala waren die Befragten aufgefordert, mit dem Mauszeiger eine ausgewählte Antwortmöglichkeit zum jeweiligen Item zu ziehen, wohingegen in der Drag-Item-Skala das jeweilige Item zur ausgewählten Antwortmöglichkeit gezogen werden sollte. Durch den Einsatz der Drag-and-Drop Technik und die damit gesteigerte kognitive Stimulation soll die Motivation des Befragten gesteigert werden, die nötige kognitive Anstrengung für eine aufmerksame und sorgfältige Bearbeitung einer Ratingskala aufzuwenden. Visuelle und interaktive Gestaltungselemente wurden hierbei gezielt eingesetzt, um die Aufmerksamkeit des Befragten auf die Items (Drag-Item-Skala) oder die Antwortmöglichkeiten (Drag-Response-Skala) zu lenken und die Verbindung zwischen dem jeweiligen Item und der ausgewählten Antwortmöglichkeit zu verstärken.

Zur Überprüfung der Effektivität der beiden Drag-and-Drop-Ratingskalen hinsichtlich einer aufmerksameren und sorgfältigeren Bearbeitung und letztlich einer Vorbeugung von systematischen Antworttendenzen wurden insgesamt sechs Experimente basierend auf einem between-subjects Design durchgeführt. Das Hauptaugenmerk lag auf der experimentellen Variation des Skalenformats und der Skalenlänge. Neben den beiden Drag-and-Drop-Ratingskalen wurden eine konventionelle Matrixfrage und zwei weitere Single-Item-Ratingskalen eingeführt. Zusätzlich wurden drei Skalenlängen unterschieden, eine Batterie aus 6 Items, 10 Items oder 16 Items. Zur Überprüfung der Effektivität beider Drag-and-Drop-Ratingskalen im Vergleich zu den konventionelleren Formaten wurden mehrere Indikatoren der Datenqualität herangezogen. Darunter fielen unterschiedliche systematische Antworttendenzen (Nichterkennen „gedrehter“ Items oder „careless responding“, Tendenz zu gleichförmigen Urteilen oder „nondifferentiation“, Ja-Sage-Tendenz oder „acquiescence“, Tendenz zu extremen Urteilen oder „extremity“, Reihenfolgeeffekte oder „primacy effects“) sowie die Nichtbeantwortung der Items („item nonresponse“), der Abbruch der Befragung („survey breakoff“) und die Antwortzeiten der Befragten („response times“).

Die Ergebnisse der vorliegenden Studie zeigten, dass systematische Antworttendenzen maßgeblich durch den Einsatz der Drag-and-Drop-Ratingskalen beeinflusst werden können. Dies bezog sich jedoch lediglich auf systematische Antworttendenzen, die im kognitiven Frage-Antwort-Prozess der zweiten und dritten Phase zugeordnet werden können: Während in der Drag-Response-Skala dem Auftreten von Reihenfolgeeffekten vorgebeugt werden konnte, konnte mittels der Drag-Item-Skala der Tendenz zu gleichförmigen Urteilen entgegengewirkt werden. Ein Einfluss der Skalenlänge war nur bedingt festzustellen. Gleichzeitig erhöhte sich das Risiko der Nichtbeantwortung aller oder zumindest einzelner Items in beiden Drag-and-Drop-Ratingskalen deutlich im Vergleich zu einer herkömmlichen Matrixfrage oder Single-Item-Ratingskala.

Diese Ergebnisse zeigen zum einen, dass durch das gezielte Lenken der Aufmerksamkeit auf die Items und Antwortmöglichkeiten einer Ratingskala mittels visueller und dynamischer Gestaltungselemente der Befragte dazu angehalten wird, sich bei jedem Item von der Angemessenheit der ausgewählten Antwortmöglichkeit zu überzeugen, um für jedes der Items eine optimale Antwort zu finden. Die Wahl einer optimalen Antwort erfolgt selbst dann, wenn dies mit einem erhöhten Navigationsaufwand verbunden ist. Zum anderen zeigen diese Ergebnisse jedoch auch, dass zumindest für einige der Befragten der Aufwand für Kognition und Navigation zu hoch ist, um sich überhaupt auf die Beantwortung einer Frage einzulassen. Dies verweist auf einen schmalen Grad zwischen einerseits den positiven Auswirkungen einer gesteigerten kognitiven Beanspruchung, die sich für

die aufmerksame und sorgfältige Bearbeitung einer Frage als förderlich erweist und den negativen Folgen einer kognitiven Überbeanspruchung andererseits, die sich schlimmstenfalls in der Nichtbeantwortung einer Frage zeigt. Demnach gilt es, das optimale Maß der Aufgabenschwierigkeit zu finden, um der Ermüdung des Befragten entgegenzuwirken ohne jedoch Unzufriedenheit und Missmut eines Befragten hervorzurufen.

ACKNOWLEDGMENTS

Besonderen Dank möchte ich Prof. Dr. Marek Fuchs aussprechen, für die fachliche Betreuung und Unterstützung in allen Phasen der Entstehung dieser Dissertation, sowie für die zahlreichen Möglichkeiten, meine Arbeiten in der Forschergemeinschaft vorstellen und mich wissenschaftlich weiterentwickeln zu können. Darüber hinaus möchte ich Prof. Dr. Mark Trappmann für die Co-Betreuung meiner Dissertation, sowie Prof. Dr. Helmuth Berking und Prof. Dr. Andrea Rapp für ihre Unterstützung im Rahmen meiner Disputation danken.

Hela Oueslati und Kevin Schaller danke ich für die technische Umsetzung der Drag-and-Drop Programmierungen. Allen Kolleginnen und Kollegen, denen ich im Verlauf meiner Beschäftigung an der TU Darmstadt begegnen durfte, danke ich für die zahlreichen konstruktiven Anregungen, die tolle gemeinsame Zeit und wichtige Unterstützung über all die Jahre.

Mein besonderer Dank gilt meinem Freund Matthias, der mir in allen Phasen mit kollegialem Rat, unzähligen Taten und rückhaltloser Unterstützung zur Seite stand. Seine Zuversicht und sein Vertrauen in mich sind von unschätzbarem Wert. Herzlichst danken möchte ich auch allen meinen Freunden, die mich immer ermutigt und auf andere Gedanken gebracht haben. Nicht zuletzt möchte ich meiner Familie von ganzem Herzen danken, die immer an mich geglaubt und mich mein ganzes Leben in allem unterstützt hat.

TABLE OF CONTENTS

Abstract.....	iii
Zusammenfassung	v
Acknowledgments	ix
List of Tables.....	xvi
List of Figures.....	xx
List of Appendices.....	xxi
1. Introduction	1
1.1 An Introduction to Survey Research	5
1.2 An Introduction to Survey Errors	7
1.3 An Introduction to Survey Errors in Web Surveys	10
1.3.1 Coverage Error.....	10
1.3.2 Sampling Error.....	12
1.3.3 Nonresponse Error	14
1.3.4 Measurement Error	17
2. Survey Responding in Web Surveys	29
2.1 Cognitive Information Processing.....	29
2.1.1 Cognitive Shortcuts in Survey Responding.....	32
2.1.2 Conditions Fostering Cognitive Shortcuts.....	33
2.1.3 Integrative Framework of Survey Responding.....	43
2.2 Perceptual Information Processing.....	47
2.2.1 Visual Perception and Attention in Survey Responding	47
2.2.2 Verbal and Visual Questionnaire Features	51
2.2.3 Processing of Visual Questionnaire Features	53
3. Design and Administration of Questionnaires in Web Surveys.....	57
3.1 Characteristics of Questionnaire Design	58
3.1.1 Question Content	58
3.1.2 Question Format	64
3.2 Characteristics of Questionnaire Administration	67
3.2.1 Questionnaire Structure	67
3.2.2 Actual and Announced Survey Length.....	70
3.2.3 Progress Indicator	72

3.2.4	Real-Time Validation.....	73
3.2.5	Question Context.....	75
3.2.6	Socio-Environmental Context.....	76
4.	Design and Administration of Rating Scales in Web Surveys	79
4.1	Form-Related Verbal Characteristics of Rating Scales	80
4.1.1	Number of Items in a Measure.....	80
4.1.2	Polarity of Items.....	80
4.1.3	Number of Response Options	81
4.1.4	Labeling of Response Options	83
4.1.5	Agree-Disagree versus Construct-Specific Response Options ..	84
4.1.6	Polarity of Response Options.....	85
4.1.7	Middle Response Option.....	86
4.1.8	'Don't Know' Response Option	87
4.2	Form-Related Visual Characteristics of Rating Scales.....	88
4.2.1	Single- versus Multiple-Item-per-Screen Formats.....	88
4.2.2	A Special Case: Grid Formats.....	90
4.3	Context-Related Characteristics of Rating Scales	99
4.3.1	Item-Order Effects	99
4.3.2	Response-Order Effects	105
5.	Assessing Data Accuracy in Rating Scales.....	111
5.1	Internal Consistency	113
5.2	Systematic Response Tendencies	116
5.2.1	Reversed-Item Bias	116
5.2.2	Nondifferentiation	119
5.2.3	Acquiescence.....	124
5.2.4	Extremity.....	126
5.2.5	Primacy Effects	127
5.2.6	Summary	131
5.3	Item Missing Data.....	133
5.4	Response Times	135
6.	Improving Data Accuracy in Rating Scales.....	143
6.1	Length of Rating Scales.....	143
6.2	Balanced Rating Scales.....	144
6.3	Interactive Rating Scales	146

7.	New Drag-and-Drop Rating Scale Designs.....	151
7.1	Drag-Response Scale.....	152
7.2	Drag-Item Scale.....	153
7.3	Expectations Regarding Data Accuracy.....	155
8.	Methods.....	169
8.1	Participants.....	170
8.1.1	Panel Survey 2012.....	170
8.1.2	University Applicants Survey 2012.....	171
8.1.3	University Applicants Survey 2013.....	172
8.2	Experimental Manipulation.....	172
8.2.1	Scale Format.....	172
8.2.2	Scale Length.....	174
8.2.3	Scale Arrangement.....	175
8.2.4	Scale Sequence.....	175
8.3	Experimental Designs.....	176
8.3.1	Experiment 1.....	176
8.3.2	Experiment 2.....	178
8.3.3	Experiment 3.....	179
8.4	Instruments.....	181
8.4.1	Experiment 1.....	181
8.4.2	Experiment 2.....	182
8.4.3	Experiment 3.....	183
8.5	Measures.....	184
8.5.1	Careless Responding.....	184
8.5.2	Nondifferentiated Responding.....	184
8.5.3	Acquiescent Responding.....	185
8.5.4	Extreme Responding.....	186
8.5.5	Primacy Effects.....	186
8.5.6	Semantic-Order Effects.....	187
8.5.7	Item Missing Data.....	188
8.5.8	Response Times.....	189
8.5.9	Respondent Evaluation.....	191
9.	Results.....	193
9.1	Item Nonresponse.....	193
9.1.1	Experiment 1.....	193
9.1.2	Experiment 2.....	198
9.1.3	Experiment 3.....	200

9.1.4	Summary	203
9.2	Survey Breakoff.....	204
9.2.1	Experiment 1	204
9.2.2	Experiment 2	207
9.2.3	Experiment 3	207
9.2.4	Summary	209
9.3	Scale Properties.....	209
9.3.1	Dimensionality	209
9.3.2	Internal Consistency Reliability and Item Means	211
9.4	Careless Responding.....	215
9.5	Nondifferentiated Responding.....	220
9.5.1	Experiment 1	220
9.5.2	Experiment 2	224
9.5.3	Experiment 3	226
9.5.4	Summary	229
9.6	Acquiescent Responding	230
9.7	Extreme Responding.....	234
9.7.1	Experiment 1	234
9.7.2	Experiment 2	237
9.7.3	Experiment 3	238
9.7.4	Summary	241
9.8	Primacy Effects.....	242
9.8.1	Experiment 2	242
9.8.2	Experiment 3	246
9.8.3	Summary	251
9.9	Semantic-Order Effects	251
9.10	Response Times	257
9.10.1	Experiment 1	257
9.10.2	Experiment 2	267
9.10.3	Experiment 3	269
9.10.4	Dragging Times.....	278
9.10.5	Summary	282
9.11	Respondent Evaluation	283
10.	Summary and Conclusions	289
10.1	Main Findings and Implications	291
10.2	Overall Assessment	303
10.3	General Discussion	308
10.4	Limitations and Further Research.....	310

References	317
Appendix A: Scale and Item Characteristics	333
Appendix B: Scale Properties	339
Appendix C: Statement of Academic Honesty	353
Appendix D: Curriculum Vitae	353

LIST OF TABLES

Table 1: Overview of the six experiments implemented in three Web surveys	170
Table 2: Description of differing rating scales formats tested.....	173
Table 3: Number of completes per experimental condition (Experiment 1.1).....	177
Table 4: Number of completes per experimental condition (Experiment 1.2).....	177
Table 5: Number of completes per experimental condition (Experiment 1.3).....	177
Table 6: Number of completes per experimental condition (Experiment 2).....	178
Table 7: Number of completes per experimental condition (Experiment 3.1).....	180
Table 8: Number of completes per experimental condition (Experiment 3.2).....	181
Table 9: Proportion of no missing, partially missing, and completely missing values (in %) depending on scale format and scale length (Experiment 1.1, $n = 771$)	195
Table 10: Proportion of no missing, partially missing, and completely missing values (in %) depending on scale format and scale length (Experiment 1.2, $n = 5,262$)	196
Table 11: Proportion of no missing, partially missing, and completely missing values (in %) depending on scale format and scale length (Experiment 1.3, $n = 5,896$)	198
Table 12: Proportion of no missing, partially missing, and completely missing values (in %) depending on scale format and scale arrangement (Experiment 2, $n = 768$)	200
Table 13: Proportion of no missing, partially missing, and completely missing values (in %) depending on scale format and scale arrangement (Experiment 3.1, $n = 5,865$)	202
Table 14: Proportion of no missing, partially missing, and completely missing values (in %) depending on scale format and scale arrangement (Experiment 3.2, $n = 5,998$)	203
Table 15: Breakoff rate (in %) depending on scale format and scale length (Experiment 1.1, $n = 812$)	205
Table 16: Breakoff rate (in %) depending on scale format and scale length (Experiment 1.2, $n = 5,486$)	206

Table 17: Breakoff rate (in %) depending on scale format and scale length (Experiment 1.3, $n = 6,250$).....	206
Table 18: Breakoff rate (in %) depending on scale format and scale arrangement (Experiment 2, $n = 833$)	207
Table 19: Breakoff rate (in %) depending on scale format and scale arrangement (Experiment 3.1, $n = 6,034$)	208
Table 20: Breakoff rate (in %) depending on scale format and scale arrangement (Experiment 3.2, $n = 6,147$)	208
Table 21: Item-pair correlations depending on scale format and scale length (Experiment 1.1, $n = 714$).....	217
Table 22: Item-pair correlations depending on scale format and scale length (Experiment 1.2, $n = 4,813$).....	218
Table 23: Item-pair correlations depending on scale format and scale length (Experiment 1.3, $n = 5,529$).....	219
Table 24: Differentiation index depending on scale format and scale length (Experiment 1.1, $n = 714$).....	221
Table 25: Differentiation index depending on scale format and scale length (Experiment 1.2, $n = 4,813$).....	223
Table 26: Differentiation index depending on scale format and scale length (Experiment 1.3, $n = 5,529$).....	224
Table 27: Differentiation index depending on scale format and scale arrangement (Experiment 2, $n = 727$)	226
Table 28: Differentiation index depending on scale format and scale arrangement (Experiment 3.1, $n = 5,211$)	228
Table 29: Differentiation index depending on scale format and scale arrangement (Experiment 3.2, $n = 5,227$)	229
Table 30: Acquiescence index depending on scale format and scale length (Experiment 1.1, $n = 714$).....	231
Table 31: Acquiescence index depending on scale format and scale length (Experiment 1.2, $n = 4,813$).....	232
Table 32: Acquiescence index depending on scale format and scale length (Experiment 1.3, $n = 5,529$).....	233
Table 33: Extremity index depending on scale format and scale length (Experiment 1.1, $n = 714$).....	235

Table 34: Extremity index depending on scale format and scale length (Experiment 1.2, $n = 4,813$).....	236
Table 35: Extremity index depending on scale format and scale length (Experiment 1.3, $n = 5,529$).....	237
Table 36: Extremity index depending on scale format and scale arrangement (Experiment 2, $n = 727$).....	238
Table 37: Extremity index depending on scale format and scale arrangement (Experiment 3.1, $n = 5,211$).....	239
Table 38: Extremity index depending on scale format and scale arrangement (Experiment 3.2, $n = 5,227$).....	241
Table 39: Server-side item-response time (mean in seconds) depending on scale format and scale length (Experiment 1.1, $n = 673$).....	259
Table 40: Server-side item-response time (mean in seconds) depending on scale format and scale length (Experiment 1.2, $n = 4,644$).....	261
Table 41: Client-side item-response time (mean in seconds) depending on scale format and scale length (Experiment 1.3, $n = 5,075$).....	264
Table 42: Adjusted item-response time (mean in seconds) depending on scale format and scale length (Experiment 1.3, $n = 5,055$).....	267
Table 43: Server-side item-response time (mean in seconds) depending on scale format and scale arrangement (Experiment 2, $n = 703$).....	269
Table 44: Client-side item-response time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.1, $n = 4,916$).....	271
Table 45: Adjusted item-response time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.1, $n = 4,889$).....	273
Table 46: Client-side item-response time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.2, $n = 4,943$).....	274
Table 47: Adjusted item-response time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.2, $n = 4,995$).....	277
Table 48: Otto and colleagues' (2001) three-dimensional scale on perceived emotional intelligence (Experiment 1.1).....	334
Table 49: Modick's (1977) three-dimensional scale on achievement's motive (Experiment 1.2/ 1.3).....	335
Table 50: Gerlitz and Schupp's (2005) Ten-Item Personality Inventory (TIPI) (Experiment 2).....	336

Table 51: Scale on reasons for social advancement published in Weinhardt and Schupp (2011) (Experiment 3.1)	337
Table 52: Scale on locus of control published in Weinhardt and Schupp (2011) (Experiment 3.2)	338
Table 53: Principal components analysis (Experiment 1.1, $n = 714$)	340
Table 54: Principal components analysis (Experiment 1.2, $n = 4,813$)	341
Table 55: Principal components analysis (Experiment 1.3, $n = 5,529$)	342
Table 56: Principal components analysis (Experiment 2, $n = 727$)	343
Table 57: Principal components analysis (Experiment 3.1, $n = 5,211$)	344
Table 58: Principal components analysis (Experiment 3.2, $n = 5,227$)	345
Table 59: Internal consistency reliability and item means depending on scale format (Experiment 1.1, $n = 714$)	346
Table 60: Internal consistency reliability and item means depending on scale format (Experiment 1.2, $n = 4,813$)	347
Table 61: Internal consistency reliability and item means depending on scale format (Experiment 1.3, $n = 5,529$)	348
Table 62: Internal consistency reliability and item means depending on scale format (Experiment 2, $n = 727$)	349
Table 63: Internal consistency reliability and item means depending on scale format (Experiment 3.1, $n = 5,211$)	350
Table 64: Internal consistency reliability and item means depending on scale format (Experiment 3.2, $n = 5,227$)	351

LIST OF FIGURES

Figure 1: Components of the total survey error (TSE) and the mean squared error (MSE) according to Biemer (2010)	8
Figure 2: Graphical layout of the drag-response rating scale	154
Figure 3: Graphical layout of the drag-item rating scale	154
Figure 4: Response distributions and item means depending on a positive-to-negative (p-n) or negative-to-positive (n-p) scale arrangement separately for scale format (Experiment 2, $n = 727$).	245
Figure 5: Response distributions and item means depending on a positive-to-negative (p-n) or negative-to-positive (n-p) scale arrangement separately for scale format (Experiment 3.1, $n = 5,211$).	248
Figure 6: Response distributions and item means depending on a positive-to-negative (p-n) or negative-to-positive (n-p) scale arrangement separately for scale format (Experiment 3.2, $n = 5,227$).	250
Figure 7: Mean shifts (left) and reliability shifts (right) depending on a forward and backward scale sequence separately for scale format (Experiment 3.1, $n = 5,211$).	254
Figure 8: Mean shifts (left) and reliability shifts (right) depending on a forward and backward scale sequence separately for scale format (Experiment 3.2, $n = 5,227$).	256
Figure 9: Initial-reaction time (mean in seconds) depending on scale format and scale length (Experiment 1.3, $n = 5,127$)	265
Figure 10: Initial-reaction time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.1, $n = 4,946$)	272
Figure 11: Initial-reaction time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.2, $n = 4,969$)	276
Figure 12: Dragging time (mean in seconds) depending on scale format and scale length (Experiment 1.3, $n = 1,736$)	279
Figure 13: Dragging time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.1, $n = 3,072$)	280
Figure 14: Dragging time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.2, $n = 3,024$)	281

Figure 15: Respondent evaluation regarding navigation-related aspects of the survey depending on scale format, separately for the six experiments	284
Figure 16: Respondent evaluation regarding design-related aspects of the survey depending on scale format, separately for the six experiments	285
Figure 17: Respondent evaluation regarding overall survey perception depending on scale format, separately for the six experiments.....	287

LIST OF APPENDICES

Appendix A: Scale and Item Characteristics	333
Appendix B: Scale Properties.....	339
Appendix C: Statement of Academic Honesty.....	353
Appendix D: Curriculum Vitae	353

1. INTRODUCTION

In recent years, Web surveys have steadily gained importance as a mode of data collection in survey research. The latest figures show that among all data collection methods in 2013, Web surveys were in a leading position with 36% of all surveys conducted by German private market and social research agencies, on par with telephone surveys. At the same time, face-to-face surveys made up no more than 22% of all survey instruments. Most noticeable, however, is the meanwhile marginal position of paper-based surveys with only 6% in 2013. This small percentage is mainly due to the fact that the majority of paper-based questionnaires have been replaced by Web surveys in recent years (ADM, 2013).

This rapid transition presumably explains why Web-based questionnaire designs still often rely on design principles that commonly stem from general guidelines for paper-based questionnaires, without making full use of the opportunities available within the scope of visual and dynamic features of questionnaire design and administration in Web surveys. Not so long ago, general questionnaire design guidelines for Web surveys advised survey researchers to “present each question in a conventional format similar to that normally used on paper self-administered questionnaires” (Dillman, 2000, p. 379). At about the same time, however, Couper and colleagues (2001) already pointed out that interactive Web surveys allow for the implementation of advanced dynamic features that can be specifically used to make survey responding more efficient and to improve data accuracy. Dynamic features can be used to draw the respondents’ attention directly to the relevant question components, to improve orientation on a Web page, and therefore, to decrease task difficulty (Couper, Tourangeau, Conrad, & Zhang, 2013; Kaczmirek, 2010). Furthermore, dynamic features can be used to increase respondent motivation by making the response task more engaging (Sikkel, Steenbergen, & Gras, 2014; Stanley & Jenkins, 2007).

In survey research, rating scales are frequently used to measure the respondents’ attitudes towards a variety of social, cultural, political, and economic issues as well as self-descriptions related to various personality traits and abilities by asking a series of related items, each of them using the same response options (Krosnick & Fabrigar, 1997; Preston & Colman,

2000). The popularity of rating scales is attributed to several factors: “Among the many practical advantages ratings are easy to construct and implement, are easy for respondents to use, are fast, and have approximately interval properties” (Coote, 2011, p. 1296). The basic decision on whether several rating scale items are presented together on the same screen, or separately with each individual item on a single screen, is commonly made in favor of presenting all items together in a grid format. In grid questions, the rating scale items are usually arranged in rows with each row being treated as a separate question, while the response options shared by all items are arranged column by column. Typically, conventional radio button scales are preferred for data input (Toepoel, Das, & Van Soest, 2009b).

Despite the widespread use and the resultant benefits of displaying multiple items neatly arranged and supposedly easy to complete on a single screen, grid questions are associated with several drawbacks repeatedly referred to in literature. A concern often expressed is that grid formats encourage the respondents to rely on cognitive shortcuts, while rushing through a set of items rather quickly with minimized navigational and cognitive effort, and without giving the necessary attention to the content of a rating scale. In turn, such survey-taking behaviors striving for effort minimization are deemed to be at the expense of processing each item carefully, resulting, among others, in less complete and less differentiated responses compared to using grid questions with fewer items or presenting each item on a single screen (Couper, et al., 2001; Toepoel, et al., 2009b; Tourangeau, Couper, & Conrad, 2004).

In self-administered surveys in general, where no interviewer is present to render assistance and motivate the respondents throughout the course of survey completion, it is up to the respondents themselves to either carefully process all survey questions and provide accurate answers, or choose not to answer at all. Hence, the likelihood of complete and accurate responses to survey questions mainly depends on how much effort a respondent is able and willing to invest in the processing of the questions. A decisive factor determining the extent of effort required to adequately answer survey questions lies in the difficulty of the response task (Dillman, Smith, & Christian, 2009, p. 69; Krosnick & Alwin, 1987). The level of task difficulty largely depends on the verbal and visual characteristics related to the design and administration of a questionnaire (Beatty & Herrmann, 2002; Couper, 2000; Ganassali, 2008; Krosnick, 1991).

In Web surveys, the use of conventional radio buttons is still prevailing as an input method because “radio buttons allow questionnaires to be designed to look very much like conventional mail surveys with which most respondents are familiar” (Heerwegh & Lossveldt, 2002, p. 471). By contrast, advanced dynamic rating scale formats are more difficult to use than conventional rating scales using radio buttons, at least for respondents who are less computer literate. For one thing, dynamic designs involve specific technical requirements such as JavaScript. For another, dynamic designs require new input methods such as slider scales, where respondents have to drag a bar with the mouse pointer to indicate their answers. Special technical requirements and higher task difficulty of dynamic rating scale designs usually increase the time and effort needed to provide an answer. As a consequence, respondent burden rises because of the higher demands on cognitive and navigational processing, which in turn is likely to result in unreasonably long response times, an increased risk of survey breakoff, and impaired data accuracy. Therefore, new rating scale designs are still used rather rarely (Couper, Tourangeau, Conrad, & Singer, 2006; Heerwegh & Lossveldt, 2002; Peytchev, 2009). So, although advanced dynamic rating scale formats are available, conventional radio buttons are still preferred as the common input method in rating scales. And, although grid questions suffer from a number of shortcomings, they are nevertheless the prevailing question format for rating scales.

The main focus of this study is on two rating scale formats using drag-and-drop as a more dynamic input method with the respondents being required to left-click on a draggable element, drag it to a desired position while holding the mouse button down, and then release the mouse button to drop the element in the desired position. The two drag-and-drop rating scales differ in the respective rating scale component which is draggable. In the drag-response scale, respondents need to drag the response options towards an item, whereas in the drag-item scale, respondents have to drag the items towards a response option. The main objective is to determine whether these newly designed drag-and-drop rating procedures can be used in Web surveys as adequate alternatives for conventional grid questions using radio buttons, and moreover, whether they can improve data accuracy by encouraging the respondents’ attentive and careful processing of rating scales. More specifically, it is asked whether the drag-response and drag-item scale can effectively prevent the respondents’ susceptibility to cognitive shortcuts in

rating scales as compared to conventional grid questions by (a) counteracting respondent fatigue, (b) drawing the respondents' attention to the items and response options as the key components of a rating scale, and (c) strengthening the link between both components to encourage a more careful matching of items and response options. Besides the respondents' susceptibility to cognitive shortcuts, the occurrence of item missing data, response times, and respondent evaluations are assessed as more indirect indicators of data accuracy.

In the remainder of this chapter, a brief introduction to survey research is provided. The basic requirements for designing effective Web surveys are discussed along the "four cornerstones of survey research", namely sampling, coverage, nonresponse, and measurement (De Leeuw, Hox, & Dillman, 2008, p. 4). The total survey error framework is introduced describing various error sources in surveys affecting the accuracy of survey data, whereas the focus in this study is on different sources of measurement error. In Chapter 2, key aspects of the respondents' cognitive and perceptual processing of survey questions are outlined to gain a better understanding of how respondents deal with survey questions. Special attention is drawn to the difficulty of the response task as a relevant determinant of a respondent's ability and motivation to provide complete and accurate survey responses. Chapter 3 presents an overview of aspects related to the design and administration of questionnaires. These aspects are further referred to in Chapter 4 with special focus on rating scales in Web surveys. Grid questions as the most common rating scale format in Web surveys are discussed with respect to their implications for the cognitive and navigational processing and data accuracy in rating scales. In Chapter 5, five different types of systematic response tendencies which are commonly used as indicators of data accuracy in rating scales are outlined. In addition, item missing data is introduced as a rather indirect indicator of data accuracy. Moreover, the meaning of response times is discussed related to data accuracy in rating scales. At this stage in the study, the theoretical foundations are laid for a comprehensive understanding of the processing of rating scales in Web surveys. In Chapter 6, previous approaches used in Web surveys to improve data accuracy in rating scales are outlined. Taking into account the theoretical and practical knowledge of rating scale design and administration, two drag-and-drop rating scale procedures are introduced with their main characteristics, potential benefits, and shortcomings being considered in Chapter 7. Key assumptions regarding the

effectiveness of the drag-response and drag-item scale in reducing the risk of systematic response tendencies and the occurrence of item missing data in rating scales are discussed and tested in the following chapters. Therefore, the study design of six experiments is described in Chapter 8. The findings are presented in Chapter 9 with an emphasis on the comparison of the two drag-and-drop rating scales with a conventional grid question. Finally, in Chapter 10, the main findings of the present study are summarized and discussed, followed by implications and directions for further research.

1.1 An Introduction to Survey Research

In survey research, questionnaires and interviews are used to systematically collect data on attitudes and behaviors, factual issues, and other socio-demographic information from a sample of individuals (De Leeuw, et al., 2008). Hence, information gathering in survey research is primarily based on self-reports. And, instead of surveying all elements of the target population the survey researcher is interested in and wants to make inference, a sample of elements—mostly a subset of individuals or households—is drawn from a sampling frame to conduct a survey and generalize the results to the population of interest (Couper, 2000). The implementation of a census in terms of surveying all units of the target population is mostly impracticable particularly because of the high costs and shortcomings in actually identifying every entity of a population. By contrast, sample surveys are characterized by lower costs, less effort for administration, and greater timeliness due to shorter field phases and turnaround times. In addition, better response rates and greater accuracy are yielded because the implementing agency is able to invest more time and effort in “maximizing responses from those surveyed, perhaps via more effort invested in survey design and pre-testing, or perhaps via more detailed non-response follow-up” (R. D. Fricker, 2008, p. 196). However, appropriate conclusions about the target population presuppose that a proper sampling frame is constructed and random sampling is applied. Taking this into account, conducting a survey comprises several steps including the definition of the key objectives of a survey and the identification of the target population which the survey researcher is interested in, the selection of a sampling frame comprising the accessible population, the determination of the method by which the sample is drawn from the sampling

frame, the decision on the mode of administration or any combination of modes, as well as the design of an adequate survey instrument, followed by the collection, processing, analysis, and reporting of the survey data (Kelley, Clark, Brown, & Sitzia, 2003).

A major concern in survey research is to optimize survey quality. Survey quality is best considered a multidimensional concept including various quality requirements for survey data. Biemer and Lyberg (2003) emphasized three dimensions of survey quality, namely, timeliness of survey data in respect that it is “available at the time it is needed”, accessibility of survey data “to those for whom the survey was conducted”, and accuracy of survey data which is commonly defined in terms of the total survey error (p. 13). According to the total survey error framework, the accuracy of survey data can be substantially interfered by different sources of error. In attitude and personality measurement, when asking respondents for their attitudes and self-descriptions, self-reports are particularly prone to measurement error as one component which may restrict data accuracy. Coverage, sampling, and nonresponse are considered further error sources of key importance to data accuracy (Biemer & Lyberg, 2003; Groves, et al., 2009). De Leeuw and colleagues (2008) referred to sampling, coverage, (non)response, and measurement as the “four cornerstones of survey research”, which in turn are based on the specification of the theoretical concepts that are intended to be measured in a survey (p. 4).

Various modes of data collection can be distinguished in survey research including face-to-face and telephone interviews as types of interviewer-administered surveys, as well as paper-based and Web-based questionnaires as kinds of self-administered surveys. Web surveys have steadily gained in importance over the last years since this data collection method offers several benefits compared to other modes of data collection. The following features are considered one of the most advantageous: By using Web surveys, a large number of potential respondents can be reached in relatively short time. Furthermore, Web surveys are more cost-effective than face-to-face, telephone, or paper-based surveys. Like in face-to-face or telephone surveys, computer-assisted survey completion enables the implementation of complex question routing and skipping, without overburdening or overstraining the respondents. Moreover, computer technology enables survey researcher to use basic or even complex visual elements and advanced dynamic features in questionnaire design and

administration making the questionnaire visually appealing and more interactive for the respondents. And, less time is needed for data preparation and data cleansing. Like in paper-based surveys, undesired interviewer effects can be avoided. Furthermore, respondents can finish the survey off when and where it suits them best (Couper & Bosnjak, 2010; Wright, 2005).

On the contrary, Web surveys present new challenges to survey researchers as related to the aspects of coverage, sampling, nonresponse, and measurement. For instance, participation in Web surveys presupposes certain computer and Internet knowledge. Apart from this, not all individuals or households in the general population have access to the Internet, certain populations are less likely to have Internet access, or certain populations are more reluctant to participate in Web surveys (Couper, 2000; Dillman & Bowker, 2001; Grandjean, Nelson, & Taylor, 2009). Furthermore, it is virtually impossible to draw a probability sample for the general population because of the lack of complete lists of e-mail addresses and an appropriate method for generating random samples of e-mail addresses (Couper, 2000; Dillman & Bowker, 2001). In the following, the framework of total survey error is explained related to the special requirements of Web surveys.

1.2 An Introduction to Survey Errors

A major concern in survey research is to optimize total survey quality which presents survey researchers with various challenges. Besides timeliness and accessibility of survey data, achieving a high level of data accuracy in terms of minimum deviations between the survey estimate and the true value of the population parameter is considered a major challenge in maximizing total survey quality (Biemer & Lyberg, 2003, p. 13). However, high accuracy of inferences derived from survey data can substantially be interfered by different sources of error potentially arising during the survey process. The total survey error (TSE) is a conceptual framework describing various error sources in surveys. Used as an indicator of data accuracy, the TSE is defined as “the difference between a population mean, total, or other population parameter and the estimate of the parameter based on the sample survey (or census)” (Biemer & Lyberg, 2003, p. 36). The TSE can be quantified by the mean squared error (MSE) described as the “average squared difference between the estimates produced by many hypothetical repetitions of the

survey process and the [true] population parameter value” (Biemer & Lyberg, 2003, p. 53). The larger the MSE, the greater the magnitude of the negative effect of one or more sources of error on the accuracy of the survey estimate. However, the MSE is rarely estimated in practice because of the considerable costs and effort involved (Biemer & Lyberg, 2003, p. 55).

The TSE framework can be applied to identify factors affecting the accuracy of survey estimates with the aim of maximizing the quality of survey data. Although current descriptions of the TSE framework make no difference in the various sources of errors included in the framework, descriptions differ in the systematization of the error sources. Groves and colleagues (2009, pp. 41, 48) systematized the different sources of error along a first dimension of measurement (‘what’ is the survey about), with measurement and processing being potential error sources, and a second dimension of representation (‘who’ is the survey about), with coverage, sampling, nonresponse, and adjustment as potential error sources. By contrast, Biemer and Lyberg (2003, p. 35) basically distinguished between sampling errors that occur as a result of drawing a sample rather than surveying the total population of interest, and nonsampling errors that occur during the collection and processing of survey data. A schematic representation of the TSE framework is depicted in Figure 1 which distinguishes the component of sampling error and the five subcomponents of nonsampling error, namely specification error, frame error, nonresponse error, measurement error, and data processing error (Biemer & Lyberg, 2003, p. 39).

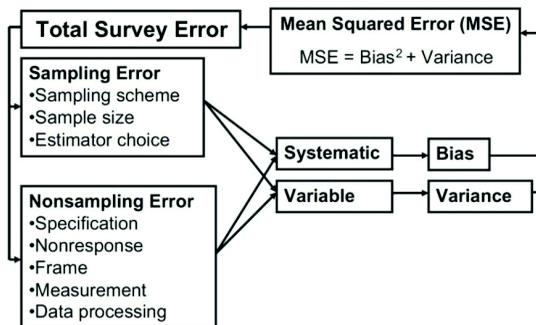


Figure 1: Components of the total survey error (TSE) and the mean squared error (MSE) according to Biemer (2010)

Sampling error refers to sampling bias in survey estimates resulting from violating the condition that each member of the target population has a known and non-zero probability of being selected, or refers to sampling variance in survey estimates resulting from the fact that different samples drawn from the sampling frame produce different survey estimates (Groves, 1989, p. 240). Specification error occurs when the concept which is actually measured by a survey question deviates from the theoretical construct intended to be measured. In such a case, a clear relation between the theoretical construct and the variables in the questionnaire is missing. Nonresponse error occurs when no substantive data is obtained from sampled individuals or households for some or all survey questions. Frame error (or coverage error) occurs when elements of the target population are missing, when they are erroneously included, or when they are listed more than once in the sampling frame which is used to draw a sample. Measurement error refers to differences between reported or recorded values and the respondents' true values with inaccurate measures potentially stemming from the interviewer (if one is present), from the respondents themselves, and/or from the questionnaire. Data processing error occurs once survey data has been collected and refers to processes of data entry, data cleaning, imputation and weighting procedures, and data reporting (Biemer & Lyberg, 2003, pp. 38-43).

These different types of error can further be classified into a systematic and random (or variable) error component, with both components having different implications for statistical analysis. Hence, each source of sampling and nonsampling errors contributes either systematically, randomly, or both to the cumulative effect of the total survey error (Biemer & Lyberg, 2003, p. 46ff). Considering the total MSE, the distinction between systematic and random errors is reflected in the decomposition in squared bias and variance of a survey estimate. In terms of random errors, an equal probability is assumed that the respondents' observed values deviate at random either positively or negatively from their true values with the effect that positive and negative deviations cancel each other out, resulting in net error effects close to zero. Although random errors have no biasing effect on survey estimates, the variance of survey estimates increases. By contrast, when positive and negative errors do not have equal probabilities and either positive or negative errors exceed, this will be reflected in systematic errors. By implication, survey estimates are systematically biased in one direction when positive errors outweigh negative errors, or vice versa. Accordingly, high data

accuracy exists when survey estimates have small bias and small variance (Biemer & Lyberg, 2003, pp. 46-51; Groves, 1989, p. 15; Weisberg, 2005, pp. 22-23). In survey measurement, accuracy and precision are often used as alternative terms: While accuracy of inferences derived from survey data refers to the inverse of the total survey error and includes both the bias and variance components, precision as the inverse of variance refers to the variance component (Biemer, 2010; Groves, 1989, p. 15f). According to this, the occurrence of systematic and random errors is closely linked to the extent of validity and reliability of survey measures, terms that are primarily applied in psychological measurement. Whereas validity refers to the extent survey measures actually measure what they are intended to measure, reliability refers to the extent survey measures provide consistent results each time they are used under the same conditions (Alwin, 1989; Viswanathan, 2005, p. 77; Weisberg, 2005, p. 22). Thus, both systematic and random errors reduce the validity of a measure, while random errors impair the reliability of a measure. In the context of experimental testing of different aspects of rating scale design and administration, the terms validity and reliability are more common, which is why these terms are used alternatively to report the findings of previous studies on variations in rating scale design and administration later in this study.

In general, coverage, sampling, nonresponse, and measurement are considered the major sources of error in surveys (Couper, 2000). In the next section, these four error sources are discussed related to the specific requirements of Web surveys, with a special focus on sources of measurement error.

1.3 An Introduction to Survey Errors in Web Surveys

1.3.1 Coverage Error

A frequently mentioned disadvantage encountered with Web surveys is that the population and sample is limited to those with access to a computer and the Internet. Thus, a major threat in Web surveys is coverage error as a result of divergences between the target population and the sampling frame used to draw a sample (Couper, 2000). In Web surveys, coverage error primarily refers to individuals or households missing from the sampling frame because they have no Internet access. Thus, the Internet penetration rate equals the

coverage rate defined as the “proportion of the target population that potentially can be reached via the Web” (Couper, 2000, p. 467).

Based on the ICT usage survey¹, information concerning the access to different electronic devices was collected at household level, while information on the use of computers and the Internet was gathered at individual level. In 2013, 83% of German households had access to a computer at home (including desktop, laptop, netbook, tablet, or handheld devices, excluding smart phones), whereas 82% of German households also had access to the Internet at home (via any device, including smart phones). Considering the individual use of a computer (at home, at work or any other place), 80% of the respondents used a computer within the last three months (including desktop, laptop, netbook, tablet, or handheld devices, excluding smart phones), of which 82% used the computer every day or almost every day. Considering the individual use of the Internet (at home, at work or any other place) via any device (including smart phones), 79% of the respondents used the Internet within the last three months, of which 80% used the Internet every day or almost every day (Statistisches Bundesamt, 2014).

Besides the proportion of the target population that can or cannot be reached via Web surveys, potential differences between these two subgroups with respect to socio-demographic variables and other substantive variables of interest are another important aspect of the risk of coverage error (Couper, 2000). The findings of the ICT usage survey 2013 showed that the proportion of households with computer and/or Internet access increased with the increase in size and the monthly net income of the household. The most common causes of not having an Internet access at home were either that an Internet access was not needed (72%) or that there was an insufficient knowledge level regarding the use of the Internet (43%). A less frequent reason for not having an Internet access at home was that an Internet access was available elsewhere, e.g., at work (13%). Considering the use of

¹ The information and communication technologies (ICT) usage survey is conducted in the EU-28 member states on an annual basis since 2002 and provides data on the availability and use of modern information and communication technologies, in particular computers and the Internet. In Germany, the survey is conducted as a postal survey comprising a household questionnaire delivered to households with at least one member aged 16 to 74 and an individual questionnaire which has to be filled in by every household member aged 10 years or older. The survey is conducted as a quota sample and delivered to approximately 12,000 households annually (Statistisches Bundesamt, 2014).

computers and the Internet within the last three months at individual level, the level of usage decreased with the increase in age, and was higher for men than for women. Furthermore, the extent of usage increased with an increase in the level of educational attainment. For school and university students, the proportion of computer and Internet usage within the last three month was the highest, at 99%, respectively. This was followed by the working population (94%, respectively), unemployed persons (81% and 79%, respectively), and retired persons (49% and 46%, respectively) (Statistisches Bundesamt, 2014). Thus, because of the high computer and Internet penetration rates among students and individuals with high levels of education, coverage is less of a problem for these two groups as it is also the case with segments of populations who have computer and Internet access at work, such as employees of certain organizations, members of professional organizations, or certain types of business people (Couper, 2000; Crawford, Couper, & Lamias, 2001; Dillman & Bowker, 2001). Previous findings on differences in substantive variables between the two subgroups that can or cannot be reached via Web surveys are less common, with the few studies revealing that differences between samples of Web respondents and samples of other survey modes do not follow a predictable pattern (Couper, 2000; Grandjean, et al., 2009).

1.3.2 Sampling Error

In order to facilitate the selection of samples that adequately represent the target population, every element of the population needs to have a known, non-zero probability of being selected. Therefore, each element of the target population needs to be part of the sampling frame from which the sample is drawn by means of probability-based sampling methods. In Web surveys, however, a major concern in the creation of a sampling frame arises because of the lack of standardized e-mail addresses and the unavailability of an appropriate method for generating random samples of e-mail addresses (Couper, 2000; Dillman & Bowker, 2001). Thus, even if presumably every individual or household had access to the Internet, probability-based sampling among units of the general population would be virtually impossible for Web surveys because not every e-mail address would be known and therefore, could not have a non-zero probability of being selected into a random sample (Couper, 2000; Dillman & Bowker, 2001).

Both coverage and sampling errors depend on the type of Web survey, i.e., which population is surveyed, and what kind of sampling method is applied. Regarding Web surveys, there is a key distinction between nonprobability surveys and probability-based surveys. The basic premise of probability sampling is that all the elements of the target population have a known and non-zero probability of being selected for a survey. Therefore, it is necessary to identify the target population, create a sampling frame, and draw a random sample from this sampling frame. This is considered essential for allowing conclusions to be drawn from a sample to the target population. For nonprobability samples, however, the relationship between the sample and the target population is unknown. As a result, there is no knowledge about how representative the sample is of the target population as a whole (Couper, 2000; Krosnick, 1999). Couper (2000) provided a comprehensive overview of the various kinds of nonprobability and probability-based surveys. The most common ones are briefly discussed further.

The most common examples of nonprobability samples in survey research are self-selected Web surveys or self-selected Web panel surveys, relying on convenience samples of Internet users. In predominantly one-off self-selected Web surveys, multiple channels (e.g., announcements on social networks or newsgroups, banner ads on high traffic websites) are used to reach potential participants. Due to the fact that access restrictions are non-existent in most of these studies, survey researchers have little or no control over who is participating in the survey and whether the same person takes part in the same survey on several occasions. In self-selected Web panel surveys, volunteers are recruited on high traffic websites or Internet portals to opt in a Web panel. Basic socio-demographic information is gathered at the time of registration. Based on a large pool of potential respondents, quota sampling or probability sampling are applied to select the respondents for later surveys. Despite the use of random sampling in this second stage, the problem remains that the initial panel is a self-selected, rather than a random sample of volunteers (Couper, 2000).

In view of the fact that not everyone in the general population has access to the Internet and no sampling frame in terms of a complete list of e-mail addresses exists for the general population from which a representative sample can be drawn, there are basically two approaches to attain probability-based Web samples. Samples are either restricted to individuals who have access to the Internet whereby the population of interest is restricted as well,

or traditional probability-based sampling methods are used to contact and recruit a panel of potential respondents. List-based samples of special populations are restricted to populations with very high or complete coverage. Intra-organizational or student sample surveys are prominent examples. Complete lists of e-mail addresses are usually available for these special population segments, enabling to draw a random sample. Pre-recruited panels of Internet users make use of probability-based sampling methods such as RDD telephone interviewing to have access to and collect background information about the general population, determine persons with Internet access, and invite those with Internet access to join the Web panel. Pre-recruited panels represent a consistent development: By RDD telephone interviewing, information is collected about the general population. All units of the sample are invited to participate in the panel, irrespective of whether they have reported to have Internet access or not. Thus, instead of limiting the recruitment of panel members to those who already have Internet access, all other persons are offered to be equipped with the necessary knowledge and resources to access the Internet in return for becoming a panel member (Couper, 2000).

Thus, probability sampling is considered indispensable to prevent sampling bias and receive samples that adequately represent the target population (Couper, 2000; Krosnick, 1999; Yeager, et al., 2011). Increasing the sample size usually reduces the sampling variance by increasing the precision of survey estimates, though larger sample sizes alone cannot necessarily overcome errors arising from undercoverage and nonresponse (Dillman & Bowker, 2001).

1.3.3 Nonresponse Error

Even if high coverage is given and a complete list of e-mail addresses is available (at least for special populations), nonresponse remains a concern in Web surveys (Couper, 2000; Dillman & Bowker, 2001). In general, nonresponse error occurs because of a failure to obtain data from individuals or households in the sample, resulting in missing data. Nonresponse is considered a specific form of response behavior and can occur at different stages of the survey process (Bosnjak & Tuten, 2001; Vehovar, Batagelj, Lozar Manfreda, & Zaletel, 2002).

Accordingly, different forms of nonresponse are distinguished which differ in the pattern of missing data either occurring at the unit level or at the item level. Unit nonresponse occurs when no data is available for a sampled individual or household. This in turn is commonly due to the fact that units in the sample refuse to take part in the survey, are physically or mentally unable to respond, or cannot be contacted during the data collection phase. As opposed to nonresponse occurring at the unit level, item nonresponse and partial nonresponse occur at the item level when substantive answers to one or more items are missing (Bethlehem, Cobben, & Schouten, 2011; Couper, 2000, pp. 5-6). Item missing data occurs after the respondents have already agreed to participate in the survey and have started to answer the questionnaire. Hence, the underlying mechanisms affecting the likelihood of item missing data differ considerably from those affecting the occurrence of unit nonresponse. The reasons for item missing data rather correspond with aspects affecting survey measurement (Groves, 1989, p. 156; Peytchev, 2009). Against this backdrop, item nonresponse and partial nonresponse are discussed related to sources of measurement error in section 1.3.4, whereas the following remarks are confined to unit nonresponse.

Generally, it is assumed that unit nonresponse is negligible provided that the group of nonrespondents is a random subset of the sample, and that unit nonresponse is not systematically related to the variables being measured in the survey. Although there are no biasing effects in that case, unit nonresponse leads to increased variance of estimates and thus, to less precise or reliable estimates as a result of the reduced effective sample size. By contrast, nonresponse biases occur when nonrespondents systematically differ from respondents with respect to the variables being measured in the survey. The extent of the biasing effects on survey data depends on the share of nonrespondents on the total sample. Even more important, however, is the actual difference between respondents and nonrespondents concerning the characteristics to be investigated in the survey. Thus, when analyses are based solely on the respondents' data, while nonrespondents differ from respondents to a non-ignorable extent with regard to the characteristics of interest, this results in a nonresponse bias in terms of inaccurate or invalid estimates of the theoretical construct (Bethlehem, et al., 2011, pp. 3-9; Groves, 1989, pp. 133-134).

As the prevention of nonresponse error is a primary objective in survey research, it is important to know the factors that discourage respondents from

complying with a survey request. Although the reasons why respondents are unable or unwilling to participate in a survey are diverse, most theoretical explanations are based on the general assumption that respondents balance the pros and cons of survey participation. This also applies to the leverage-salience theory of survey participation, according to which the respondents' decision on acceptance or refusal of a survey request depends on the importance attached to different aspects of the survey request and how strongly these aspects are emphasized in the respective situation (Groves, et al., 2009, p. 199).

Several factors influencing unit nonresponse are frequently mentioned. In general, a respondent's interest and involvement in the survey topic and the salience of this topic on the one hand are considered supportive of the decision to participate in a survey. On the other hand, being over-surveyed as a result of the sheer number of survey requests in everyday life is likely to trigger the respondent's reluctance to take part in a survey, particularly among student populations (Peytchev, 2011; Vehovar, et al., 2002). With respect to Web surveys, insufficient computer knowledge and incompatibilities in hardware and/or software prevent respondents from participating in a Web survey right from the outset (Couper, 2000; Dillman & Bowker, 2001; Lozar Manfreda, et al., 2008). In a relatively early meta-analysis including 56 Web surveys, Cook and colleagues (2000) revealed that pre-notification of the survey request and the number of contacts with the respondent are the factors most associated with higher response rates in Web surveys, though several contact attempts—including pre-notification and reminders—are only feasible up to a certain number of attempts. Based on a meta-analysis comparing Web surveys with other modes of data collection, Lozar Manfreda and colleagues (2008) similarly found that the number of reminders in a Web survey reaches an early saturation point with several contact attempts, resulting in rejecting rather than cooperating response behaviors. Both meta-analyses also showed that response rates remain unaffected by whether or not incentives are provided (Lozar Manfreda, et al., 2008; Shih & Xitao Fan, 2008).

There are several factors that can have an impact on a respondent's ability and willingness to take part in a Web survey. However, most of these factors can be influenced by survey researchers to only a limited extent. This fact is also proven by Peytchev (2011) revealing a high consistency of unit nonresponse across Web surveys: 61% of the students that were nonrespondents in a prior survey also refused to take part in a follow-up

survey they were invited to. This finding suggests that a certain percentage of sampled individuals are basically negative about survey requests because of their underlying personal predisposition.

It is generally assumed that “in the extreme, a sample will be nearly perfectly representative of a population if a probability sampling method is used and if the response rate is 100%” (Krosnick, 1999, p. 540). However, previous research has also found that the representativeness of survey data does not increase monotonically with a higher response rate, particularly because respondents with a lower propensity to participate in a survey are more susceptible to item nonresponse and systematic responding (see next section) (Kaminska, McCutcheon, & Billiet, 2010; Olson, 2006; Tourangeau, Groves, & Redline, 2010; Yan & Curtin, 2010). Hence, higher response rates are not necessarily indicative of higher data accuracy. Nonetheless, response rates are still considered an important quality indicator as low response rates raise concerns about nonresponse bias and low sample representativeness (Bethlehem, et al., 2011, p. 11; Cook, et al., 2000; Couper, 2000; Krosnick, 1999). In this respect, meta-analyses showed that Web surveys have a lower response rate than traditional survey modes, such as telephone or paper-based surveys (Lozar Manfreda, Bosnjak, Berzelak, Haas, & Vehovar, 2008; Shih & Xitao Fan, 2008). Also, larger sample sizes are needed in Web surveys compared to other modes of data collection to achieve the same precision of survey estimates (Lozar Manfreda, et al., 2008).

1.3.4 Measurement Error

Each of the aforementioned sources of error is of great importance when striving for high data accuracy. In Web surveys, however, the various sources of measurement error are of special importance (Couper, et al., 2001; Dillman & Bowker, 2001). In principal, measurement error refers to the difference between a respondent’s true value concerning the underlying theoretical construct the survey question is intended to measure and the answer actually obtained from the respondent. Such a difference between a respondent’s true value and the observed value is reflected in an inaccurate measurement of what is intended to be measured (Alwin, 2010; Viswanathan, 2005, p. 3).

It is generally assumed that a respondent’s observed value is composed of the respondent’s true value, a systematic measurement error, and a random measurement error (Alwin, 1991). Hence, measurement errors in terms of

deviations of the respondent's answer from his or her true value are divided into a random and systematic error component. While random measurement errors have no consistent effects across the entire sample but result in an increased response variance and a less reliable measurement of an underlying construct, systematic measurement errors occurring when observed values consistently deviate from the true values in either a positive or negative direction result in a response bias and an inaccurate or invalid measurement of the theoretical construct (Viswanathan, 2005, pp. 97-99). Thus, random measurement error affects the precision of survey estimates negatively in terms of reduced reliability of survey measurement attenuating observed relationships between variables of interest. Systematic measurement error interferes with the accuracy of survey estimates by negatively affecting the validity of survey measurement in two ways: (a) concerning estimates of the means of variables by inflating or deflating a respondent's observed value on a measure, and (b) concerning estimates of the relationship between variables by inflating or deflating correlations between a respondent's observed values on different measures (Alwin, 1991, 2010; Andrews, 1984; Baumgartner & Steenkamp, 2006; Viswanathan, 2005, pp. 98-99; Weisberg, 2005, p. 22). Hence, reliable measures equate to the absence of random measurement error, whereas valid and thus, accurate measures imply the absence of both random and systematic measurement error (Alwin, 2010; Groves, 1989, p. 15f; Viswanathan, 2005, p. 97).

In principal, systematic and random measurement error can stem from different sources, including the interviewer, the respondent, the questionnaire and respective questions, as well as the mode of data collection, or the setting in which the survey is conducted (Biemer & Lyberg, 2003; Groves, 1989, p. 295). Following the classification of Viswanathan (2005), systematic and random measurement errors in Web surveys—where no interviewer is present—can be basically ascribed either to certain individual characteristics of the respondent, to certain characteristics of the data collection method, or to the interaction between both respondent-related and method-related characteristics. Method-related characteristics can be further divided into specific questionnaire-related characteristics and other administration-related or contextual factors accompanied with the mode of data collection (Baumgartner & Steenkamp, 2006; Biemer & Lyberg, 2003, p. 117; Groves, 1989, p. 295; Viswanathan, 2005, pp. 135-148). In this regard, measurement error can take various forms of response errors in terms of different kinds of

systematic response tendencies. Moreover, nonresponse error occurring at the item level arises from the measurement process, which is why it often falls within the scope of sources of measurement error.

Response Errors

In reference to factors inducing systematic measurement error, different kinds of response biases can be distinguished. Generally, a “response bias is a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content” (Paulhus, 1991, p. 17). Such systematic response tendencies induce consistent but inaccurate answers to survey questions (Baumgartner & Steenkamp, 2001; Viswanathan, 2005, p. 141). Various kinds of systematic response tendencies can be distinguished, such as nondifferentiation, acquiescence, and extremity, to mention the ones most commonly examined in relation to rating scales (Baumgartner & Steenkamp, 2001; McCarty & Shrum, 2000; Van Herk, Poortinga, & Verhallen, 2004). Reversed-item biases and primacy effects are further kinds of systematic response tendencies encountered with regard to rating scales (Krosnick, 1991; Weijters, Geuens, & Schillewaert, 2009). Each of these systematic response tendencies is proven to have a negative impact on the accuracy of survey estimates. They will be discussed in detail in Chapter 5.

Although the underlying causes of these different types of systematic response tendencies can vary considerably, they have one thing in common. As previous research revealed, the occurrence of systematic response tendencies often cannot be attributed to a single source of error, which is why in most instances respondent-related and method-related factors cannot be clearly distinguished either from one another (Baumgartner & Steenkamp, 2001; McGee, 1967). Instead, it is assumed that the respondent-related factors can be encouraged or discouraged by situational determinants, such as the characteristics of the method or the context in which survey data is collected, and vice versa (Hui & Triadis, 1985; Jackson & Messick, 1958; Shulman, 1973; Snyder & Ickes, 1985; Viswanathan, 2005, p. 143). By implication, in order to enable a comprehensive understanding of systematic response tendencies and their underlying mechanisms, they are best understood as an interaction of individual respondent-related and situational method-related characteristics.

Basically, the respondent ability and the respondent motivation are considered the key respondent-related factors determining the likelihood of

systematic responding in a survey which will be examined in greater detail in section 2.1. Apart from the basic objective to gain a comprehensive theoretical understanding of systematic response tendencies, a survey researcher's primary emphasis is still on method-related sources of error because factors related to the respondent are beyond a survey researcher's control, or can only be influenced indirectly by factors related to the method.

Method-related sources of measurement error are further divided into questionnaire-related and administration-related characteristics. Questionnaire-related error sources refer to characteristics attributable to the survey instrument itself and can be subdivided into content-related and form-related characteristics of a questionnaire (Ganassali, 2008; Viswanathan, 2005, p. 143). Content-related factors refer, among others, to the clarity or ambiguity of the question wording. An insufficient understanding of the meaning of a survey question due to a complex and/or ambiguous item wording requires respondents to create their own meanings of a survey question, which in turn encourages random responding or other systematic response tendencies (Lenzner, Kaczmirek, & Lenzner, 2010; Schober & Conrad, 1997). Form-related sources of error are directly related to the format of a survey question and its visual appearance. Rating scales, for instance, are a frequently used question format in attitude and personality measurement, with the rating scale items being often arranged in a so-called grid format. However, grid formats are more likely to be associated with systematic responding compared to most other question formats because of the considerable amount of information presented simultaneously in rows and columns (Couper, et al., 2013; Kaczmirek, 2010). Administration-related sources of measurement error refer primarily to factors related to the mode of data collection and the setting in which the survey is conducted. Among others, the respondent's answers to survey questions can differ depending on whether the survey is interviewer-administered or self-administered (Bowling, 2005; De Leeuw, 2005). In the case of Web surveys, the technical equipment which is available to the respondent for processing the survey has to be taken into account (e.g., type of terminal device, screen resolution, type of Internet connection) (Bowling, 2005; Dillman, Tortora, & Bowker, 1998). Respondent distraction is another factor that further affects measurement (De Leeuw, 2005). Administration-related sources of error also comprise errors arising from the measurement procedure, e.g., the order of questions within a questionnaire, the number of questions in a questionnaire, or in the case of a

Web survey, the number of questions on a screen (Couper, et al., 2001; Krosnick & Presser, 2010; Tourangeau & Rasinski, 1988).

Nonresponse Errors at the Item Level

The occurrence of item missing data due to item nonresponse and partial nonresponse represents a further error arising from the measurement process. Item missing data poses a threat to survey inferences in case the respondents who failed to complete one or more survey questions systematically differ with regard to the survey variables of interest from those who fully completed the questionnaire (De Leeuw, Hox, & Huisman, 2003; Peytchev, 2011). As distinct from unit nonresponse but similar to different kinds of systematic response tendencies, the likelihood of item missing data occurring once respondents have begun to answer the questionnaire is decisively affected by characteristics of the questionnaire design, mainly the question content and question format, as well as by characteristics of the questionnaire administration (Beatty & Herrmann, 2002; Peytchev, 2009).

Item nonresponse occurs when data to one or more items is missing, though it was not intended or controlled by the survey researcher. According to the classification of De Leeuw, Hox, and Huisman (2003), item nonresponse refers to the absence of an answer to one or more questions either as a result of (1) a respondent's failure to provide an answer at all, (2) a respondent's failure to provide an answer usable for analysis, e.g., responses outside the range of permissible answers or 'don't know' responses being conceivably indicative of the respondent's difficulties in question processing, which in turn can be prevented by thorough pretesting, or (3) a loss of usable answers during data preparation and data cleansing which can be prevented by more careful data processing (De Leeuw, et al., 2003; Dillmann, Eltinge, Groves, & Little, 2002; Groves, 1989, p. 156; Shoemaker, Eichholz, & Skewes, 2002).

The first-mentioned form of item nonresponse occurring when a respondent provides no answer at all is considered the most problematic manifestation of item nonresponse since different missing data mechanisms take effect, with each of them having a different impact on the accuracy of survey data. This is why in the present study item nonresponse refers to a respondent's failure to provide an answer at all. The simplest explanation for the occurrence of item nonresponse is that respondents just overlook a survey question by mistake. In the case of unintentional skipping, item nonresponse

results in data missing completely at random (De Leeuw, et al., 2003; Weisberg, 2005, p. 137). Data is missing completely at random, if it is unrelated to the unobserved variable of interest, or more precisely, if it is independent of item values which would have been obtained had the data been complete (target variable) and independent of other observed variables measured in the survey (auxiliary variables). In such a case, missing values have no biasing effect on survey estimates but enlarge variance and reduce the statistical power of applied testing procedures (Bethlehem, et al., 2011, p. 123; Little & Rubin, 2002, pp. 11-13).

In most cases, however, item nonresponse is a conscious response decision that is related to the respondent's ability and/or motivation. After the initial consent to take part in a survey, respondents have to decide for each survey question "whether they can respond, and whether they will respond" (Beatty & Herrmann, 2002, p. 72). The former source of item nonresponse concerns the respondent's ability and means that a respondent is willing but unable to provide a substantive answer, resulting in data that is commonly missing at random (e.g., because of his or her declining memory, an elderly respondent is unable to remember an event, which in turn results in a missing value that is related to the respondent's age but not to the event itself). Data is missing at random when missing values are related to the auxiliary variables measured in the survey but unrelated to the unobserved values of the target variable. Although there is no direct relationship between the target variable and the probability of missing data, there will be a risk of biased estimates, if there is a strong relationship between the target variable and auxiliary variables (Bethlehem, et al., 2011, p. 123; De Leeuw, et al., 2003; Little & Rubin, 2002, pp. 11-13). By contrast, the latter source of item nonresponse refers to the respondent's motivation and the fact that a respondent decides to consciously not answer which typically leads to data that is not missing at random (De Leeuw, et al., 2003). Sensitive or difficult survey questions "demanding on respondent memory, requiring access to information (such as financial records) or involving complex question formats" have a greater risk of inducing item missing data (Dixon & Tucker, 2010, p. 611). In such a case, the probability of item missing data is systematically related to the unobserved values of the target variable, which in turn results in biased survey estimates (e.g., a respondent is unwilling to disclose sensitive information and consciously decide to skip the question) (Bethlehem, et al., 2011, p. 123; Little & Rubin, 2002, pp. 11-13). In either case, a respondent's

lack of ability and/or motivation to provide a substantive answer indicates the respondent's difficulties in processing a survey question or express the extent of the respondent burden associated with the respective survey questions (Beatty & Herrmann, 2002; Peytchev, 2009).

Partial nonresponse occurs when data is missing after a certain point in time which is often considered a more aggravated form of item nonresponse (De Leeuw, et al., 2003). Partial nonresponse refers most frequently to a respondent's premature termination of the survey, also known as survey breakoff. Survey breakoff is considered another outcome of the response decisions. Since respondents continuously re-evaluate their initial decision to participate in a survey, survey breakoff can potentially occur at every survey question (Beatty & Herrmann, 2002; Peytchev, 2009). If the respondents who prematurely abandoned a survey systematically differ with regard to the survey variables of interest from those who completed the questionnaire, there is a risk of biased survey estimates (Peytchev, 2011). Similar to item nonresponse, the reasons for a respondent's premature termination of the survey can particularly be ascribed to characteristics of the questionnaire design and administration (Peytchev, 2011; Sakshaug & Crawford, 2010). For instance, the ambiguity and sensitivity of question content can increase respondent burden, or the length of a questionnaire can increase respondent annoyance which are both likely to result in an increased risk of survey breakoff. Burdensome question formats such as grid questions make premature termination of a survey more likely as well (Galesic & Bosnjak, 2009; Peytchev, 2009; Sakshaug & Crawford, 2010).

Sources of Measurement Error in Web Surveys

Web surveys feature a number of key characteristics, including self-administration, computerization, interactivity, visual enhancement, and decentralization, whereas the "combination of [these] attributes makes Web surveys unique in terms of measurement" (Couper & Bosnjak, 2010, p. 540). The many opportunities and challenges related to the design and administration of Web-based questionnaires and the implications for measurement errors in Web surveys are discussed further.

In self-administered surveys in general, no interviewer is present who can guide and motivate the respondent throughout survey completion, or more precisely, who renders assistance and probes after inadequate answers in order to receive complete and accurate responses (Cannell, Miller, &

Oksenberg, 1981; Heerwegh, 2009; Schwarz, Strack, Hippler, & Bishop, 1991). All the information presented to the respondent usually needs to be processed visually instead of verbally which poses an additional challenge for at least the less educated respondents. Apart from this, completion of an interviewer-administered questionnaire follows a clear sequential order which is fixed by the interviewer. Also, the speed at which the survey questions are processed is decisively affected by the interviewer. By contrast, self-administered surveys allow respondents to go back and forth on their own, to take each question at their own speed, to skip questions, to turn their attention to several other things alongside, or even to interrupt the processing of the questionnaire for a while. Thus, self-administration certainly implies some advantages concerning the lack of time pressure, the absence of undesired interviewer effects, and greater perceived privacy and confidentiality. At the same time, however, there is a greater risk that respondents process the survey questions merely superficially to quickly come to the end, the potential lack of necessary explanations and clarification, and a higher risk that respondents are distracted from the actual response task (Schwarz, Strack, Hippler, et al., 1991).

These general findings concerning self-administered paper-based surveys that are well documented by Schwarz and colleagues (1991) similarly apply to Web surveys. The greater respondent privacy and confidentiality in Web surveys reduce the respondents' susceptibility to socially desirable responding to survey questions dealing with sensitive topics as compared to face-to-face or telephone interviews (Sakshaug, Yan, & Tourangeau, 2010; Tourangeau & Yan, 2007). Thus, similar to paper-based surveys, respondents are more willing to report sensitive information in self-administered Web-based questionnaires than in interviewer-administered questionnaires (Kreuter, Presser, & Tourangeau, 2008; Tourangeau & Yan, 2007). On the downside, Web surveys also suffer from impaired measurement. For instance, Heerwegh and Loosveldt (2008) found that respondents completed a questionnaire considerably faster via the Web (about 32 minutes) compared to a face-to-face setting (about 48 minutes), while at the same time, the Web respondents yielded a lower degree of differentiation on rating scales, a higher item nonresponse rate, and a 'don't know' rate that was 2.6 times higher compared to their face-to-face counterparts (Heerwegh & Loosveldt, 2008). Fricker and colleagues (2005) also found less differentiation among rating scale items when the items were arranged in a grid format and administered

via the Web rather than over the telephone, whereas no differences were found between the two modes concerning the extent of acquiescent responding. Comparing the self-administered surveys in terms of paper-based and Web-based questionnaires with an interviewer-administered telephone survey, Grandjean and colleagues (2009) found more elaborated answers to open-ended questions when these were asked over the telephone compared to in a Web survey, whereas least elaborated answers were found in the paper-based questionnaire. Based on a meta-analysis including 68 Web surveys, Lozar Manfreda and Vehovar (2002) found survey breakoff rates ranging from 0% to 73% with an average rate of 16% which is considered higher than in face-to-face or telephone interviews where breakoff rates usually do not exceed the 5% rate. However, studies systematically contrasting survey breakoff rates in Web surveys with face-to-face or telephone surveys are not known to date. In this regard, it should be emphasized that the findings reported on mode differences are far from being exhaustive. This fact is due to “gaps in our knowledge” because mode comparisons between Web surveys and face-to-face or telephone surveys are still less common than comparisons between modes that are conceptually more similar (e.g., telephone and face-to-face, Web and paper-based) (Couper, 2011, p. 895).

The differences in measurement properties and the prevalence of item missing data between the self-administered and interviewer-administered modes of data collection are generally considered to be indicative of respondents being less inclined to focus their attention on the response task and to expend the effort required for complete and accurate responses in self-administered surveys. Thus, the various kinds of systematic response tendencies, item nonresponse and survey breakoff are considered a particular concern in self-administered surveys since no interviewer is present to assist and encourage the respondents to finish a questionnaire in an attentive and careful manner (Couper, 2000; Krosnick, 1991; Peytchev, 2011; Sakshaug & Crawford, 2010; Toepoel, et al., 2009b). As compared with self-administered paper-based surveys, the problem of diminished respondent attention and reduced respondent effort is expected to be even more serious in Web surveys. Survey completion is more likely to be interfered with multitasking because multiple programs and browser windows can be run concurrently which enables the respondents to perform multiple activities at the same time. As a consequence of more respondent distraction, the probability of superficial cognitive processing of survey questions is increased (De Leeuw,

2005; Heerwegh & Loosveldt, 2008). Additionally, respondent attention to survey questions suffers from just scanning rather than carefully reading the information presented on a Web page, from generally shorter attention spans in screen-based activities, and from respondents tending to lose interest in the Web survey faster compared to other modes of data collection (Bauman, Jobity, Airey, & Atak, 2000; Gräf, 2002). Thus, self-administration in a Web environment has a decisive impact on the cognitive question-answer processing and imposes additional requirements on several aspects of the questionnaire design and administration. As Couper and colleagues (2001) already stated for self-administered surveys in general, “the design of the instrument may be extremely important in obtaining unbiased answers from respondents” (p. 231). This is considered even more pronounced when dealing with the wide range of visual and dynamic features available for the design and administration of survey questions in Web surveys (Couper, et al., 2001).

In interactive Web surveys where the survey questions are predominantly presented page-by-page, survey researchers have plenty of (audio)visual and dynamic questionnaire design and administration options available that can be specifically used to counteract at least some of the aforementioned shortcomings of self-administered surveys. Rather basic visual questionnaire features, like colors, symbols, sounds, or pictures, right up to advanced audio-visual features, such as video-interviewers or virtual-interviewers can be used to improve the respondents’ understanding of a survey question, to ease orientation and navigation on a Web page, to direct the respondents’ attention to the relevant components of a survey question, and to enhance their motivation over the entire course of survey completion (Conrad, et al., 2008; Couper, Tourangeau, & Conrad, 2007; Fuchs & Funke, 2007; Tourangeau, Conrad, & Couper, 2013, pp. 77-98). Furthermore, respondents can be prevented from going back to a previous page once they have submitted it; respondents can be prompted to answer each question before they can continue to the next Web page; progress indicators keep the respondents informed about their advance within the survey; or, additional clarifying information can be displayed when respondents actually need it (Conrad, Schober, & Coiner, 2007; Derouvray & Couper, 2002; Yan, Conrad, Tourangeau, & Couper, 2011). Concerning rating scales, for instance, dynamic highlighting can help guide the respondent’s attention to the relevant question components, thus enabling a better understanding and navigation

(Couper, et al., 2013; Kaczmirek, 2010). Moreover, dynamic prompting allows for immediate referral to the respondents' actions and, if necessary, ask them to slow down their response speed or to differentiate more among the rating scale items (Kunz & Fuchs, 2014; Zhang & Conrad, 2013). Hence, as distinct from paper-based questionnaires, interactive Web surveys allow for a visually enhanced, and particularly, more interactive respondent-survey interaction, suited to assist and encourage the respondents to provide complete and accurate responses (Couper, et al., 2001). In this way, Web surveys can also be regarded as a "hybrid mode, sharing features with self-administered and interview-administered surveys" (Christian, Parsons, & Dillman, 2009, p. 395).

Following the aforementioned findings, it will be shown further throughout this study that "although measurement error effects represent one of the most serious threats to the conduct of quality web surveys", these can be addressed by basic visual and advanced dynamic questionnaire design and administration (Dillman & Bowker, 2001, p. 170). However, it should also be noted that new sources of measurement error may be introduced in Web surveys. Since the implementation of Web surveys is decentralized, respondents use their own hardware and software to fill in the questionnaire. As a consequence, the visual presentation and physical placement of survey questions can vary depending on the type of hardware, operating system, and browser the respondents use, as well as depending on their screen configurations, screen resolution, and browser window size (Couper, 2000; Dillman & Bowker, 2001). In the case of rating scales, for instance, the number of items being in full view without the need for vertical scrolling depends on the size of the browser window. The same is true for the number of horizontally arranged response options. The number of initially visible items and response options in turn affects the cognitive processing of the rating scale as a whole and influences how attentively single components of the rating scale are processed (Couper, Tourangeau, Conrad, & Crawford, 2004; Galesic & Yan, 2010; Toepoel, et al., 2009b).

Besides differences in the visual appearance, the need for scrolling in rating scales is likely to result in greater respondent confusion and increased respondent burden, whereby respondents are at greater risk of accidentally skipping one of the items or relying on systematic response tendencies, resulting in less accurate answers (Couper, et al., 2013; Norman, Friedman, Norman, & Stevenson, 2001; Toepoel, et al., 2009b). Furthermore,

burdensome scrolling increases the risk of premature termination of the survey (Schonlau, Fricker, & Elliott, 2002). Besides poor questionnaire design, incompatibilities in hardware or software as well as insufficient computer knowledge are a further source of respondent frustration and thus, an increased risk of survey breakoff (Dillman & Bowker, 2001). Hence, although the various visual and dynamic features available in Web surveys offer numerous advantages related to the design and administration of Web surveys, they also involve a certain risk of impaired measurement and need to be carefully tested before being implemented in Web surveys (Vicente & Reis, 2010, p. 251).

In order to prevent measurement error in terms of various systematic response tendencies and item missing data, and thus to increase data accuracy in Web surveys, the main focus in survey research is on the sources of measurement error that are related to the respondents themselves, or to the method used to gather data from the respondents (Biemer & Lyberg, 2003, pp. 63, 116; Couper, 2000; Dillman, et al., 1998; Groves, 1989, pp. 11-12; Lozar Manfreda, Batagelj, & Vehovar, 2002). Thus, the likelihood of receiving complete and accurate responses in self-administered Web surveys mainly depends on how much effort a respondent is able and willing to invest in the processing of survey questions, which in turn is decisively influenced by features of the questionnaire design and administration (Beatty & Herrmann, 2002; Krosnick, 1991). To better understand the impact of these visual and dynamic features on data accuracy and to systematically examine the benefits and shortcomings associated with the implementation of such features in Web surveys, i.e., how they can promote the respondents' ability and motivation, or if they hamper or even prevent respondents from providing complete and accurate responses, we need to understand the basic mechanisms underlying survey responding and the cognitive and perceptual processing of survey questions discussed in the next sections.

2. SURVEY RESPONDING IN WEB SURVEYS

In order to prevent measurement error in terms of various systematic response tendencies and item missing data effectively, a comprehensive understanding of their underlying determinants is necessary. Besides various kinds of systematic response tendencies reflecting systematic measurement error, item missing data occurring after the respondents have agreed to participate in the survey is considered a further error arising from the measurement process. Since the underlying determinants of systematic response tendencies and item missing data are largely attributable to the same sources, these different kinds of response and nonresponse errors, both arising from the measurement process are discussed along the same dimensions of respondent-related and method-related sources of measurement error (Groves, 1989, p. 156; Peytchev, 2009).

But first of all, in order to gain a clear understanding of the determinants of complete and accurate survey responses, the question-answer process a respondent generally undergoes when answering survey questions needs to be better understood (section 2.1). In this regard, potential cognitive shortcuts in survey responding (section 2.1.1), and conditions fostering these cognitive shortcuts (section 2.1.2) are discussed. Against this background of a general understanding of response behaviors in surveys, an integrative framework of survey responding is described that incorporates explanations for the respondents' decisions to provide complete and accurate responses to survey questions (section 2.1.3). While cognitive information processing is of key importance irrespective of the mode of data collection, the issues of perceptual information processing (section 2.2) are particularly important in designing self-administered (Web) questionnaires.

2.1 Cognitive Information Processing

In general, respondents undergo various cognitive processes when answering survey questions. Established models of the question-answer process differ only slightly in the formulation of the following four stages: (1) comprehending and interpreting the question, (2) retrieving relevant

information, (3) recalling or computing a judgment, and (4) formatting and editing the response, with the latter being reported as two separate stages in some models (Cannell, et al., 1981; Groves, 1989, p. 407ff; Sudman, Bradburn, & Schwarz, 1996, p. 56ff; Tourangeau, Rips, & Rasinski, 2000, p. 165ff). These four steps of the question-answer process are best understood as parallel, rather than strictly consecutive stages. Hence, cognitive processing at these stages mostly takes place simultaneously, with feedback loops being very likely to occur (Sudman, et al., 1996, p. 56; Tourangeau, et al., 2000, p. 15). The following discussion refers to the cognitive processing of attitudinal information, while further specifics of the cognitive processing of behavioral information can be found in Sudman and colleagues (1996) and Tourangeau and colleagues (2000).

Stage I: Comprehension

Understanding and interpreting survey questions in a consistent manner and in accordance with the meaning intended by the survey researcher are essential premises for high data accuracy. Besides the syntax being too complex for some of the respondents, the use of unknown, unfamiliar, or ambiguous terms in question wording is the most prevalent reason for misunderstanding a survey question (Alwin, 1991; Bradburn, 2004; Podsakoff, MacKenzie, Jeong-Yeon, & Podsakoff, 2003). Besides the various studies examining the causes and prevention of misinterpretation of survey questions (Conrad & Schober, 2000; Fowler, 1992; Graesser, Cai, Louwerse, & Daniel, 2006; Suessbrick, Schober, & Conrad, 2000; Tourangeau, et al., 2006), Belson (1981) very clearly illustrated that even commonly used key terms or concepts like 'you', 'regularly', or 'on a weekday' can provoke quite divergent interpretations, up to being completely ignored by the respondent. Thus, a clear and unambiguous question wording is essential for a correct question comprehension and interpretation. However, what is even more important for the correct understanding and interpretation of survey questions in self-administered questionnaires and thus, in Web surveys as well, is to attract the respondents' attention to a survey question and promote their careful processing of all the relevant components (Jenkins & Dillman, 1997; Tourangeau, et al., 2000, p. 9).

Stage II: Information Retrieval

Based on their interpretation of the question content, respondents have to retrieve relevant information from memory to answer a survey question. When answering attitude questions, the extent of effort required for information retrieval mainly depends on how accessible the relevant information is (Schwarz, Strack, & Mai, 1991; Tourangeau, Rasinski, & D'Andrade, 1991; Tourangeau, et al., 2000, pp. 172-180). Saliency of information determines the accessibility and ease of retrieval, with information being highly salient if it is of high personal relevance for a respondent (e.g., being part of one's self-schema) and when a respondent often thinks about it (Groves, 1989, p. 432; Tourangeau, et al., 1991).

Stage III: Judgment

Respondents formulate their answers on the basis of the retrieved information. Therefore, they either recall an already existing judgment, or in most cases when answering attitude questions, they have to compute a judgment "on the spot" on the basis of the information currently accessible (Sudman, et al., 1996, p. 70). Concerning the computation of a judgment, it can simply be assumed that "the output from the judgment component is a simple average of the considerations that are the input to it" (Tourangeau, et al., 2000, p. 180). These considerations comprise information from long-term memory as well as information available in the respective situation. Again, accessibility of information decisively determines which information is incorporated (Sudman, et al., 1996, pp. 70-73; Tourangeau, et al., 2000, pp. 180-184).

Stage IV: Reporting

Within this stage, two different cognitive processes can be distinguished: formatting (or mapping) and editing of responses. Depending on the mode of data collection, these two aspects are of differing importance. Whereas in self-administered surveys formatting has priority, editing is particularly relevant in interviewer-administered surveys (Sudman, et al., 1996, p. 74). When respondents have to communicate their answers directly to an interviewer, they are likely to edit their answers first to conform to socially desirable standards or other situational circumstances (Biemer & Lyberg, 2003, p. 144; Groves, 1989, p. 437; Tourangeau & Yan, 2007). Despite the fact that there is no interviewer-respondent interaction or communication, editing is of certain importance in self-administered surveys as well

(Tourangeau & Yan, 2007). Concerning the process of response formatting, a key difference in the extent of effort required to answer a question is whether respondents need to provide their answers in an open-ended or closed question format. Depending on the respective question format, the cognitive and navigational requirements differ concerning the translation of a judgment into a response (Reja, Manfreda, Hlebec, & Vehovar, 2003; Tourangeau, et al., 2000, p. 231).

2.1.1 Cognitive Shortcuts in Survey Responding

In order to ensure high data accuracy, survey researchers want respondents to think carefully about survey questions and potential answers before providing a specific survey response. An optimal answer to a survey question, in terms of complete and accurate responses, requires respondents to consciously go through all four steps of the question-answer process which, however, is associated with a certain extent of cognitive effort. In order to reduce the cognitive effort required to answer survey questions, respondents quite often take various cognitive shortcuts, resulting in merely satisfying rather than optimal survey responses (Cannell, et al., 1981; Krosnick, Narayan, & Smith, 1996; Sudman, et al., 1996, p. 71). Such deviations from optimal answers due to cognitive shortcuts and their negative effects on data accuracy are described by Cannell, Miller, and Oksenberg's (1981) process theory and Krosnick and Alwin's (1987) theory of survey satisficing. Although both theories largely agree on their core statements, they differ in terminology and specifications. Here, the key concepts of Krosnick and Alwin's (1987) satisficing theory are adopted.

In general, Krosnick (1991, 1999) and colleagues' (Krosnick & Alwin, 1987) theory of survey satisficing refers to the respondents' shortcuts within the question-answer process, potentially resulting in suboptimal answers and impaired data accuracy. The satisficing theory provides a useful theoretical framework to gain a better understanding of these cognitive shortcuts and the conditions fostering them. However, satisficing cannot be observed directly. Instead, several systematic response tendencies have been identified that are deemed to aim at satisfying, rather than optimal survey responses (Krosnick, 1991). While these satisficing behaviors can take various forms, all of which reflecting measurement error and reduced data accuracy, they differ in the degree of the adverse effects on data accuracy. Accordingly, there is a

distinction between weak and strong satisficing. Weak satisficing differs from optimizing in the fact that respondents are less thorough in executing the four steps of the question-answer process which increases the risk of inaccurate responses to survey questions. In this regard, respondents “may be less thoughtful about a question’s meaning, they may search their memories less thoroughly, they may integrate retrieved information more carelessly, and/or they may select a response choice more haphazardly” (Krosnick, et al., 1996, p. 31). Although the respondents go through all four steps of the question-answer process, they minimize their cognitive effort by being satisfied with the first acceptable answer rather than searching for an optimal answer. Among others, acquiescence (see section 5.2.3) and primacy effects (see section 5.2.5) fall within the category of weak satisficing behaviors. In the case of strong satisficing, respondents reduce their cognitive effort even further by skipping the second step of information retrieval and/or the third step of information integration and judgment. Hence, the respondents dispense with any memory search or information integration, which results in a superficial selection of an answer that appears to be reasonable based on external cues rather than any internal cues. Among others, the selection of a ‘don’t know’ option or nondifferentiation among rating scale items (see section 5.2.2) is considered a form of strong satisficing behavior (Krosnick, 1991, 1999; Krosnick, et al., 1996; Krosnick & Presser, 2010).

If no explicit ‘don’t know’ option is provided, item nonresponse is, by definition, not considered a satisficing behavior (Holbrook, Green, & Krosnick, 2003). Nevertheless, item nonresponse as well as survey breakoff—as a more aggravated form of item nonresponse—are taken into account throughout the following remarks on the conditions fostering cognitive shortcuts because, first, the conditions encouraging systematic response tendencies and item missing data are much the same, and second, a comprehensive outline of the conditions determining accurate as well as complete survey responses as a prerequisite for accurate responses is pursued.

2.1.2 Conditions Fostering Cognitive Shortcuts

According to Krosnick (1991, 1999) and colleagues’ (Krosnick & Alwin, 1987) theory of survey satisficing, three different conditions can be distinguished that foster cognitive shortcuts in survey responding: respondent ability, respondent motivation, and task difficulty. Hence, respondents differ

in their ability and motivation to expend the effort required to process all the steps of the question-answer process with sufficient care. Whereas respondents with high cognitive ability and high motivation are more likely to optimize their response behavior, less able and less motivated respondents have a higher risk to satisfice, in terms of falling back on various kinds of systematic response tendencies when answering survey questions. Apart from this, the respondents' susceptibility to satisfice rather than optimize is thought to be influenced by task difficulty, whereas a higher level of task difficulty increases the risk of satisficing. Accordingly, the likelihood of satisficing can be described by the following formula (Krosnick, 1991):

$$p(\text{Satisficing}) = \frac{\alpha_1 (\text{Task Difficulty})}{\alpha_2 (\text{Ability}) \times \alpha_3 (\text{Motivation})}$$

While the likelihood of satisficing increases with growing task difficulty, it decreases with higher respondent ability and/or respondent motivation. Furthermore, the interactions between task difficulty, respondent ability, and respondent motivation are specified in this formula, assuming that the effects of all three determinants on the likelihood of satisficing are rather multiplicative than simply additive. Overall, the impact of a respondent's ability and motivation is deemed comparably low in case of low task difficulty, whereas high task difficulty increases the influence of low ability and motivation on the likelihood of satisficing behaviors. Similarly, the effects of low respondent ability on the likelihood of satisficing are reduced given that respondent motivation is high, and vice versa (Krosnick, 1991).

Respondent ability is greater among respondents who have practice in performing complex cognitive tasks, who have practice in thinking about varying topics, and who also have previously thought about the issues in question, or among those who have predefined judgments at hand. Some of the most important factors determining the respondent's motivation lie outside the survey researcher's sphere of influence. Those factors are related to the respondent, such as dispositional differences in the need for cognition, the perceived value of the survey, and the extent of personal importance of the survey topic (Krosnick, 1991; Krosnick, et al., 1996). By contrast, the extent of the actual and perceived respondent burden as another important factor influencing respondent motivation can be decisively affected by several

characteristics of the questionnaire design and administration, such as the format of a survey question, the length of the questionnaire, or the position of the respective question within the questionnaire (Galesic, 2006; Krosnick, 1991). Task difficulty depends in large part on the difficulty of interpreting the meaning of a question and retrieving relevant information from memory which is influenced by the complexity and familiarity of the language used and the concepts or constructs addressed by the survey question. The task difficulty is also a function of the ease of making a judgment and, in the case of closed questions, mapping this judgment onto the response options available (Krosnick & Presser, 2010; Podsakoff, et al., 2003; Schwarz 1994). Furthermore, situational factors like distraction can have an influence on task difficulty (Krosnick, 1991).

Respondent Ability

According to the satisficing theory, more capable and highly sophisticated respondents are more likely to provide an optimal answer because optimizing demands less cognitive effort from them and implies more enjoyment compared to less able and less sophisticated respondents, for whom an optimal answer implies accelerated effort (Krosnick & Alwin, 1987). Thus, it is commonly assumed that “respondents who have lower cognitive ability and cognitive sophistication may have more difficulty with the respondent role and would likely break off at higher rates, may yield higher item nonresponse, and may induce more error in responses” (Peytchev, 2009, p. 81). A respondent’s cognitive ability and cognitive sophistication are commonly measured by proxy variables on the basis of the respondent’s age and educational attainment as two socio-demographic characteristics that are not under the survey researcher’s control (Galesic, 2006; Krosnick & Alwin, 1988; Krosnick, et al., 1996; McCarty & Shrum, 2000; Peytchev, 2009). In fact, previous findings related to the occurrence of item nonresponse consistently confirm correlations with age and education, according to which elderly and less educated respondents tend to have higher item nonresponse rates (Messer, Edwards, & Dillman, 2012; Yan & Curtin, 2010). Furthermore, a respondent’s age actually has a significant effect on survey breakoff, however, contrary to expectations, in terms of older respondents being more likely to complete a questionnaire (Galesic, 2006; Peytchev, 2009). Similarly, respondents with higher education tend to be less susceptible to prematurely break off than respondents with lower education levels (Galesic, 2006;

Peytchev, 2009). Based on the respondents' age and their level of educational attainment, most studies showed that respondents with lower cognitive abilities and lower cognitive sophistication are more likely to engage in satisficing behaviors. This is expressed in various forms of systematic response tendencies, including primacy effects, acquiescent responding, extreme and midpoint responding, 'don't know' responses, and nondifferentiation (Kaminska, et al., 2010; Knäuper, 1999; Knäuper, Belli, Hill, & Herzog, 1997; Krosnick & Alwin, 1987; Krosnick, et al., 1996; McCarty & Shrum, 2000; Narayan & Krosnick, 1996; Toepoel, Das, & Van Soest, 2009a).

In Web surveys, however, age and education may not serve as ideally suitable proxies of a respondent's cognitive ability and cognitive sophistication since Web survey respondents have at least the cognitive ability to use computer technologies (Peytchev, 2009). In general, Web survey respondents report higher computer and Internet use and less need for assistance or difficulties in using computers and the Internet than respondents in a paper-based survey (Smyth, Dillman, Christian, & O'Neill, 2010). Accordingly, the level of computer and Internet usage is regarded as one of the most important predictors of participating in a Web survey (Vehovar, et al., 2002, p. 234). Conversely, respondents who are rather inexperienced in dealing with computers have difficulties providing and erasing answers to differing question formats in Web surveys, e.g., radio buttons, check boxes, and drop-down menus which can lead to a respondent's frustration and premature termination of a Web survey, as well as less complete and accurate responses (Dillman & Bowker, 2001; Dillman, et al., 1998). Thus, besides a respondent's age and education, a respondent's computer literacy and the extent of computer and Internet usage experience are considered relevant factors determining a respondent's cognitive ability and cognitive sophistication to complete a Web survey.

Furthermore, a respondent's ability to provide complete and accurate responses to survey questions may interact with the task difficulty which mainly depends on characteristics of the questionnaire design and questionnaire administration. Hence, a respondent's ability to answer a survey question is not exclusively attributable to a respondent's socio-demographic characteristics but also to specific characteristics of the questionnaire. In particular, respondent ability is influenced by characteristics of the question content and its accessibility and clarity, as well as by the appropriateness of

the question format and its response options (see section 3.1) (Beatty & Herrmann, 2002; Kalton & Schuman, 1982; Podsakoff, et al., 2003; Schwarz, Strack, & Mai, 1991). A respondent's cognitive ability and cognitive sophistication are also assumed to be related to the respondent's motivation to engage in the response task. For respondents with low ability, the response task may require greater cognitive effort, which is why the respondents' motivation needs to be comparatively higher in order to cope with these enhanced cognitive requirements (Krosnick & Alwin, 1988; McCarty & Shrum, 2000). Thus, motivation as a further respondent-related characteristic decisively co-determines the likelihood of satisficing behaviors which is discussed next.

Respondent Motivation

The respondents' motivation to take part in a survey, to conscientiously process all survey questions, and to completely and accurately answer them is mainly determined by two factors: respondent interest in the survey topic and respondent burden perceived while processing the questionnaire (Bauman, et al., 2000; Beatty & Herrmann, 2002; Chang & Krosnick, 2009; Galesic, 2006; Herzog & Bachman, 1981; Jenkins & Dillman, 1997; Krosnick, 1991; Lozar Manfreda & Vehovar, 2002; Peytchev, 2009). In this regard, Peytchev (2009), Galesic (2006), and Krosnick (1991, 1999) emphasized the dynamic influence of respondent interest and respondent burden on the likelihood of complete and accurate survey responses because both factors are deemed to change throughout the completion of a survey depending on the dispositional characteristics of a respondent as well as depending on method-related characteristics of questionnaire design (see section 3.1) and questionnaire administration (see section 3.2).

Previous findings commonly revealed that respondent motivation decreases over the course of survey completion (Galesic, 2006; Herzog & Bachman, 1981; Hui & Triadis, 1985; Krosnick, 1991; Peytchev, 2009). It is assumed that as a result of an increase in the burden perceived over the course of survey completion, there is a cumulative effect of the negative aspects of survey participation which accumulates, and finally, leads to survey breakoff. Consequently, response decisions may be re-evaluated for each survey question and a respondent's initial interest in the survey and preference for participation may be gradually superseded by giving preference to stop participation. While respondents tolerate a certain extent of burden, they will

break off the survey when this individual threshold is reached (Galesic, 2006; Heerwegh & Loosveldt, 2006; Herzog & Bachman, 1981; Peytchev, 2009).

Even if the critical threshold of survey breakoff has not yet been reached and a respondent still continues with the questionnaire, the increase in respondent burden is likely to have a negative effect on the respondent's motivation to expend the extent of effort required to undergo the question-answer process consciously. This in turn may result in a decreased accuracy in the respondents' answers to survey questions, particularly to questions immediately preceding the point of breakoff, with higher item nonresponse rates being one possible indicator of decreased data accuracy (Galesic, 2006; Herzog & Bachman, 1981). Furthermore, the risk of the respondents relying on cognitive shortcuts may increase during questionnaire completion because the respondents' motivation gradually disappears as "they are likely to become increasingly fatigued, disinterested, impatient, and distracted" (Krosnick, 1991, p. 214). Because of this decline in motivation, respondents are more prone to answer survey questions with little concern for accuracy, potentially resulting in different kinds of satisficing behaviors or other systematic response tendencies which are particularly likely to occur towards later sections of a questionnaire (Beatty & Herrmann, 2002; Galesic, 2006; Galesic & Bosnjak, 2009; Herzog & Bachman, 1981; Hui & Triadis, 1985; Krosnick, 1991). In rating scales, for instance, the degree of differentiation among items is likely to decrease the later a rating scale is presented in the questionnaire (Galesic & Bosnjak, 2009). Consequently, the completeness and accuracy of the respondents' answers vary depending on the burden perceived while answering the current question as well as on the cumulative burden arising during survey completion.

In sum, intrinsic respondent motivation for providing high accuracy responses either does not exist right from the beginning because of the respondents' low interest in the survey topic, or the respondents' initial motivation gradually disappears as a result of the respondent burden growing over the course of survey completion. Based on these assumptions, Galesic (2006) measured the respondents' interest and perceived burden throughout the survey by asking them after any of the 20 blocks of thematically related survey questions to evaluate how interesting the questions were and how much burden the respondents experienced while answering them. By randomizing the order of the question blocks, a respondent's interest and perceived burden were expected to be independent of the topic and position of

the survey questions. The author found a significantly negative effect of a respondent's overall interest in the survey topic on breakoff rates: Respondents who expressed an above-median interest had a 40% lower risk of survey breakoff than respondents with a below-median interest in the survey topic. Concerning a respondent's overall perceived burden, a significantly positive effect on breakoff rates was found: Respondents with a perceived burden above the median had a 20% higher risk of breakoff than respondents with a below-median perceived burden. In addition, the respondents who dropped out at any later point expressed a significantly lower initial interest and a significantly higher perceived burden right from the beginning of the survey, whereas their interest further decreased and their burden continued to rise during survey completion which was indicative of their growing preference for breakoff before termination actually occurred.

Galesic (2006) further showed that a respondent's increasing preference for breakoff was also reflected in higher overall item nonresponse rates among respondents who dropped out at any later point in time compared to respondents who completed the questionnaire; and, in higher item nonresponse rates for question blocks immediately preceding survey breakoff compared to question blocks not immediately preceding survey breakoff (Galesic, 2006). By contrast, other studies found no effect of a respondent's interest in the survey topic on survey breakoff or item nonresponse rates (Crawford, et al., 2001; Tourangeau, Groves, Kennedy, & Yan, 2009). Also, findings concerning the relation between the respondent's interest in the survey topic and the measurement properties were mixed as some studies provided initial evidence of more accurate responses in terms of, for example, higher differentiation in rating scales in case of greater interest in and higher personal relevance of the question topic (Chang & Krosnick, 2009; Herzog & Bachman, 1981), whereas other studies found little or no effect of a respondent's interest in the topic of a Web survey and his or her carefulness in survey responding in terms of more complete and accurate answers to survey questions (Tourangeau, et al., 2009). Thus, based on these findings, it can be concluded that a respondent's interest in survey topic may, but do not necessarily need to have a positive effect on the completeness and accuracy of survey data. Presumably, it is rather the case that the respondent's interest in the survey topic primarily influences a respondent's initial decision on survey participation, whereas major effects on succeeding response decisions at question level are deemed less likely.

Contrary to the respondent's interest in the survey topic as a disposition that is not under a survey researcher's control, respondent burden is mainly attributable to specific characteristics of the questionnaire design and questionnaire administration respectively, which in turn can be directly influenced by the survey researcher. Based on their classification of features inherent in questionnaire design and questionnaire administration affecting respondent burden, Wenemark and colleagues (2010) distinguished five categories: (a) cognitive burden, (b) unnecessary work, (c) distrust, (d) offending questions, and (e) distress. Cognitive burden arises from difficulties within one or more steps of the question-answer process. These difficulties in turn result from, among others, ambiguous question wordings, unrealistic demands with respect to information retrieval and integration, or inappropriate response formats (Krosnick & Presser, 2010; Schwarz 1994). A high level of cognitive challenge implied by the response task increases the cognitive burden, which in turn can have negative effects on the respondent's motivation to spend the effort required to provide complete and accurate responses (Beatty & Herrmann, 2002; Cannell, et al., 1981; Krosnick, 1991; Peytchev, 2009). On the contrary, questionnaires that are too long and offer little variety as a result of monotonous and uniform questions asking repeatedly for the same content, pose little or no cognitive challenge to the respondents, which in turn can also reduce their motivation to expend sufficient effort on consciously and thoroughly processing the questionnaire and providing complete and highly accurate answers (Galesic, 2006; Gräf, 2002; Hui & Triadis, 1985; Krosnick, 1991). Thus, a low cognitive challenge may provoke respondent fatigue in terms of a respondent's tendency to lose interest in the response task with progressive completion of a survey, which in turn is directly related to Wenemark and colleagues' (2010) second category of respondent burden: unnecessary work. Respondent burden is likely to be increased by asking repetitive questions, questions that are not applicable to a respondent, or questions perceived as irrelevant, all of which contribute to the unnecessary lengthening of a questionnaire. The remaining categories of Wenemark and colleagues' (2010) classification of aspects affecting respondent burden refer to: respondent's distrust related to potential misuse of the data by the survey researcher and a feeling of manipulation and control in consequence of vague and repetitive questions; offending questions asking for disclosure of sensitive information or information that is too personal; distress

caused by survey questions that elicit sad or frightening thoughts about a respondent's present or future life.

Closely related to respondent interest in the survey topic and respondent burden experienced during survey completion is the concept of respondent satisfaction with the survey. Respondent satisfaction is also indicative of a respondent's motivation to continue with a questionnaire and to expend the effort required to answer each survey question properly (Ganassali, 2008; Lozar Manfreda, et al., 2002). Respondent satisfaction in turn is directly influenced by formal characteristics of the questionnaire as Ganassali (2008) showed that, asking respondents to justify their overall satisfaction with the survey, aspects related to the response format (17%) and question wording (16%) were most frequently mentioned as factors influencing satisfaction. In total, formal characteristics of the questionnaire seemed to be particularly influential on a respondent's satisfaction because 47% of the respondents mentioned at least one formal characteristic (Ganassali, 2008). Thus, questionnaire design plays a major role in promoting respondent satisfaction, which in turn can help maintain a respondent's motivation to provide complete and accurate responses, resulting in a high level of data accuracy.

Since respondent burden is a negative aspect of survey participation and may lower respondent motivation to provide highly accurate answers or to provide an answer at all, survey researchers usually agree on the need to keep the extent of respondent burden at a low level. Disagreement still exists, however, concerning the impact of high or low cognitive challenge evoking either respondent frustration or respondent fatigue. Implications of high and low cognitive load as related to rating scales are further discussed in section 4.2. Various aspects of questionnaire design and questionnaire administration affecting respondent burden need to be well understood in order to prevent negative effects on respondent motivation. The extent of cognitive burden as a key aspect of overall respondent burden may especially affect respondent motivation, which in turn decisively depends on the complexity of the response task. Task difficulty as a third factor fostering cognitive shortcuts in survey responding is discussed next.

Task Difficulty

In general, the task difficulty is largely determined by characteristics of both questionnaire design and questionnaire administration. When trying to assess

the determinants of task difficulty, it is advised to consider various respondent difficulties, potentially arising within each stage of the question-answer process (Krosnick, 1991; Krosnick & Presser, 2010). Concerning the content-related characteristics of questionnaire design, a lack of clarity of question content can make it harder for a respondent to get a clear understanding of the survey question (Alwin, 1991; Bradburn, 2004; Podsakoff, et al., 2003). Similarly, retrieval-intensive questions, asking for information of low accessibility can make information retrieval more demanding. Also, judgment-intensive attitude questions are likely to require more cognitive effort to come to a thorough answer than, for instance, simple socio-demographic questions on sex, age, or occupation (Beatty & Herrmann, 2002; Krosnick, 1991; Peytchev, 2009; Schwarz, Strack, & Mai, 1991; Shoemaker, et al., 2002; Tourangeau, et al., 1991). Concerning the form-related aspects of questionnaire design, specific question formats such as open-ended questions or grid questions usually require more cognitive and navigational effort in providing an answer than other question formats (Couper, et al., 2013; Krosnick & Presser, 2010; Peytchev, 2009; Reja, et al., 2003). Some characteristics of questionnaire administration also determine the level of task difficulty. For example, the perceived task difficulty can vary depending on the overall length of the questionnaire and the respective question position (Galesic & Bosnjak, 2009; Heerwegh & Loosveldt, 2006; Herzog & Bachman, 1981; Yan, et al., 2011). Also, the presence of distracting events in the survey situation interferes with the level of task difficulty, with interruptions and external interference increasing the risk of cognitive shortcuts in terms of various systematic response tendencies (Krosnick, 1991).

Thus, the level of task difficulty is a function of different features inherent in the questionnaire design and administration which can increase the extent of cognitive and navigational effort required to provide complete and accurate responses. This increase in respondent effort interacts with the level of respondent ability and respondent motivation, with the perceived difficulty of a response task varying from respondent to respondent, and over the course of survey completion (Beatty & Herrmann, 2002; Peytchev, 2009; Schuman & Presser, 1996, p. 205). It is generally assumed that the increasing task difficulty interferes with a respondent's ability and motivation to provide complete and accurate responses (Beatty & Herrmann, 2002; Heerwegh & Loosveldt, 2008; Krosnick, 1991; Peytchev, 2009; Shoemaker, et al., 2002).

The various characteristics of questionnaire design and questionnaire administration affecting task difficulty and by association, respondent ability and respondent motivation are discussed in greater detail in section 3.1 and section 3.2. Factors of the rating scale design and rating scale administration that affect task difficulty are systematically discussed in section 4.2 and section 4.3.

2.1.3 Integrative Framework of Survey Responding

In the question-answer process and its four steps of cognitive information processing which respondents need to undergo when striving for complete and accurate survey responses, different kinds of cognitive shortcuts can occur comprising various forms of systematic response tendencies. In this regard, respondents are likely to abbreviate the question-answer process as they are not able or willing to expend the effort required to answer a survey question accurately. Otherwise, the respondents are not able or motivated to provide an answer at all and instantly decide to refuse to answer or drop out before even engaging in the question-answer process, resulting in item nonresponse or, in its extreme form, in survey breakoff. All these potential outcomes of the response decisions result in measurement error and impaired data accuracy. The integrative framework of survey responding which is stated further aims at explaining the determinants of systematic response tendencies, item nonresponse, and survey breakoff. These determinants are largely attributable to the same dimensions of respondent-related and method-related sources of measurement error: respondent ability, respondent motivation, and task difficulty (Beatty & Herrmann, 2002; Groves, 1989, p. 156; Krosnick, 1991; Peytchev, 2009). When respondents are invited to take part in a survey, they are required to repeatedly take multiple response decisions at different stages of the question-answer process. Provided that a respondent's initial decision is in favor of participating in the survey, positively influenced by, for example, the respondent's personal interest in the survey topic, various response decisions will ensue which have a different impact on data accuracy in terms of item missing data and various kinds of systematic response tendencies (Beatty & Herrmann, 2002; Krosnick, 1991; Peytchev, 2009).

Peytchev's (2009) framework of Web survey participation provides an integrative model for the theoretical explanation of response decisions in Web

surveys. These response decisions refer to (1) a respondent's decision whether to continue with the questionnaire or not, with the latter finding its expression in survey breakoff, (2) a respondent's decision on providing a substantive answer to a survey question or leaving it blank, resulting in item nonresponse, and (3) a respondent's decision on how thoroughly to answer the survey question, being reflected in the measurement properties of substantive answers. Each of these response decisions can be regarded as the result of some common but also unique causes. Relevant factors influencing these response decisions are assigned to (1) respondent-related characteristics and (2) method-related characteristics which can be further divided into characteristics of (2a) questionnaire design and (2b) questionnaire administration (Viswanathan, 2005, pp. 143-147). Whereas method-related factors are under the control of the survey researcher, factors related to the respondent typically cannot be directly influenced by the researcher (Baumgartner & Steenkamp, 2006; Beatty & Herrmann, 2002; Biemer & Lyberg, 2003, p. 116ff; Groves, 1989, p. 295; Peytchev, 2009).

In addition to Peytchev's (2009) framework of Web survey participation, Beatty and Herrmann's (2002) response decision model referring to surveys in general can be used for explaining the occurrence of item missing data. Moreover, Krosnick (1991, 1999) and colleagues' (Krosnick & Alwin, 1987) theory of survey satisficing is applied to explain mechanisms influencing measurement properties of the respondents' answers in greater detail as a function of respondent ability, respondent motivation, and task difficulty. In general, when respondents take part in a survey, they pass through several cognitive processes which confront them with tasks of different levels of complexity (Beatty & Herrmann, 2002; Krosnick, 1991; Peytchev, 2009). For a more detailed explanation of these cognitive processes and their underlying mechanisms, the three explanatory approaches refer to the more general framework of the question-answer process described above. It comprises the four steps of (1) comprehending and interpreting the survey question, (2) retrieving relevant information, (3) recalling or computing a judgment, and (4) formatting and editing the survey response. According to the question-answer process, survey responding is deemed a four-stage process, with measurement error in terms of incomplete and inaccurate responses potentially occurring within each stage (Beatty & Herrmann, 2002; Krosnick, 1991; Sudman, et al., 1996, p. 56ff; Tourangeau, et al., 2000, p. 165ff).

Within the first stage of the question-answer process, a respondent has to understand and interpret the survey question and its intended meaning. Whether a respondent provides a substantive answer and continues with the questionnaire principally depends on a respondent's ability and motivation to understand the question content. If respondents fail to understand the question meaning because of their insufficient ability and/or motivation, they will leave the answer to a question blank or even quit the survey (Beatty & Herrmann, 2002; De Leeuw, et al., 2003; Ganassali, 2008; Peytchev, 2009). Or, alternatively, respondents will provide an answer although a sufficient understanding of the question is not at hand. In order to give a substantive answer without having a clear understanding of the question meaning, respondents may be tempted to respond randomly or fall back on systematic response tendencies, resulting in either random or systematic measurement error (Ganassali, 2008; Lenzner, et al., 2010; Podsakoff, et al., 2003; Schober & Conrad, 1997; Viswanathan, 2005, p. 145; Weijters & Baumgartner, 2012). A third possibility would be that even if a respondent has an appropriate interpretation of the question on hand, a respondent will not be willing to spend the effort required to provide an answer or to give too much personal information. As a result, he or she may quit the questionnaire, leave the question blank, or deliberately give an inaccurate answer (Beatty & Herrmann, 2002; Peytchev, 2009; Shoemaker, et al., 2002; Tourangeau & Yan, 2007).

In the second stage, a respondent faces the task of retrieving relevant information from his or her memory. In attitude questions, it is mostly the case that a judgment has to be computed first on the basis of the information retrieved before an answer can be provided. Thus, in the third stage, a judgment needs to be computed on the basis of the relevant information retrieved from long-term memory as well as from the situational context. In this respect, information which is most accessible "is usually the information that has been used most recently, for example, for the purpose of answering a preceding question" (Schwarz, Strack, & Mai, 1991, p. 5). According to Beatty and Herrmann (1995, 2002), four different cognitive states are distinguished differing in the extent to which the information requested is available. A respondent's ability and motivation to provide complete and accurate responses vary depending on the respective state of knowledge. A respondent's knowledge can range between: (1) available (no apparent effort is needed to retrieve information that is known and available), (2) accessible

(special effort is needed to retrieve information that is not known but not instantly available), (3) generatable (information is not known but an attitude can be generated on the basis of other information), and (4) inestimable (information is neither known nor is there other information to generate an answer). Considering state (1), item nonresponse is due to a respondent's unwillingness to disclose his or her 'true' answer even though it is available. This commonly occurs when asking respondents for personal or particularly sensitive information and respondents try to avoid a potentially embarrassing situation by refusing to answer. In this case, item missing data is not considered to be missing at random. Alternatively, respondents provide an answer either on a random basis or based on systematic response tendencies. By contrast, if an answer is neither known nor can be computed on the basis of other information in state (4), item nonresponse would be the only appropriate response decision. However, at least some respondents still provide an answer. In this instance, providing an answer without the knowledge of relevant information reflects a misrepresentation of the respondent's knowledge, and again, results in random or systematic measurement error. In the more common intermediate states (2) and (3), the information requested is not known but can be retrieved and computed which, however, entails greater cognitive effort. Hence, whether an answer is provided at all, whether this answer is accurate, and whether the respondent continues with the questionnaire is again a matter of respondent motivation (Beatty & Herrmann, 1995, 2002; Peytchev, 2009). If respondents are unwilling to spend the effort required to answer a survey question, they may instantly decide to refuse an answer or actually drop out before even starting cognitive processing (Peytchev, 2009; Shoemaker, et al., 2002). If, however, respondents decide to provide an answer, inaccurate responses are likely, in case of performing the stages of the question-answer process merely superficially (weak satisficing), or in case of actually skipping the entire retrieval and judgment steps (strong satisficing) (Krosnick, 1991; Paulhus, 1991).

Once a respondent has come to a judgment, he or she needs to report this judgment in terms of formatting and editing the response in the fourth stage of the question-answer process. When answering closed questions in self-administered surveys, a respondent has to format the answer by mapping their judgment onto the response options provided, without the assistance of an interviewer. In this regard, a respondent is faced with certain constraints

attributable to the question format and its predefined response options. If a respondent fails to assign his or her answer to a response option because an appropriate response option is simply missing, this will result in item nonresponse or even survey breakoff. In either case, the item missing data is not missing at random (De Leeuw, et al., 2003). Otherwise, further measurement errors can take effect and impair data accuracy. A lack of appropriate response options may encourage respondents to randomly select between the response alternatives provided. Furthermore, the predefined response options may encourage respondents to rely on various kinds of systematic response tendencies, resulting in a systematic preference of, among others, the most positive or the most extreme response options (Krosnick, 1991; Paulhus, 1991; Podsakoff, et al., 2003). Other systematic response tendencies refer to a respondent's susceptibility to edit the final response in terms of "consistency, acceptability, desirability, or other criteria" (Podsakoff, et al., 2003, p. 886). For instance, social desirability can result in the systematic underreporting of socially undesirable information and a systematic overreporting of socially desirable information, inducing systematic measurement error (Paulhus, 1991; Sakshaug, et al., 2010; Shoemaker, et al., 2002; Tourangeau & Yan, 2007).

2.2 Perceptual Information Processing

2.2.1 Visual Perception and Attention in Survey Responding

According to the question-answer process described in section 2.1, respondents first need to comprehend a survey question in order to retrieve relevant information, recall or compute an appropriate judgment, and report an accurate answer. Whereas in interviewer-administered surveys, the interviewer plays a crucial role with respect to the comprehension of information that is predominantly conveyed verbally, this responsibility pertains solely to the respondent in self-administered surveys where information is commonly presented visually (Cannell, et al., 1981). In general, the comprehension of information presented visually presupposes prior visual perception of and attention to this information. Based on the work of Dillman and colleagues (1997; 2002), an additional step of 'perceiving and attending' which precedes the other four stages of the question-answer process is added in self-administered surveys and Web surveys as well.

Within this preceding stage of perceptual processing, respondents are expected to pass through several steps of visual perception. These steps are basically assigned to the stage of pre-attentive and attentive processing of the verbal and visual (or nonverbal) features of a questionnaire. Hence, a distinction can be made between the macro level of an entire questionnaire and the micro level of individual survey questions (Ganassali, 2008; Jenkins & Dillman, 1997; Toepoel & Dillman, 2010).

Within the initial stage of pre-attentive processing, respondents perceive the features of a survey at a holistic level. This rather superficial processing of the survey proceeds automatically and often even unconsciously. That is why it does not require much effort from the respondent. As a result of pre-attentive processing at both the questionnaire and question level, respondents should ideally be able to comprehend the overall structure of a questionnaire and of individual survey questions at first glance, just as they should understand the general instruction on how to navigate through the questionnaire and provide an answer to individual survey questions (Jenkins & Dillman, 1997). Already at this stage of pre-attentive processing, questionnaire-related and administration-related characteristics have great impact on the respondent's motivation to continue with the survey and to provide complete and accurate responses to the survey questions because the respondents get a first impression of the length and visual appearance of the questionnaire as well as a rough impression of the extent of effort required to answer the survey questions (Ganassali, 2008; Jenkins & Dillman, 1997; Vicente & Reis, 2010). Accordingly, in interactive Web surveys where the survey questions are generally presented page-by-page, it is frequently the case that the highest proportion of survey breakoff occurs within the first few Web pages of the questionnaire. For example, Ganassali (2008) showed that, irrespective of the overall questionnaire length, about 50% of all respondents who broke off over the course of survey completion abandoned the questionnaire already on the second page. Thus, a brief glance at the questionnaire within the first few seconds can decide about the respondents' impression and participation in a Web survey.

In general, after the respondents have gained a first impression of the questionnaire, they need to comprehend the questionnaire details by paying attention to aspects of the wording and the layout of survey questions in the second stage of attentive processing of verbal and visual features of the questionnaire. By focusing on smaller areas of the visual field, i.e., individual

words or input fields of a survey question, respondents can consciously analyze the information presented (Dillman, et al., 2009, p. 168; Jenkins & Dillman, 1997). A respondent is mainly expected to start from the beginning of a questionnaire and to consciously process and answer all survey questions in the prescribed order as intended by the survey researcher, instead of answering the survey questions in a more or less arbitrary order without having fully read and understood them (Jenkins & Dillman, 1997). By means of paying sufficiently high attention to the relevant verbal and visual features of the survey questions, the respondent's ability to obtain a clear understanding of the question meaning, the kind of data being expected, and the respective data input requirements should be enabled (Ganassali, 2008; Jenkins & Dillman, 1997; Toepoel & Dillman, 2010).

However, during both pre-attentive and attentive processing, different components of a survey question are competing for the respondents' attention, with respective question components varying in the extent to which they are spontaneously noticed and carefully attended to by the respondents (Jenkins & Dillman, 1997; Redline & Dillman, 2002; Tourangeau, Couper, & Conrad, 2007b). Since a respondent's sharp vision is limited to about 2 degrees or 8 to 10 characters, a respondent's focus of attention is severely limited (Kahneman, 1973, p. 50). Moreover, not every information provided to a respondent is of equal interest, which is why the respondent's attention is likely to vary between the various survey question components depending on their perceived relevance (Jenkins & Dillman, 1997; Tourangeau, et al., 2007b). Consequently, the visibility or visual prominence of the question components can differ, which in turn affects whether respondents perceive and use the information provided, as well as how they use this information (Redline & Dillman, 2002; Tourangeau, et al., 2007b). However, the visibility of the respective survey question components and the amount of attention respondents assign to each component can be increased or decreased by the way the information is presented, i.e., by means of visual questionnaire features, such as the element's size, color, or location (Christian, et al., 2009; Dillman, et al., 2009, p. 91).

In self-administered surveys, no interviewer is present who can encourage the respondents over the course of the questionnaire completion to give sufficient attention to all relevant components of a survey question. Furthermore, no interviewer is available to render assistance on how respondents have to navigate through the questionnaire and answer the survey

questions, or to remove any uncertainty concerning the meaning of a question and its response options. Respondents rather draw on verbal and visual questionnaire features to gain additional information about each survey question and its meaning, about general requirements concerning navigation within the questionnaire, and about special requirements concerning data input (Christian & Dillman, 2004; Christian, et al., 2009; Couper, et al., 2001; Heerwegh, 2009; Schwarz, 1990; Schwarz, Grayson, & Knäuper, 1998; Schwarz, Strack, & Mai, 1991; Toepoel & Dillman, 2008, 2010; Tourangeau, Couper, & Conrad, 2007a). Therefore, a better understanding of the impact of verbal and visual questionnaire features and their specific use may help motivate respondents to engage in a survey and to spend sufficient effort during questionnaire completion, keeping respondent burden within reasonable limits. Furthermore, verbal and visual questionnaire features can be used to help respondents gain a clearer understanding and enable a smooth processing of the questionnaire as a whole as well as of each survey question, by drawing respondent attention to the relevant question components and providing guidance throughout the entire questionnaire (Jenkins & Dillman, 1997; Smyth, Dillman, Christian, & Stern, 2006; Toepoel & Dillman, 2010).

In Web surveys, the technical opportunities are available for the implementation of a variety of visual questionnaire features. That is why these visual questionnaire features have increasingly been used to assist respondents by improving navigation, to motivate them by making the survey more interesting and enjoyable, and to ensure a better question comprehension by drawing the respondents' attention to the key components of a survey question. Altogether, these aspects lay the foundations for complete and accurate survey responses (Jenkins & Dillman, 1997; Toepoel & Couper, 2011; Toepoel & Dillman, 2010; Tourangeau, et al., 2004). Thus, besides an appropriate use of verbal questionnaire features, a targeted implementation of visual features may determine the accuracy of the respondents' answers, whether respondents provide an answer at all, and whether they continue until the questionnaire is completed (Dillman, et al., 1998; Ganassali, 2008; Jenkins & Dillman, 1997; Toepoel & Dillman, 2010; Vicente & Reis, 2010). Consequently, successive stages of the respondent's perception of verbal and visual questionnaire elements and their impact on response decisions need to be examined in detail, with the visual issues being of specific importance for the design and administration of self-administered questionnaires in order to prevent the risk of item missing data and impaired

measurement properties. In the following, the focus is primarily on visual questionnaire features, with some studies exemplifying their possible applications and the respondents' processing of such features.

2.2.2 Verbal and Visual Questionnaire Features

Verbal and visual questionnaire features are assigned to the categories of verbal language in terms of spoken or written words and visual (or nonverbal) language. Verbal language as the predominant form of communication in all modes of data collection is related to the words being used in survey questions, instructions or other auxiliary verbal information, with careful wording in terms of clear, understandable, and unambiguous verbalization being essential to convey the correct question meaning. By contrast, visual language including numerical, symbolic, and graphical languages refers to the visual presentation of words (Christian & Dillman, 2004; Jenkins & Dillman, 1997). Particularly in self-administered questionnaires when information is usually presented visually, visual language is of special importance because the visualization of survey questions and their components "may affect whether questions are read, the order in which they are read, and the meaning conveyed to respondents" (Christian & Dillman, 2004, p. 59). Thus, visual cues can be used in conjunction with verbal language, whereas previous research showed that both verbal and visual languages influence respondent behavior separately or in interaction with each other (Christian & Dillman, 2004; Christian, et al., 2009; Jenkins & Dillman, 1997; Redline & Dillman, 2002; Smyth, et al., 2006; Toepoel & Couper, 2011; Toepoel & Dillman, 2008, 2010; Tourangeau, et al., 2004, 2007a).

When speaking about visual (or nonverbal) language, a distinction is made between numerical, symbolic, and graphical languages (Redline & Dillman, 2002). Numerical language conveys meaning in addition to verbal language by the use of numbers. For example, assigning negative numbers (-1 to -5) instead of positive numbers (+1 to +5) to the response options of a rating scale can alter a respondent's interpretation of the response scale which is reflected in fewer respondents selecting a response option from the low end of the scale when negative numbers are added (Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991; Tourangeau, et al., 2007b). Based on these findings, Schwarz and colleagues (1991) concluded that respondents use

numerical response scale labels to disambiguate the meaning of the verbal response scale labels.

Symbolic language refers to the use of signs and symbols with shared cultural meaning. For example, arrows can be used to clarify the navigational path and to emphasize skip instructions directing to a follow-up question. Studies have shown that the use of such symbols can provably reduce the proportion of unintentional item nonresponse or mistakenly answered survey questions not provided for some of the respondents (Christian & Dillman, 2004; Redline & Dillman, 2002).

Graphical language acts like a form of paralanguage and includes graphical features, “such as size, brightness and color, shape, location, and spatial arrangement of words, numbers, and symbols” (Christian & Dillman, 2004, p. 59). In this regard, it is important to note that in self-administered questionnaires, the verbal, numerical, and symbolic languages are always transmitted visually by means of graphical language. And, even in the case of almost identical verbal, numerical, and symbolic languages, minor differences in the graphical language of a survey question can be sufficient to provoke differences in the respondents’ interpretation and their answers to a survey question, respectively. (Redline & Dillman, 2002). For example, whether verbal clarification features such as definitions or instructions are placed before or after the question text may affect the likelihood that this additional information is recognized as a relevant part of a survey question and the order in which different question components are processed. This in turn can have differing effects on data accuracy (Christian & Dillman, 2004; Kunz & Fuchs, 2012; Redline & Dillman, 2002). Depending on the size of answer boxes in open-ended questions, respondents may alter their answer length, the number of themes they fill in, and their elaboration of answers (Christian & Dillman, 2004; Emde & Fuchs, 2012; Smyth, Dillman, Christian, & McBride, 2009). As related to the design of rating scales, the respondents’ interpretation of the rating scale items and response options may be altered depending on variations in the visual presentation. For example, presenting several rating scale items sharing the same response options on a single screen, compared to placing each item on separate screens, can lead to more similar answers and is likely to increase the number of missing values (Toepoel, et al., 2009b). Or, the likelihood that a response option is selected is depend on the order in which response options are arranged in a rating scale, i.e., whether the positive scale end is presented first or vice versa (Malhotra, 2009). These are

just two examples of how variations in the graphical language, in this case changes in the context of rating scale items and in the arrangement of response options, affect the measurement properties of rating scales. In Chapter 4, additional findings related to verbal and visual characteristic of rating scale design and administration are discussed.

Although visual questionnaire features, such as the spacing of different question components, their order, or the grouping of survey questions evidently have an effect on the respondents' interpretation and answering of survey questions, previous studies also suggested that there is "a hierarchy of features that respondents attend to, with verbal labels taking precedence over numerical labels and numerical labels taking precedence over purely visual cues" (Tourangeau, et al., 2007a, p. 109). Features of graphical language, such as size, color, spacing, or location as kinds of purely visual cues seem to have a noticeable impact on the respondents' answers when either verbal language is not clear and leaves any uncertainty concerning the meaning of a survey question or its response options, or when no rival cues in terms of verbal or numerical language are present (Toepoel & Dillman, 2008, 2010; Tourangeau, et al., 2007a). Thus, respondents will attend to visual questionnaire features, in particular if they search for additional information about the survey question, or if their attention is not distracted by other question components with a higher perceived visual prominence. In order to get a clearer understanding of the respondents' interpretation of visual questionnaire features in Web surveys, different processing stages and cognitive heuristics are distinguished and described in the next section.

2.2.3 Processing of Visual Questionnaire Features

In self-administered surveys in general, and more specifically, in interactive Web surveys where the survey questions are mainly presented page-by-page, respondents pass through three stages of processing visual questionnaire features (Dillman, et al., 2009; Jenkins & Dillman, 1997; Toepoel & Dillman, 2010). Within the initial stage of pre-attentive processing, respondents use visual properties such as luminance, color, and texture to subdivide each of the Web pages of a survey into basic regions. As a result, respondents get a first overview of the general structure of a respective Web page. Within the second intermediate step of element organization, respondents divide the Web page in basic regions and establish relations between the different visual

elements. For this purpose, page elements are grouped together which helps respondents improve their understanding of a survey question and accelerate its processing. This step of pattern recognition is accomplished through both pre-attentive and attentive processing and is explained by the design principles of Gestalt psychology, two of which are relevant in the context of the present study: first, the principle of proximity which refers to objects that are perceived as a group because they are presented closely together; second, the principle of similarity which refers to objects that are perceived as a group because they have similar properties in terms of, among others, their shape, size, contrast, or orientation. In the third stage of attentive processing, it is generally assumed that the respondents follow five cognitive heuristics in interpreting the visual features of survey questions: (a) middle means typical, (b) left and top mean first, (c) up means good, (d) near means related, and (e) like means close (Tourangeau, et al., 2004). In this stage of task-oriented processing and task completion, respondents use these cognitive heuristics in interpreting and answering survey questions, with the response task being facilitated and the extent of respondent burden being decreased (Toepoel & Dillman, 2010; Tourangeau, et al., 2004). In general, the respondents' use of these cognitive heuristics can have differing effects on the interpretation and answering of survey questions as ascertained, among others, by Tourangeau and colleagues (2004). That is why these cognitive heuristics should be kept in mind when designing a Web questionnaire. In the following, the cognitive heuristics are explained with reference to the design of rating scales.

According to heuristic (a) 'middle means typical', respondents deem the middle item of an item list or the middle option of a set of response categories as the most typical one. Thus, respondents use the middle item or response category as an anchor, or reference point for making their judgment (Tourangeau, et al., 2004). An example concerning a respondent's visual processing of the rating scale response options was provided by Tourangeau and colleagues (2004) showing that when the visual midpoint of a rating scale deviates from the conceptual midpoint (e.g., when a nonsubstantive 'don't know' option is added but not visually separated from the substantive options, when response options are unevenly spaced), respondents seem to use the visual, rather than the conceptual midpoint of the scale as a reference point for their responses as reflected in the respondents' answers shifting towards the visual midpoint.

Heuristic (b) ‘left and top mean first’ describes the fact that respondents consider the leftmost or topmost response options in a horizontally or vertically arranged set of response options as the conceptually first one, while the opposing response option represents the conceptual opposite, and the remaining response options in between follow a logical order (Tourangeau, et al., 2004). In accordance with their hypothesis, Tourangeau and colleagues (2004) found longer response times in an agree-disagree rating scale when the order of the response options deviated from the order prescribed by this heuristic which is indicative of the respondents’ problems in processing response options deviating from an assumed logical order.

Heuristic (c) ‘up means good’ is a variant of heuristic (b). One implication for the design of rating scales is that respondents expect the response options to appear from positive to negative, with the positive end of the scale to be presented first, i.e., topmost in vertically and leftmost in horizontally arranged rating scales. Designing Web surveys in compliance with this heuristic can facilitate the response task as indicated by the respondents’ answers taking less time and being more reliable. At the same time, deviations from this heuristic entail additional cognitive effort to process the information that is not presented in the expected manner (Christian, et al., 2009; Tourangeau, et al., 2004). However, compliance with heuristic (b) and (c) can also provoke systematic response tendencies in terms of primacy effects (see section 4.3.2) because one of the first response options are more likely to be selected in a rating scale than later response options (Christian, et al., 2009; Toepoel, et al., 2009a; Toepoel & Dillman, 2008).

Heuristic (d) ‘near means related’ states that respondents expect items that are listed close to each other on a Web page to be in fact conceptually related. This heuristic draws on the design principle of proximity as objects are grouped together conceptually when these objects are presented closely together (Tourangeau, et al., 2004). One implication of the above for the design of rating scales is that respondents may see stronger interconnections among items that are arranged on a single Web page than among items that are presented separately on several Web pages. Among others, Tourangeau and colleagues (2004) found initial evidence of this kind of proximity effect (or grouping effect) (see section 4.2.2).

The final heuristic (e) ‘like means close’ traces back to the design principle of similarity and describes the fact that respondents expect items or response options that are visually similar as conceptually closer (Tourangeau,

et al., 2004). By implication, using different numerical labeling of the rating scale response options (e.g., numerical labels ranging from -5 to +5 compared with 1 to 11) or variations in visual labeling (e.g., shadings ranging from dark red to dark blue compared with dark blue to light blue) results in endpoints being perceptually more distinct and hence, responses being more extreme as indicated by means shifted to the right side of the scale when the endpoints differ in color or in sign and value (Schwarz, Knäuper, et al., 1991; Toepoel & Dillman, 2008; Tourangeau, et al., 2007a).

The consideration of these cognitive heuristics in developing general principles for questionnaire design and administration can help promote complete and accurate responses to survey questions because the understanding of survey questions can be enhanced, while the extent of respondent burden can be kept within reasonable limits. However, the implementation of these simplified heuristics can also increase the respondents' susceptibility to cognitive shortcuts and the risk of misinterpreting survey questions, which in turn is likely to induce measurement error further discussed in section 4.3 (Toepoel & Dillman, 2010; Tourangeau, et al., 2004).

In general, detailed knowledge of the respondents' cognitive and perceptual information processing helps understand response behaviors in surveys. As the above considerations have shown, the likelihood of complete and accurate survey responses is decisively determined by respondent ability, respondent motivation, and task difficulty. Whereas task difficulty depends to a great extent on specific characteristics of questionnaire design and administration which can be directly influenced by a survey researcher, task difficulty is an important factor influencing a respondent's ability and motivation to provide complete and accurate survey responses. Hence, when striving for the construction of a respondent-friendly questionnaire in order to ensure high data accuracy in Web surveys, it is important to understand how task difficulty, respondent ability and respondent motivation to attentively and carefully process survey questions can be positively affected by means of verbal and visual characteristics of questionnaire design and administration discussed in the next sections.

3. DESIGN AND ADMINISTRATION OF QUESTIONNAIRES IN WEB SURVEYS

In the design and administration of a Web survey, the primary objective is the implementation of a respondent-friendly questionnaire generally defined as “the construction of web questionnaires in a manner that increases the likelihood that sampled individuals will respond to the survey request, and that they will do so accurately” (Dillman, et al., 1998, p. 3). Respondent-friendliness can also be described in terms of “a form that is easy for respondents to complete, avoids confusion about what or how to answer it, and results in respondents feeling neutral or positive, as opposed to negative, about the form itself” (Dillman, Sinclair, & Clark, 1993, p. 290). Against the background of the theory of survey satisficing, respondent-friendly questionnaire design aims at ensuring an appropriate level of task difficulty that corresponds with the respondents’ abilities and keeps their motivation at a high level over the course of survey completion, which in turn reduces the respondents’ susceptibility to systematic response tendencies and the risk of item missing data (Couper, 2000; Dillman, et al., 1998; Jenkins & Dillman, 1997; Krosnick & Presser, 2010).

In general, the characteristics of both questionnaire design and questionnaire administration determine the level of task difficulty. The complexity of the response task, in turn, interacts with the respondent’s ability and motivation to provide complete and accurate responses, with the latter being decisively influenced by the extent of the respondent burden perceived at the macro level of the entire questionnaire and at the micro level of individual questions (Beatty & Herrmann, 2002; Galesic, 2006; Krosnick, 1991; Peytchev, 2009). Thus, incomplete and inaccurate survey responses may arise from several characteristics of the survey. Besides specific respondent-related characteristics, several factors related to questionnaire design and questionnaire administration are decisive for high data accuracy. In this regard, verbal and visual characteristics of questionnaire administration are particularly important for a respondent’s initial impression of the survey and his or her assessment of the extent of effort required to answer the questionnaire, while verbal and visual characteristics of questionnaire design determine more specifically the extent of effort required for answering each

individual survey question accurately (Jenkins & Dillman, 1997). Key factors related to questionnaire design and questionnaire administration are discussed in the following sections, potentially affecting the difficulty of the response task as well as respondent ability and respondent motivation to provide complete and accurate survey responses.

3.1 Characteristics of Questionnaire Design

In interactive Web surveys, factors related to the characteristics of questionnaire design are page-specific, as survey questions are commonly presented page-by-page. Thus, each survey question will be visible only after the respondent has moved to the next page. That is why verbal and visual questionnaire features primarily affect a respondent's decision to continue with the survey rather than his or her initial decision to participate in the survey (Peytchev, 2009). Presenting survey questions page-by-page also implies that respondents may instantly re-evaluate their initial decision to provide complete and accurate responses on every new Web page, which in turn is influenced by the verbal and visual features of a survey question (Galesic, 2006; Peytchev, 2009). These verbal and visual features decisively influence the difficulty of the response task and a respondent's ability and motivation to overcome these difficulties (Christian, et al., 2009; Couper, 2008; Dillman, et al., 2009; Dillman, et al., 1998). With respect to the broad category of verbal and visual questionnaire features applied in questionnaire design and administration, these can be further classified according to content-related and form-related characteristics of the questionnaire design (Alwin, 2010; Ganassali, 2008; Viswanathan, 2005, p. 143).

3.1.1 Question Content

In general, questionnaires and interviews in survey research are applied to gather data concerning factual information which is related to socio-demographic questions and other knowledge-based issues as well as nonfactual information including attitudes, beliefs, values, and self-descriptions. Whereas factual questions refer to objective information that is characterized by a certain degree of clarity and specificity and is verifiable against objective records (e.g., date of birth, educational achievement, household income, and quantifiable information with regard to certain

behaviors), nonfactual questions ask for subjective information which requires personal judgment concerning more abstract concepts or constructs (Alwin, 1989; 2007, pp. 122-124; Kalton & Schuman, 1982).

Concerning nonfactual information, a distinction can be made between beliefs, values, attitudes, and self-descriptions. Beliefs, as subjective assessments made about the physical and social environment that an individual generally holds to be true ('what is') and values, as higher-level standards ('what is good') guiding behavioral choices by determining desirable aims and legitimate means of attaining them, are both relatively stable over time and situations. In contrast, attitudes as expressions of a positive or negative evaluation towards an attitude object (e.g., person, object, event, activity, idea) in terms of acceptance, favorability, or agreement are derived from more basic beliefs and values. Attitudes are more strongly influenced by situational factors and thus, less stable over time (Alwin, 2007, p. 124; Chaiken & Stangor, 1987; Tourangeau & Rasinski, 1988; Tourangeau, et al., 2000, pp. 169-170). Three components of attitudes can be distinguished with an affective, cognitive, and behavioral dimension, whereby all the various attitudes have in common the fact that they are evaluative in nature (Alwin, 2007, p. 124; Chaiken & Stangor, 1987). As distinct from attitudes where the evaluation object is external and not the individual's self, self-descriptions refer to either self-evaluations of personal traits, experiences, attitudes or behaviors, or to self-perceptions without any implied evaluative dimension (Alwin, 2007, p. 124). In the following, the term attitude question refers to nonfactual questions comprising attitudes towards another person or an object as well as self-descriptions in terms of attitudes towards oneself (Alwin, 2007, p. 124).

Attitudes and self-descriptions as kind of nonfactual information most frequently measured in surveys are of primary interest to survey researchers because the underlying theoretical concepts and constructs regarding the respondent and regarding other persons or objects cannot be observed and measured directly. For that reason, a lot of design issues have to be considered when constructing attitude scales and personality measures (Kalton & Schuman, 1982). As compared to factual questions, the respondent's answers to nonfactual questions are less reliable in general. Moreover, answers to survey questions assessing attitudes or self-descriptions are even less reliable compared to answers to questions on beliefs or values (Alwin, 1989; 2007, p. 149). One reason for the reduced reliability of the

respondent's answers is that in most instances answers to an attitude question are based on information that is retrieved in the respective situation rather than based on ready-made judgments, which is why "such evaluations are often easily manipulated and are subject to situational factors" (Alwin, 2007, p. 124; Tourangeau & Rasinski, 1988; Tourangeau, et al., 1991). Furthermore, low reliability of responses to attitude questions is explained by a lack of clarity and specificity in question meaning since nonfactual constructs are commonly more abstract and complex than factual issues. Therefore, an adequate translation into carefully defined indicators allowing for the precise and accurate measurement of a theoretical construct is more difficult (Alwin, 2010; Kalton & Schuman, 1982). Another reason often quoted is the fact that respondents may have difficulties translating their internal cues into the conceptual framework, or to be precise, their judgment into the response options provided (Alwin, 2007, p. 149; 2010). Thus, respondent ability and respondent motivation to provide complete and accurate responses is principally affected by the difficulty of a survey question and the cognitive effort required to cope with the complexity of the response task, which in turn is in large part related to the content of a question, or more precisely, the accessibility, sensitivity, and clarity of question content. While aspects of question accessibility are central to the present study and addressed again in section 4.3.1, aspects of question sensibility and clarity are briefly mentioned for the sake of completeness.

Question Accessibility

When asking respondents for nonfactual information on attitudes or self-descriptions, they usually have no ready-made judgment available. Instead, survey researchers have to rely on the respondents' ability and motivation to retrieve the relevant information and—based on this information—to make an informed judgment (Tourangeau, et al., 1991). Thus, the respondents' answers to attitude questions are decisively affected by information that is temporarily accessible and that can be easily retrieved from long-term memory and from the context of respective survey questions (Schwarz, Strack, & Mai, 1991; Sudman, et al., 1996, p. 70; Tourangeau, et al., 1991).

Temporal accessibility of relevant information determines the effort needed to retrieve the information and to compute a judgment concerning the issue in question. Temporal accessibility in turn depends on the salience of question content and the frequency and recency of prior activation:

Depending on whether the issue in question is of high personal importance or relevance, and therefore, highly salient for a respondent, whether a respondent has discussed or thought about this issue several times before, or whether a respondent has considered this issue only recently, determines the accessibility of relevant information. And this in turn has immediate implications for the complexity of the response task and a respondent's ability and motivation to invest the effort necessary to answer a survey question accurately (Groves, 1989, p. 432; Schwarz, Strack, & Mai, 1991; Tourangeau, et al., 1991; Tourangeau, et al., 2000, pp. 172-180). By implication, asking for issues being inaccessible because information is of low personal relevance for respondents or even not applicable to them reduces a respondent's ability and motivation to retrieve relevant information and may increase the risk of item missing data and systematic response tendencies, resulting in reduced data accuracy (Alwin, 1989; Beatty & Herrmann, 2002; Herzog & Bachman, 1981; Messick, 1967; Peabody, 1961; Peytchev, 2009; Tourangeau, 1992).

In general, the cognitive effort required to answer attitude questions is greater since respondents usually have to compute a judgment first before they can answer the questions. Consequently, this type of judgment-intensive survey questions per se carries an increased risk of item nonresponse and survey breakoff compared to survey questions requiring less effort (Beatty & Herrmann, 2002; Krosnick, 1991; Peytchev, 2009; Shoemaker, et al., 2002). This is explained by the fact that when respondents instantly realize the difficulty of a survey question, they decide to refuse an answer or actually drop out before even starting cognitive processing (Beatty & Herrmann, 2002; Lozar Manfreda & Vehovar, 2002; Peytchev, 2009; Shoemaker, et al., 2002). Alternatively, if respondents decide to provide a substantive answer although their motivation to spend a lot of cognitive effort on retrieval and judgment is low, respondents will abbreviate the question-answer process by shortening or even skipping the stage of information retrieval and judgment, resulting in satisfying rather than optimized answers, e.g., acquiescent responding, nondifferentiation, or selecting the 'don't know' option (Herzog & Bachman, 1981; Krosnick, 1991; Shoemaker, et al., 2002).

Temporal accessibility of relevant information, and thus, the ease of information retrieval and judgment on attitude questions can be increased by asking a set of several items on related topics or issues. By answering preceding survey questions on the same or related issue, the cognitive effort needed to retrieve relevant information for answering the current question can

be reduced. Likewise, judgment on attitude questions can be facilitated when questions related in content are presented consecutively (Schwarz, Strack, & Mai, 1991; Tourangeau, 1992; Tourangeau & Rasinski, 1988; Tourangeau, et al., 1991). Regarding rating scales, facilitation effects due to the thematic relationship between successive rating scale items were found in terms of a speedup in the retrieval process reflected in decreased response times for an item when respondents already answered an earlier item on the same topic or issue (Couper, et al., 2001; Knowles, 1988; Knowles & Byers, 1996; Tourangeau, 1992; Tourangeau & Rasinski, 1988; Tourangeau, et al., 1991). Facilitation effects and implications for measurement properties of rating scales are further discussed in section 4.3.1.

Question Sensitivity

Asking the respondents sensitive questions about issues that are perceived as too personal or socially undesirable also carries an increased risk of item nonresponse and impaired measurement properties because respondents are not willing to disclose personal information to third parties or to present themselves in an undesirable light (Beatty & Herrmann, 2002; Shoemaker, et al., 2002; Tourangeau & Yan, 2007). In this regard, misreporting on sensitive questions is primarily considered a motivated process, in which respondents edit their responses in order to avoid public exposure and invasion of privacy (Paulhus, 1991; Tourangeau & Yan, 2007).

Based on a convenience sample of 242 questions from various surveys including items asking about socially desirable and undesirable characteristics related to a respondent's academic performance and items neutral in content (e.g., age or years since graduation), Shoemaker and colleagues (2002) found a significantly positive correlation ($r = .18$) between question sensitivity and item nonresponse. The more sensitive a survey question was considered, the more respondents refused an answer to that question. Consistent with previous research on sensitive topics, such as income, personal traits, attitudes (e.g., racism, sexism), and behaviors (e.g., voting and energy conservation as socially desirable behaviors and alcohol, tobacco, or other drug use as socially undesirable behaviors), there is a positive correlation between sensitive questions and item nonresponse (Kays, Gathercoal, & Buhrow, 2012; Moore, Stinson, & Welniak, 1997; Tourangeau & Yan, 2007).

Instead of refusing an answer to such sensitive questions, respondents may systematically misreport relevant information and induce large biases

into survey estimates by consistently underreporting socially undesirable information, while socially desirable information is consistently overreported (Tourangeau & Yan, 2007). Sakshaug and colleagues (2010) found evidence for measurement error in terms of misreporting as the largest source of error for items asking for socially undesirable characteristics, whereas item missing data mostly occurred for items asking for socially desirable or neutral characteristics. Moreover, meta-analysis revealed that misreporting on sensitive questions occurs more frequently than refusing answers to sensitive questions (Tourangeau & Yan, 2007).

Question Clarity

In questionnaire design, the prime objective is the development of clear and specific survey questions that allow respondents an easy comprehension and answering which, however, is not always achieved in practice. In some cases, a lack of clarity or specificity in questionnaire measures is certainly due to a higher level of abstraction and complexity of some theoretical constructs (Alwin, 2010; Kalton & Schuman, 1982; Podsakoff, et al., 2003). More often, difficulties in understanding a survey question arise from complex, ambiguous, or imprecise question wordings (e.g., double-barreled questions using words with contextual or several different meanings), or from unfamiliar words. This in turn can be prevented by a more careful question construction, using simple syntax and unambiguous wording clearly specified by context (Alwin, 1991; Bradburn, 2004; Podsakoff, et al., 2003).

The key problem with unclear or ambiguous wordings of survey questions is that they entail difficulties in comprehension and high cognitive burden for respondents. Ambiguity in question meaning leaves room for the respondents' own idiosyncratic interpretations. This comes along with the fact that respondents have to expend greater cognitive effort to come to an adequate understanding and answering of a survey question (Lenzner, et al., 2010; Podsakoff, et al., 2003). This in turn can lower data accuracy because of variations in the interpretation of a survey question, because of cognitive shortcuts in cognitive processing, or because of the refusal of an answer (Ganassali, 2008; Lenzner, et al., 2010; Podsakoff, et al., 2003; Schober & Conrad, 1997; Viswanathan, 2005, p. 145). Hence, a lack of question clarity can impair data accuracy in terms of random responding or in terms of the increased risk of systematic response tendencies. Furthermore, the likelihood of item nonresponse is likely to be increased.

3.1.2 *Question Format*

Different kinds of survey question formats can be distinguished, each one having different implications for the occurrence of item missing data and impaired measurement properties. Besides open-ended numerical questions requiring short numeric information like dates, numbers, frequencies or counts within small text boxes (e.g., number of visits to the doctor in the past 12 months) and closed questions with unordered response options (e.g., marital status), closed rating scale questions with ordered response options (e.g., degree of satisfaction with life on a 5-point rating scale) are the question formats most commonly used in questionnaire design (Tourangeau, et al., 2000, p. 231). Although widely used, each of these question formats presents respondents with certain challenges. And, respondent ability and respondent motivation to provide complete and accurate responses depend in large part on the respective question format. In particular, open-ended questions and rating scale questions such as conventional grid questions or newer visual analogue scales considerably increase the risk of item nonresponse, survey breakoff, and in the case of rating scales, the risk of systematic response tendencies (Alwin & Krosnick, 1985; Beatty & Herrmann, 2002; Couper, et al., 2006; Crawford, et al., 2001; Funke, Reips, & Thomas, 2011; Galesic, 2006; Knapp & Heidingsfelder, 1999; Lozar Manfreda & Vehovar, 2002; Millar & Dillman, 2012; Peytchev, 2009; Van Dijk, Datema, Piggen, Welten, & Van de Vijver, 2009; Weijters, Cabooter, & Schillewaert, 2010). All these question formats have in common that they require more cognitive and navigational effort in providing an answer, or even special respondent abilities and particular software requirements as compared to other question formats (Beatty & Herrmann, 2002; Galesic, 2006; Ganassali, 2008; Knapp & Heidingsfelder, 1999; Krosnick, 1991; Krosnick & Presser, 2010; Peytchev, 2009). This increase in task difficulty entails an increase in perceived and actual respondent burden, and in the end, an increased number of respondents who are unable or unwilling to spend the necessary extent of effort to provide complete and accurate responses (Beatty & Herrmann, 2002; Ganassali, 2008; Krosnick, 1991; Peytchev, 2009). Thus, the decision on the format of a survey question needs to be carefully considered because of its great impact on the prevalence of item missing data and the impairment of measurement properties.

Open-ended versus Closed Question Formats

A major distinction in question formats is made between open-ended and closed question formats. In general, the advantages of open-ended question formats are that they allow respondents to provide their answers spontaneously; unaffected by predefined response options, thereby enabling a more diverse set of self-penned answers. Since predefined response options which may provide certain respondent orientation are missing, whilst at the same time, no interviewer is present who can probe for an answer in the desired format, the need for extensive coding and large item nonresponse rates are major disadvantages of open-ended question formats (Reja, et al., 2003). Thus, whether open-ended question formats yield adequate answers mainly depends on respondent motivation to think carefully about the issue in question and provide complete answers (Dillman, 2000, p. 41).

Although the specification of predefined response options implies certain limitations of possible respondent answers, a major advantage of closed question formats is the “standardization of response and economy of processing” (Kalton & Schuman, 1982, p. 49). In self-administered questionnaires, not the interviewer but the respondents themselves put their answers down in writing. Compared to open-ended questions, closed question formats offer predefined response options which provide orientation on how respondents are expected to answer the survey question (Reja, et al., 2003; Schwarz 1994). On the one hand, predefined response options facilitate the response task and make it easier for respondents to provide an answer. On the other hand, predefined response options steer respondents in a particular direction and tempt them to rely on systematic response tendencies, or in case of inappropriate response options, to leave the question unanswered (De Leeuw, et al., 2003; Kalton & Schuman, 1982; Schwarz 1994). So, the careful choice and the appropriateness of response options in closed questions decisively influence a respondent’s decision to provide complete and accurate responses.

Rating Scale versus Rank-Order Scale Formats

In questionnaire design, a major distinction is made between nominal categorical questions with unordered response categories (e.g., favorite leisure time activities) and ordinal categorical questions with ordered response categories (e.g., importance of different environmental protection issues). Rating scale questions as the best known example of the latter question format

are prevailing in attitude and personality measurement, while dichotomous questions (e.g., yes/no, agree/disagree) and rank-order scales (e.g., hierarchy from most to least important) are far less popular (Alwin & Krosnick, 1985; McCarty & Shrum, 2000).

When answering rating scales, respondents are asked to indicate their level of agreement or disagreement in terms of agree-disagree, like-dislike, or true-false to several statements about an issue, with these ordinal response options being identical for the whole set of items (Krosnick, Judd, & Wittenbrink, 2005; Preston & Colman, 2000; Schwarz, Knäuper, et al., 1991; Toepoel, et al., 2009a). As compared to dichotomous question formats solely distinguishing between agreement and disagreement without any gradation, rating scales allow respondents to differentiate in their responses and thus, to provide more refined data concerning the issues in question, while their cognitive burden can be kept within a limit owing to the use of identical response options (Bethlehem & Biffignandi, 2012, p. 155; Krosnick, et al., 2005). As opposed to rank-order scale formats, however, rating scales do not necessarily require the respondents to distinguish clearly between various items of the same rating scale. Hence, the respondents' answers to rating scales as a kind of non-comparative scaling technique often exhibit little differentiation (Alwin & Krosnick, 1985; McCarty & Shrum, 2000). Moreover, responses to rating scales are much more affected by various kinds of systematic response tendencies such as acquiescence or nondifferentiation than rank-order scales (Alwin & Krosnick, 1985; Krosnick & Alwin, 1988; McCarty & Shrum, 2000). However, major drawbacks of rank-order scales as a kind of comparative scaling technique are that respondents are forced to differentiate between the issues, even if no difference actually exists. Moreover, ranking procedures require considerably more respondent effort, are complex to administer, and take significantly longer time to be completed as compared to ratings scales (Alwin & Krosnick, 1985; McCarty & Shrum, 2000). Regarding the first argument, the question arises "whether the ranking approach may in fact create artificial contrasts among the latent content of the measures" (Alwin & Krosnick, 1985, p. 549; Krosnick & Alwin, 1988). Furthermore, the analysis of rank-order data is limited to the use of nonparametric statistical procedures, whereas rating data allows for the use of parametric statistical procedures because rating scales are commonly treated as an interval scale, not as an ordinal scale (Alwin & Krosnick, 1985; Brill, 2008; McCarty & Shrum, 2000). Despite the fact that both scale formats have

advantages but also certain disadvantages, rating scale formats are more prevalent in survey practice than rank-order scale formats. This is particularly due to the fact that rating scale formats facilitate a faster and less burdensome scale completion (Alwin & Krosnick, 1985; McCarty & Shrum, 2000).

3.2 Characteristics of Questionnaire Administration

The characteristics frequently cited in the context of questionnaire administration include the type and amount of incentives (Bosnjak & Tuten, 2003; Deutskens, Ruyter, Wetzels, & Oosterfeld, 2004; Galesic, 2006; Lozar Manfreda & Vehovar, 2002; O'Neil, Penrod, & Bornstein, 2003), pre-notification of the survey request (Kaplowitz, Hadlock, & Levine, 2004), and reminders to the survey request (Deutskens, et al., 2004; Kaplowitz, et al., 2004). However, these factors are more likely to determine initial consent to survey participation rather than affect the response decisions over the course of survey completion (Peytchev, 2009). The following remarks are confined to the characteristics of questionnaire administration that also have an effect on the item missing data and the measurement properties by primarily affecting overall respondent burden perceived during survey completion and respondent motivation. The basic decision for an interactive Web survey has implications for the general structure of a questionnaire (section 3.2.1), the length of a questionnaire (section 3.2.2), and the use of dynamic feedback such as process indication (section 3.2.3) and real-time validation (section 3.2.4). Situational context factors comprising question context (section 3.2.5) and other socio-environmental context factors (section 3.2.6) are of particular interest in the present study because these issues have a substantial impact on task difficulty as well as on the respondent's ability and motivation to provide complete and accurate survey responses.

3.2.1 Questionnaire Structure

In Web surveys, questionnaires can be implemented in a one-page design (scroll-based form) or in a multiple-page design like an interactive Web survey (screen-based form). In a one-page design, on the one hand, all survey questions are presented on a single Web page with a submit button at the end of the questionnaire which requires periodically scrolling by the respondent until the end of a questionnaire is reached. In a multiple-page design, on the

other, each survey question is placed on a single Web page with a submit button on every page. In this latter case, there is the advantage that survey data is simultaneously transmitted to the server and even in case of survey breakoff a respondent's answers are available for all questions answered before termination occurred (Crawford, et al., 2001; Peytchev, Couper, McCabe, & Crawford, 2006).

Since both page designs feature certain advantages, a general preference for one of the two has yet not been established. Nevertheless, a survey researcher's decision on the questionnaire structure may have differing implications, particularly for the occurrence of item missing data. On the one hand, in a one-page design respondents have an overview of the entire survey and receive an impression of the extent of effort required to answer the questionnaire. In a multiple-page design, however, respondents often can see only one survey question at a time (Dillman & Bowker, 2001; Ganassali, 2008; Norman, et al., 2001). Hence, if no explicit progress indicator is implemented in a multiple-page design, the respondents' lack of overview of the overall survey length and their progress within the questionnaire will enhance the risk of survey breakoff (see also section 3.2.3) (Couper, et al., 2001; Dillman & Bowker, 2001; Lozar Manfreda, et al., 2002; Peytchev, et al., 2006). On the other hand, one-page designs require periodical scrolling in order to process the questionnaire. And, although scrolling should facilitate the navigation within a questionnaire, respondents often perceive scrolling more burdensome than button navigation by mouse clicking. In this regard, more burdensome scrolling is likely to increase the risk of survey breakoff (Schonlau, et al., 2002). Moreover, periodical scrolling may result in the respondent's confusion about navigation and loss of position within the questionnaire which particularly applies to less computer literate respondents. A respondent's confusion about navigation as a result of presenting all survey questions on a single page makes it more likely that respondents accidentally skip a question (Norman, et al., 2001). Thus, although one-page designs enable a complete overview of the entire questionnaire, this may be at the expense of the ease of navigation within the questionnaire, potentially resulting in increased item nonresponse or even increased survey breakoff. One drawback of multiple-page designs is that overall completion times may be increased. Additional time due to longer page loading times, more time-consuming navigation requiring extra clicks, and constant reorientation on following Web pages increase the respondent burden perceived during survey

completion, which in turn may increase the risk of quitting the questionnaire prematurely (Couper, et al., 2001; Dillman & Bowker, 2001; Lozar Manfreda, et al., 2002; Peytchev, et al., 2006; Thorndike, et al., 2009).

In fact, previous findings showed that item nonresponse rates are increased when implementing one-page designs compared to using multiple-page designs (Lozar Manfreda, et al., 2002; Vehovar, et al., 2002). Although multiple-page designs tend to result in higher breakoff rates, previous studies either failed to prove significant differences between multiple-page and one-page designs (Lozar Manfreda, et al., 2002; Peytchev, et al., 2006), or they found significant differences solely for longer questionnaires (Vehovar, et al., 2002). Irrespective of the overall length of the questionnaire, completion times are commonly increased with multiple-page designs as compared to one-page designs (Lozar Manfreda, et al., 2002; Vehovar, et al., 2002). Furthermore, at question level, when considering a set of rating scale items, spreading the items over multiple pages rather than presenting them on a single Web page increases the overall time needed to answer the whole set of items (Couper, et al., 2001; Tourangeau, et al., 2004).

Concerning a respondent's indirect evaluation of the two designs, the one-page design tends to be evaluated more positively. Asking for the respondents' intention to take part again in another survey, the proportion of respondents endorsing 'very likely' was significantly higher in the one-page design. Respondents indicating 'unlikely' cited more often survey length and duration as a reason for refusing another survey when the current survey was asked in a multiple-page rather than a one-page design (Peytchev, et al., 2006). Obviously, the multiple-page design is perceived as more time-consuming and more burdensome than the one-page design. This conclusion was supported by Lozar Manfreda and colleagues (2002) who used the respondents' comments on a survey and the problems associated with it as an indirect measure of respondent satisfaction with the multiple-page design. Respondents not only criticized the long page loading times and a lack of orientation within the questionnaire, but they explicitly requested more questions on one Web page. These findings suggest that not implementing a strict multiple-page design with each item on a single page is preferable. Presumably, the ideal questionnaire structure is somewhere in between a strict one-page and multiple-page design.

3.2.2 *Actual and Announced Survey Length*

Previous studies showed that the actual survey length affects the prevalence of item missing data, with more item nonresponse and survey breakoff occurring with the increasing length of a questionnaire (Deutskens, et al., 2004; Galesic & Bosnjak, 2009; Ganassali, 2008; Heerwegh & Loosveldt, 2006; Herzog & Bachman, 1981; Yan, et al., 2011). For example, Galesic (2006) varied the overall questionnaire length (10, 20, and 30 minutes) and found that the risk of survey breakoff significantly increased with the announced survey length, which in turn was identical with the actual survey length. Compared to respondents completing a 10-minute questionnaire, respondents had a 20% and 40% higher risk of survey breakoff when receiving a 20-minute and 30-minute questionnaire, respectively (Galesic, 2006). In addition, Yan and colleagues (2011) found that the actual survey length affected the risk of item nonresponse. Based on their finding of significantly higher item nonresponse rates in short questionnaires compared to long questionnaires, they suggested that “survivors of a long questionnaire may be more conscientious than those who complete a shorter questionnaire because the longer task provided more opportunity for less conscientious respondents to quit” (Yan, et al., 2011, p. 144). Thus, the authors promoted greater sensibility towards a possible trade-off between lower breakoff rates but higher item nonresponse in shorter questionnaires compared to longer questionnaires (Yan, et al., 2011). In relation to impaired measurement properties, Deutskens and colleagues (2004) found significantly more ‘don’t know’ answers in a long questionnaire of about 30 to 45 minutes compared to a short questionnaire of approximately 15 to 30 minutes. As demonstrated by an older study of Herzog and Bachman (1981), the risk of nondifferentiation in a rating scale is a function of both the position of the scale within the questionnaire and the overall length of the questionnaire. In extremely long questionnaires of about 2 hours, rating scales presented later showed a higher extent of nondifferentiation than those appearing earlier. Interestingly, irrespective of the question position within a long questionnaire, rating scale items addressing issues of high personal relevance were hardly affected by nondifferentiation. In shorter questionnaires of about 45 minutes, however, the extent of nondifferentiation did not vary over the course of the questionnaire (Herzog & Bachman, 1981).

As previous studies showed, a longer announced survey length decreases the likelihood that respondents start the survey (Crawford, et al., 2001; Deutskens, et al., 2004; Peytchev, 2011; Trouteaud, 2004; Yan, et al., 2011). This is explained by varying levels of respondent commitment depending on the alleged survey length. A respondent's consent to a 5-minute questionnaire requires a lower level of commitment than a consent to a 20-minute questionnaire, with the latter representing a major investment in terms of respondent effort (Yan, et al., 2011). In addition, Peytchev (2011) revealed that a shorter announced survey length was associated with a higher proportion of respondents who started the survey but also with more respondents who broke off, compared to a longer announced survey length. In accordance with these findings, Crawford and colleagues (2001) previously showed in a survey with an actual average length of 20 minutes that respondents who were informed that the survey would take 8 to 10 minutes were significantly more likely to start the questionnaire than respondents who were announced a survey length of 20 minutes, however, at the expense of a higher breakoff rate in the former condition. Therefore, underestimating the actual survey length encourages more respondents to start the questionnaire because of a lower perceived initial respondent burden, but also evokes a higher risk of survey breakoff once respondents realize the actual survey length (Crawford, et al., 2001). Consequently, a respondent's initial decision in favor of starting the survey is most likely to be affected by the announced survey length, whereas actual completion rates are more likely to be affected by the actual survey length and the discrepancy between the announced and actual survey length (Peytchev, 2011). These results are in accordance with Yan and colleagues (2011) who argued that the level of respondent commitment may differ depending on the announced survey length where the commitment of respondents consenting to a short questionnaire of up to 10 minutes would not be sufficient for completing an actual 20-minute questionnaire (Yan, et al., 2011).

Thus, besides the actual and announced survey length by themselves, the discrepancy between both dates has a determining influence on the overall completion rate. Indications of a respondent's process within survey completion by means of progress indicators can facilitate a comparison between the announced and actual survey length and help prevent premature termination.

3.2.3 *Progress Indicator*

Unlike scrollable or static Web designs, interactive Web surveys with survey questions being presented page-by-page hardly provide respondent orientation concerning the actual length of the questionnaire and a respondent's progress within the questionnaire. High levels of premature termination of the survey may be the result of the respondent's perceived burden, increasing over the course of the survey completion which, by implication, results in decreased respondent motivation to continue with the survey (Dillman & Bowker, 2001; Heerwegh & Loosveldt, 2006). Thus, the basic idea of presenting progress indicators in interactive Web surveys is that they are expected to prevent premature termination of the survey because respondents are constantly kept informed about the current completion status and are able to trace the progress with each survey question answered (Dillman, 2000, p. 398; Heerwegh & Loosveldt, 2006; Yan, et al., 2011).

However, mixed results suggest that the mere presence of a progress indicator does not necessarily reduce survey breakoff rates. Findings of Yan and colleagues (2011) revealed a more differentiated view on the effects of progress indicators. They pointed out that a positive effect of progress indicators depends on a respondent's expectations regarding the overall length of the questionnaire and the extent to which these expectations are met. In accordance with their results, Yan and colleagues (2011) recommended against providing a progress indicator in actually long questionnaires, or in questionnaires that are actually longer than expected as a result of the misrepresentation of the survey length in the invitation. This conclusion is in line with the findings of Crawford and colleagues (2001) revealing that a progress indicator has a negative effect on the completion rate when the feedback on the proportional number of survey questions that have already been processed underestimates the actual progress as a consequence of a series of more burdensome open-ended questions at the beginning of the questionnaire. In this case, respondents may overestimate the remaining time and effort required to complete the survey. Accordingly, Conrad and colleagues (2003) applied experimentally manipulated progress indicators, pretending different speeds of progress and showed that fast-to-slow progress indicators indicating faster progress at the beginning of the questionnaire than respondents would have assumed were most effective in reducing survey breakoff rates, while conventional progress indicators with constant speed

display had adverse effects on completion rates in long questionnaires. Other studies showed a positive effect of progress indicators on an increase in completion rates, however, differences did not reach statistical significance (Couper, et al., 2001; Heerwegh & Loosveldt, 2006). Concerning the occurrence of item nonresponse, previous research found that the implementation of progress indicators tended to decrease item nonresponse rates as compared to questionnaires without a progress indicator. However, differences between the experimental conditions were either small (Heerwegh & Loosveldt, 2006) or failed to reach statistical significance (Conrad, et al., 2003; Couper, et al., 2001).

Thus, the added value of progress indicators seems to be rather limited relating to a reduction in item missing data, whereas in particularly long questionnaires, they even have counter-productive effects. Again, a high consistency in the actual and announced survey length seems to be most important in order to keep item nonresponse and survey breakoff at a low level. Also, the overall length of questionnaires should be kept as short as possible.

3.2.4 Real-Time Validation

Another difference between one-page and multiple-page designs is that the latter allow for the implementation of dynamic real-time validation of respondent input. Real-time validation helps reduce routing error and completeness error, which in turn has a positive impact on the occurrence of item nonresponse but potentially differing effects on the incidence of survey breakoff and measurement properties (Peytchev & Crawford, 2005). Routing errors are reduced and navigation within the questionnaire can further be facilitated by implementing automated skip logics, initiated as a function of prior inputs. As a result, survey researchers do not need to exclusively rely on the respondent's ability and motivation to correctly follow skip instructions or hyperlinks. Thereby, item missing data due to unintentional skips of survey questions or an unnecessary increase in respondent burden due to accidentally answering questions not applicable to the individual respondent can reliably be prevented in multiple-page designs (Dillman, et al., 1998; Peytchev, et al., 2006; Peytchev & Crawford, 2005). This is directly related to completeness errors that can be reduced with the aid of required-response validations being executed each time a respondent clicks on the 'Continue' button. In order to

reduce intentional or unintentional item nonresponse, automated prompts can be implemented either in terms of forced prompts inevitably requiring a respondent input before allowing to go on to the next Web page, or in terms of soft prompts enabling the respondent to proceed even though no input is made (Blumenstiel & Roßmann, 2013; Derouvray & Couper, 2002; Lozar Manfreda, et al., 2002; Peytchev, et al., 2006). Although required-response validation is successful in reducing item nonresponse, its use is controversially discussed because of potentially negative effects on the respondent's motivation to continue with the questionnaire. In some cases, respondents do not know an answer or have good reasons to purposely not provide an answer. In such cases, forcing respondents to provide an answer may cause frustration and annoyance which at worst results in premature termination of the survey (Blumenstiel & Roßmann, 2013; Derouvray & Couper, 2002; Dillman & Bowker, 2001; Dillman, et al., 2009, p. 209; Peytchev & Crawford, 2005). Forced prompts involve a particular risk of survey breakoff because of the increased respondent burden and may increase the risk of 'over-editing' since the respondent possibly provides an inappropriate response just to continue with the survey questions on the next Web page (Derouvray & Couper, 2002; Peytchev & Crawford, 2005). Hence, the benefits of required-response validation concerning reduced item nonresponse need to be traded off against the costs of increased survey breakoff and impaired measurement properties. Similarly, the likelihood of accurate responses may be affected by automated consistency checks, asking respondents to adjust their responses in case of any contradictions. As a result, although potential consistency errors can be reduced by actively countering inconsistent responses, the risk of over-editing remains because of the strong assumption that a survey researcher can judge whether responses are inconsistent or not. Furthermore, respondents may become aware of the importance of answering consistently and therefore, adapt their response behavior in accordance with the consistency checks (Peytchev & Crawford, 2005). Thus, although interactive Web surveys enable the use of such dynamic real-time validation, these various kinds of active intervention in the question-answer process are still used reservedly in the practice of questionnaire administration because their negative effects on data accuracy still seem to outweigh their potential beneficial effects.

3.2.5 *Question Context*

At the macro level of the questionnaire, the position of a survey question within the questionnaire may affect a respondent's perceived burden, which in turn is likely to have differing effects on data accuracy (Galesic, 2006; Herzog & Bachman, 1981; Krosnick & Presser, 2010; Peytchev, 2009). In addition to an immediate or delayed effect of respondent burden on the occurrence of incomplete and inaccurate responses, a 'cumulative' effect of respondent burden is composed of the total number of survey questions already answered, including the extent of burden perceived with the specific characteristics of the current question as well as the burden experienced during the processing of previous questions (Galesic, 2006; Peytchev, 2009). Thus, the risk of item missing data and impaired measurement properties is likely to vary depending on the position of the respective question and consequently, the number of questions already answered (Galesic, 2006; Galesic & Bosnjak, 2009; Herzog & Bachman, 1981; Krosnick & Alwin, 1988; Peytchev, 2009). However, a systematic examination of the effect of the item position on the extent of survey breakoff, item nonresponse, and systematic response tendencies in rating scales is scarce because the question order in Web surveys has rarely been experimentally varied in previous studies. One exception is the study of Galesic (2006) and colleagues (Galesic & Bosnjak, 2009) where a Web questionnaire was divided into 20 question blocks, with the arrangement of these blocks being systematically varied. Results showed that survey breakoff was neither at a constant level, nor increased or decreased monotonically over the course of survey completion. Nonetheless, survey questions appearing later in the questionnaire were associated with an increased perceived burden and higher breakoff rates (Galesic, 2006). By contrast, Galesic and Bosnjak (2009) found no significant increase in item nonresponse over the course of survey completion. Concerning the prevalence of impaired measurement properties, they found that the time spent on answering survey questions decreased over the course of the survey completion. Also, the length of responses to open-ended questions was shorter in later parts of the questionnaire and responses to rating scales were less differentiated, when answering them in the last third compared to the first third of the questionnaire (Galesic & Bosnjak, 2009). Based on their findings, Galesic and Bosnjak (2009) concluded that "as fatigue and boredom accumulate throughout the survey, the respondents may

be less and less willing to invest the effort needed for good quality answers” (p. 358).

At the micro level of individual survey questions, the order of survey questions within a questionnaire may affect the respondents’ cognitive question-answer processing because preceding survey questions provide a context in which the comprehension of the current question, the retrieval of information, the formation of a judgment, and the communication of the response take place. As a result, the context in which survey questions are processed has a determining influence on task difficulty, with the question context either facilitating or complicating the execution of the four steps of the question-answer process (Krosnick, 1991; Krosnick & Presser, 2010; Strack, 1992; Tourangeau & Rasinski, 1988). Variations in task difficulty may in turn have a decreasing or increasing effect on the prevalence of the item missing data and the impairment of measurement properties because of their interaction with the respondent’s ability and motivation to provide complete and accurate responses (Krosnick, 1991; Krosnick & Presser, 2010). The implications of question context for the processing of rating scales are discussed in detail in section 4.3.1.

3.2.6 Socio-Environmental Context

Further factors influencing the level of task difficulty are related to the socio-environmental context of survey responding. The presence of distracting events in the respective survey situation can increase task difficulty, which in turn interferes with a respondent’s ability and motivation to carry out the question-answer process thoroughly (Krosnick, 1991). Thus, interruptions and external interference by, for instance, the presence of other persons or an incoming phone call make it more difficult for respondents to provide complete and accurate responses (De Leeuw, 2005, p. 244; G. E. Kennedy, 2004; Schwarz, Strack, & Mai, 1991).

The occurrence of distracting events may be even more pronounced in Web surveys. Web survey respondents are highly likely to be involved in multitasking which may distract respondents from their actual response task (De Leeuw, 2005; Heerwegh & Loosveldt, 2008; G. E. Kennedy, 2004). “Multitasking and quickly skipping from one topic to the next [...] may lead to more superficial cognitive processing, more top of the head answers, and more satisficing in responding to survey questions” (De Leeuw, 2005, p. 244).

This is aggravated by the fact that any distractions occurring in the survey-taking environment of self-administered surveys are completely outside the control of the survey researcher (Schwarz, Strack, Hippler, et al., 1991). However, as opposed to question context, distracting factors of the socio-environmental context may induce random measurement error rather than systematic measurement error (Viswanathan, 2005, p. 140).

The above considerations have clearly shown that the characteristics of both questionnaire design and questionnaire administration decisively influence the level of task difficulty and hence, a respondent's ability and motivation to provide complete and accurate responses to survey questions in general. It has also been pointed out that rating scales are one of the most commonly used question formats in attitude and personality measurement although it is known that rating scales suffer from the respondents' susceptibility to systematic response tendencies and the risk of item missing data. Specific characteristics of the design and administration of rating scales may help promote the respondents' attentiveness and carefulness in processing a set of several rating scale items. For this reason, the key aspects of rating scale design and rating scale administration are expanded in the next sections.

4. DESIGN AND ADMINISTRATION OF RATING SCALES IN WEB SURVEYS

Likert-type scales are the most prominent example of rating scales and presumably the most commonly used type of rating scales for measuring attitudes and self-descriptions in survey research (Brill, 2008). Likert-type scales denote multi-item measures using an identical rating scale format. Likert-type items are traditionally answered on a 5-point fully labeled agree-disagree scale with the bipolar scale ends reflecting opposing alternatives and a clear conceptual midpoint in between (Alwin, 2007, p. 186; Brill, 2008). Thus, when using a Likert-type scale format, several key decisions concerning form-related verbal characteristics of a rating scale are already specified, i.e., design specifications concerning the number and labeling of response options, the polarity of the response options, and the provision of an explicit midpoint.

The basic design principles concerning form-related verbal characteristics of rating scales and their implications for the occurrence of item missing data and the impairments of measurement properties are discussed in section 4.1. As related to the form-related visual characteristics of rating scales discussed in section 4.2, survey researchers have to decide whether rating scale items are presented as stand-alone items, in a series, or in a battery. In practice, rating scale items are commonly arranged in batteries, so-called grid (or matrix) questions. Besides benefits from using grid questions displaying multiple items neatly arranged on a single screen and allowing for a fast and efficient scale completion, grid formats are at higher risk of evoking cognitive shortcuts within the question-answer process as respondents may rush through the items and try to minimize their effort necessary to answer the items. Grid formats and potential causes for their susceptibility to cognitive shortcuts are further discussed in section 4.2.2. The decision on using a single-item-per-screen or multiple-item-per-screen format decisively influences the context-related characteristics of rating scales. How rating scales are organized, i.e., the order of rating scale items and the arrangement of response options as key aspects of rating scale administration can have differing implications for a respondent's cognitive question-answer processing and the accuracy of survey data in terms of item-order and response-order effects further discussed in section 4.3.

4.1 Form-Related Verbal Characteristics of Rating Scales

4.1.1 Number of Items in a Measure

The distinction between single-item and multi-item measures refers to the number of rating scale items used in an attitude or personality measure. Typically, rating scale questions are composed of several rating scale items that are intended to measure the same underlying construct. Multi-item measures have two major advantages. First, they are expected to evoke higher scale validity since a latent construct can be captured in all its facets. Second, multi-item measures produce higher scale reliability due to repeated measurements and averaging out of random errors (Andrews, 1984; Diamantopoulos, Sarstedt, Fuchs, Wilczynski, & Kaiser, 2012; Schriesheim & Hill, 1981). Although multi-item measures are commonly preferred, they also carry certain risks. Correlations between items of the same measure are likely to be spuriously inflated because of “the use of the same response format, similar stems, or the completion of items in close proximity” (Viswanathan, 2005, p. 110). And, instead of capturing all facets of a construct, there is rather the risk that the items of multi-item measures represent mere semantic redundancies (Diamantopoulos, et al., 2012; Drolet & Morrison, 2001). Such item redundancies are likely to be recognized by respondents, which in turn may increase the risk of diminishing respondent motivation to attentively consider specific item meanings (Drolet & Morrison, 2001). In any case, “an increase in the number of the scale items can lead to participant fatigue, boredom, and inattention, which, in turn, can lead to inappropriate response behavior” (Drolet & Morrison, 2001, p. 198). Nevertheless, a preference for multi-item measures still deems appropriate, in particular when dealing with abstract, multifaceted constructs (Bergkvist & Rossiter, 2007; Sarstedt & Wilczynski, 2009).

4.1.2 Polarity of Items

In attitude and personality measurement, multi-item measures are often composed of items that are all worded in the same direction to measure the underlying construct which implies a uniform general direction of the wording of rating scale items. However, the polarity of rating scale items can also be varied by incorporating items worded in the opposite direction. To ensure clarity about the use of terms, reversed items are items whose meaning

is opposite to the general direction of the rating scale, i.e., contrary to the original items of a rating scale. Counterbalanced scales are rating scales which contain both original items that are worded in the general direction of the rating scale and reversed items that are worded in the opposite direction. Balanced scales have the same number of original and reversed items (Weijters & Baumgartner, 2012).

Although (counter)balanced rating scales including reverse worded items have several advantages, their use is still controversially discussed (Barnette, 2000; Weijters & Baumgartner, 2012; Weijters, et al., 2009). One advantage of (counter)balanced rating scales is that scale validity is improved because of a more comprehensive coverage of the underlying construct (Weijters, et al., 2009). Furthermore, the validity or accuracy of rating scale measures can be increased in two ways: While data gathering is ongoing, careless or nonattending respondents are urged to place greater emphasis on individual item characteristics by alternating original and reversed items, whereas once data gathering has been completed, measurement error due to systematic response tendencies such as acquiescent responding, careless responding, or nondifferentiation can be assessed by the survey researcher (Barnette, 2000; Paulhus, 1991; Podsakoff, et al., 2003; Weijters & Baumgartner, 2012; Weijters, et al., 2009). Thus, (counter)balanced rating scales can be applied to draw the respondents' attention to the item content already during rating scale completion as well as to identify systematic response tendencies in retrospect. However, one shortcoming of the inclusion of reversed items in (counter)balanced rating scales is the reduced internal consistency of rating scale measures due to lower correlations between the original and the reverse worded items. A so called reversed-item bias due to inconsistent responses or misresponses to pairs of original and reversed items finds its expression in a smaller Cronbach's alpha, lower factor loadings, or even distorted factor structures (see section 5.2.1) (Barnette, 2000; Harrison & McLaughlin, 1993; Hinkin, 1995; Weijters, et al., 2009).

4.1.3 Number of Response Options

After the respondents have made their mental judgment, they are required to map their judgment onto the response options provided. Aside from the inherent complexity of the underlying construct being intended to measure, the ease of mapping a judgment onto a rating scale decisively depends on the

number of response options provided and the resultant coverage of subtle gradations in the respondents' judgments allowing the respondents to express their opinion in a sufficiently differentiated and stable manner over time (Krosnick & Fabrigar, 1997; Krosnick, et al., 2005). Rating scales in attitude and personality measurement should basically allow respondents to report neutral, moderate, and extreme responses, which is typically true in the case of a 5-point rating scale. The additional benefit of providing more than five response options depends on the complexity of a respondent's judgment (Krosnick, et al., 2005; Viswanathan, 2005, pp. 244-245).

Theoretically, a higher number of response options may enable greater differentiation in the respondent's judgment, resulting in more valid or accurate responses provided that, first, a respondent's mental representation of the theoretical construct is sufficiently detailed, and second, a respondent actually makes use of the full scale range (Krosnick, et al., 2005). Failing these premises, respondents are likely to ignore large parts of the scale when they are confronted with a larger number of response options because the scale is too detailed by contrast to a less detailed mental representation of the theoretical construct, or simply because respondents are unwilling to make more fine-grained distinctions. As a consequence, no additional information is obtained by means of a higher number of response options when respondents reduce their effort and restrict their answers to a small range of the rating scale fostering less differentiation among the rating scale items (Krosnick, et al., 2005; Viswanathan, 2005, pp. 244-245). Furthermore, an inadequate detailedness of rating scales may result in a respondent's confusion about the meaning of each response option. Considering that any additional response option needs to be interpreted by a respondent, while each extra need for interpretation will increase the risk of inconsistencies in both meaning and use over time and across respondents, the risk of random responding increases in unreasonably long response scales. This in turn results in less reliable and thus, less precise responses to a rating scale (Alwin, 1989; Friedman & Amoo, 1999; Krosnick & Fabrigar, 1997; Krosnick, et al., 2005). In view of an increased task difficulty and the accompanying effort required to interpret the meaning of additional response options and to map a judgment, cognitive shortcuts in the question-answer process may be fostered, resulting in differing kinds of systematic response tendencies (Krosnick, et al., 2005).

Thus, using more response options than a respondent is able or willing to manage, is likely to come along with a decrease in the precision and

accuracy of survey estimates. Therefore, a tradeoff between sufficient distinction and adequate clarity of the meaning of response options has to be considered when reaching a decision concerning the optimal number of response options in rating scales. As a rule of thumb, the use of moderately long rating scales with 5 or 7 response options is recommended when using bipolar scales for attitude and personality measurement to cover the full range of slight, moderate, and substantial judgments, while keeping the cognitive effort within a limit in order to ensure a high level of reliability and validity of rating scale measures. Analogously, 4 to 7 response options are proposed for unipolar scales (Krosnick & Fabrigar, 1997; Krosnick, et al., 2005; Preston & Colman, 2000; Weijters, et al., 2010).

4.1.4 Labeling of Response Options

In general, the numeric and verbal labels assigned to each response option may help respondents interpret the response scales and clarify the meaning of a survey question (Krosnick, et al., 2005; Schwarz, Knäuper, et al., 1991). From a practical point of view, the numeric labeling of response options in Web surveys is deemed a relic from self-administered paper-based questionnaires where it was originally implemented for simplification of manual data input. In view of this, numeric labels have become superfluous in Web surveys (Callegaro, Wells, & Kruse, 2008). From a theoretical perspective, verbal labeling is deemed more advantageous than numeric labeling because respondents can directly deduce the meaning of the verbally labeled response options without the additional need for interpreting each numeric label at first. By contrast, when solely presenting numeric labels, the respondents have to generate a verbal definition for each response option first, before they can match these definitions against their mental representation of the issue in question. Thus, uniformity in terms of the meaning of response options may be increased by verbal labeling, whereas the respondent burden perceived during the cognitive question-answer processing is limited (Krosnick & Fabrigar, 1997; Krosnick, et al., 2005). In view of this theoretical assumption, Callegaro and colleagues (2008) showed that adding numbers to verbally labeled bipolar rating scales did not change the response distributions of rating scale items. Hence, the respondents' answers appear more likely to be affected by verbal labels than by numeric labels. Similarly, Toepoel and Dillman (2008) showed that adding numbers to fully-verbally

labeled rating scales did not reveal significant differences, neither in terms of response distributions nor concerning inter-item correlations. In line with the general assumption of a hierarchy of verbal and visual questionnaire features with verbal features taking precedence over numerical features, which in turn exceeds the effect of purely visual features as proposed by Tourangeau and colleagues (2007a), Toepoel and Dillman (2008) suggested that—irrespective of whether or not adding numbers—fully-verbally labeled scales should be generally preferred in order to reduce differential effects of the visual features of rating scales on the interpretation of response scales and the meaning of a survey question.

Thus, a further decision has to be made concerning the verbal labeling of all response options or simply labeling the endpoints of a rating scale, as this choice may have differing implications for data accuracy: While the full labeling of rating scales may help respondents clarify the meaning of the response options and provide more precise responses, it may also increase the cognitive effort required to read and process all response option labels before mapping a judgment onto them. Therefore, cognitive shortcuts within the question-answer process may be promoted, resulting in less accurate survey responses (Christian, Dillman, & Smyth, 2008; Krosnick & Fabrigar, 1997). Whereas it is incontestable that verbal labeling of the endpoints of a rating scale is indispensable to have a meaningful rating scale, previous studies provided mixed findings concerning the comparison of fully labeled and partially labeled rating scales and their advantages in terms of increased reliability and validity of rating scale measures (Krosnick & Fabrigar, 1997). Altogether, however, results suggest fully labeling of response options in rating scales because verbal labels actually help respondents clarify the meaning of the response options and translating their judgment into the response options which results in higher data accuracy (Alwin, 2010; Alwin & Krosnick, 1991; Christian, et al., 2008; Krosnick & Berent, 1993; Krosnick & Fabrigar, 1997; Weijters, et al., 2010).

4.1.5 Agree-Disagree versus Construct-Specific Response Options

One frequently mentioned advantage of using agree-disagree response options is that “agree-disagree items are simple to construct and easy to answer” (Schaeffer & Presser, 2003, p. 80). Furthermore, agree-disagree response options can be universally deployed for measuring almost any theoretical

construct without changing the response options (Krosnick & Presser, 2010; Saris, Revilla, Krosnick, & Shaeffer, 2010). However, it should be taken into account that rating scale items with agree-disagree response options involve more cognitive effort because, as related to the four stages of the question-answer process, it is more difficult for respondents to come to a thorough understanding of the response options and an appropriate mapping of their judgment onto it. Respondents first have to identify the evaluation dimension in the question stem, form a judgment on the basis of this evaluation dimension, and then translate this construct-specific judgment into the more abstract, construct-unspecific, agree-disagree response options. Thus, mapping a judgment onto the response options provided requires more cognitive effort in a rating scale using agree-disagree response options as compared to construct-specific response options (Saris, et al., 2010). This increased respondent burden accompanied by the use of agree-disagree response options may foster systematic responding such as a respondent's tendency to acquiesce and more extreme responses than when applying construct-specific response options (Saris, et al., 2010; Schaeffer & Presser, 2003). Although it is often advised against the use of agree-disagree response options for answering rating scale items, they are frequently applied in attitude and personality measurement mostly for pragmatic reasons (Krosnick & Presser, 2010; Saris, et al., 2010)

4.1.6 Polarity of Response Options

A rating scale can be unipolar or bipolar. While in a unipolar rating scale, the two endpoints of the scale differ in intensity but not in valence, both endpoints are of the same intensity but opposite valence in a bipolar rating scale. With proportional gradations in-between, bipolar rating scales measure both the valence of a respondent's judgment indicated by the direction to one end of the scale, and the intensity indicated by the distance to the neutral middle of the scale (Alwin, 2007, p. 195; C. Kennedy, 2008; Peabody, 1962). When measuring bipolar constructs like attitudes and self-descriptions where "more than one concept is presented and the underlying continuum is abstract" (Alwin, 2007, p. 195), survey researchers are advised to use bipolar scales (Lubian, 2010). One shortcoming of bipolar rating scales is, however, that respondents are more reluctant to choose the most negative response option because its meaning is considered more extreme in a bipolar rating

scale ('completely disagree') compared to a unipolar rating scale ('not agree at all') (Krebs & Hoffmeyer-Zlotnik, 2009). Notwithstanding this limitation, in practice, survey questions asking for nonfactual content such as attitudes and self-descriptions are commonly measured on bipolar scales, whereas unipolar scales are usually used for behavioral questions (Alwin, 2007, p. 195; 2010; Krosnick & Fabrigar, 1997).

4.1.7 Middle Response Option

The decision on whether a neutral midpoint should be provided in a rating scale or not is often discussed with regard to an increased risk of different kinds of systematic response tendencies arising from a respondent's susceptibility to cognitive shortcuts in survey responding. While an uneven number of response options including a midpoint may encourage respondents to simply select the midpoint when they are less motivated, less interested in the topic, or have less pronounced opinions towards the issue in question, an even number of response options without a midpoint may encourage such respondents to randomly select a response option on either side of the rating scale. Thus, in the former case, respondents are more prone to midpoint responding as an easy way out of the cognitive question-answer processing, whereas in the latter case, a kind of forced choice for or against the issue in question is provoked despite a respondent's actual ambivalence which is likely to foster random responding in terms of mental coin flipping (Bishop, 1990; Kalton & Schuman, 1982; Krosnick, 1991; Krosnick & Fabrigar, 1997; Krosnick, et al., 2005; O'Muirheartaigh, Krosnick, & Helic, 2000). However, O'Muirheartaigh and colleagues (2000) found that providing an explicit midpoint actually prevented respondents from randomly selecting one of the moderate response options near the middle of the rating scale which led to an improvement in the reliability of rating scale measures, while their validity remained unaffected. Based on these findings, offering an explicit midpoint in rating scales is recommended. Similarly, based on their findings of reduced extreme responding and reduced reversed-item biases due to more consistent responses to pairs of original and reversed items when adding a neutral midpoint to rating scales, Weijters and colleagues (2010) suggested "avoiding scales without a midpoint, unless particular, relevant reasons present themselves" (p. 245).

4.1.8 *'Don't Know' Response Option*

One of the questions which arises with rating scale construction and is strongly related to the offering of a scale midpoint, is whether an explicit 'don't know' option should be provided or not. In interviewer-administered surveys, there is no need to provide an explicit 'don't know' option because of an interviewer's ad hoc intervention to record a 'don't know' response if necessary. In self-administered surveys, however, a survey researcher has to decide in advance whether a 'don't know' option should be included in a survey question, and in this case, in a rating scale or not (Beatty & Herrmann, 2002; Derouvray & Couper, 2002; Kalton & Schuman, 1982). Respondents may regard an explicit 'don't know' option as a cue about the acceptability of item nonresponse (Beatty & Herrmann, 2002; Krosnick & Fabrigar, 1997). As a consequence, offering an explicit 'don't know' option in self-administered surveys may considerably enhance the proportion of nonsubstantive responses and shift the response distributions of substantive answers. For that reason, the decision for or against providing a 'don't know' option has to be weighed carefully in advance (Beatty & Herrmann, 2002; Christian, et al., 2009; De Leeuw, et al., 2003; Krosnick & Fabrigar, 1997; Krosnick, et al., 1996; Tourangeau, et al., 2004). Additionally, it is well known that 'don't know' responses "often result not from genuine lack of attitudes but rather from ambivalence, question ambiguity, satisficing, intimidation, and self-protection" (Krosnick, et al., 2005, p. 49) which implies that even though respondents would be able to provide a substantive answer, they refuse it for various reasons. Thus, in accordance with the satisficing theory, respondents may select the 'don't know' option to abbreviate the question-answer process and to complete the response task quicker without any major cognitive effort, or because they want to avoid disclosure of sensitive or too personal information (Krosnick, 1991; Shoemaker, et al., 2002). In short, respondents select a 'don't know' option because of a lack of motivation rather than a lack of opinion. On the downside, when no explicit 'don't know' option is provided, respondents who actually have no opinion on the issue in question may be more prone to some systematic response tendencies affecting data accuracy negatively. For example, in the absence of a pronounced opinion on the issue in question, respondents are likely to express a moderate opinion by simply selecting the midpoint of a rating scale (Friedman & Amoo, 1999).

A practical advice concerning the decision whether to provide an explicit ‘don’t know’ option or not, is dictated if there is good reason to believe that respondents truly have no opinion on the issue in question. In this case, the survey researcher should decide in favor of providing a ‘don’t know’ option to ensure high data accuracy (Friedman & Amoo, 1999). Conversely, the decision against providing an explicit ‘don’t know’ option is reasonable if there are good reasons to assume that almost every respondent has an opinion on the issue in question. In nonfactual questions dealing with attitudes and self-descriptions relevant to each of the respondents to a greater or lesser extent, a true lack of opinion is not expected. This is why a ‘don’t know’ option is not needed. Thus, the direction and strength of attitudes and self-descriptions can be measured with sufficient precision and accuracy by applying Likert-type rating scales with 5 or 7 response options and an explicit midpoint allowing respondents for sufficient differentiation and expression of a moderate or neutral position towards the issue in question.

4.2 Form-Related Visual Characteristics of Rating Scales

In attitude and personality measurement, rating scale measures are mostly composed of several rating scale items (see also section 4.1.1). Thus, in interactive Web surveys where individual questions are commonly presented page-by-page, survey researchers have to decide whether they present a set of related rating scale items in a single-item-per-screen or in a multiple-item-per-screen format. In the former case, each Web page includes a stand-alone item, whereas in the latter case, multiple items are arranged on a single Web page either in the form of a series where response option labels are repeated for each item, or in the form of a battery where response option labels are presented only once. Thus, single-item-per-screen and multiple-item-per-screen formats mainly differ in terms of the number of items grouped together on a single Web page, which in turn can differently affect the risk of item missing data and the impaired measurement properties of rating scales.

4.2.1 Single- versus Multiple-Item-per-Screen Formats

In a single-item-per-screen format, rating scale items are presented one by one which may allow respondents to focus more on each item and its respective response options as if all items are presented together in a multiple-item-per-

screen format. On the other hand, it is often argued that single-item-per-screen formats may lead to a loss of a sense of context. This context, however, is deemed necessary to get a complete understanding of the set of rating scale items as a whole and to be in a continuous flow of cognitive processing (Callegaro, Shand-Lubbers, & Dennis, 2009; Christian & Dillman, 2004; Couper, et al., 2001; Dillman, et al., 2009, pp. 203-204; Dillman, et al., 1998; S. S. Fricker, et al., 2005; Krosnick, 1991; Schwarz, Strack, & Mai, 1991; Toepoel, et al., 2009b). This implies that a multiple-item-per-screen format may allow respondents to process a set of rating scale items in its entirety without any unnecessary interruptions (Couper, et al., 2001; Dillman & Bowker, 2001; Dillman, et al., 1998; Norman, et al., 2001; Toepoel, et al., 2009b). On the contrary, facing respondents with a large amount of information in a multiple-item-per-screen format may also carry an increased risk of respondents losing track and missing fine distinctions between single rating scale items (Callegaro, Shand-Lubbers, et al., 2009; Couper, et al., 2001; Dillman, et al., 2009, pp. 203-204; Gräf, 2002; Tourangeau, et al., 2004). So, both the single-item-per-screen format with its risk of a loss of context and the multiple-item-per-screen format with its potential information overload require a certain degree of cognitive effort and thus, pose a risk of increased respondent burden.

With regard to the extent of navigational effort, scrolling allows respondents a smooth navigation through a list of rating scale items presented in a multiple-item-per-screen format which actually reduces overall time needed to complete a rating scale (Couper, et al., 2001; Thorndike, et al., 2009; Toepoel, et al., 2009b; Tourangeau, et al., 2004). In general, however, excessive scrolling through a seemingly endless list of survey questions is likely to increase the perceived respondent burden because of the impression that a questionnaire is too long to be completed (Schonlau, et al., 2002; Toepoel, et al., 2009b). Furthermore, scrolling can also increase the actual respondent burden, particularly for those respondents who are not overly practiced in using a computer and the Internet (Norman, et al., 2001). Similarly, in single-item-per-screen formats, the actual respondent burden can be increased because of the excessive clicking from one Web page to the next (Dillman & Bowker, 2001; Toepoel, et al., 2009b). Hence, both the necessity for considerable scrolling in a multiple-item-per-screen format as well as a large number of mouse clicks required to navigate through a questionnaire in a single-item-per-screen format increase the extent of navigational effort and

the respondents' perceived and actual burden, at least for less computer literate respondents (Dillman & Bowker, 2001; Norman, et al., 2001; Toepoel, et al., 2009a). Attempts to reduce the extent of scrolling and the number of clicks when presenting rating scale items in either a multiple- or single-item-per-screen format seem to result in a dilemma because decreasing the need for scrolling inevitably results in an increase in clicking and vice versa (Peytchev, 2006).

With the objective of combining the advantages of a single-item-per-screen and multiple-item-per-screen format while, at the same time, keeping the extent of the cognitive and navigational effort within reasonable limits, a seemingly advantageous solution is the arrangement of rating scale items in a battery (Toepoel, et al., 2009b). A key argument for the use of so-called grid questions as a sort of multiple-item-per-screen format is that grid questions enable respondents to process a set of rating scale items in its entirety, with fewer intervening events, i.e., necessary question context is provided, whereas unnecessary interruptions (e.g., page breaks) are missing that are likely to distract respondents from their actual response task (Callegaro, Shand-Lubbers, et al., 2009; Christian & Dillman, 2004; Couper, et al., 2001; Dillman, et al., 1998; S. S. Fricker, et al., 2005; Krosnick, 1991; Peytchev, 2006; Schwarz, Strack, & Mai, 1991; Toepoel, et al., 2009b). Further implications of using a grid format on the completeness and accuracy of the respondents' answers to rating scales are discussed in the next section.

4.2.2 *A Special Case: Grid Formats*

When multiple rating scale items are related in content, these items are prevalently arranged in a grid (or matrix) format where all items are presented on one screen, with each row corresponding to an item and each column to a response option. Although being a special sort of multiple-item-per-screen formats, grid formats are assumed to have several advantages for both survey researchers and respondents as compared to a series as another multiple-item-per-screen format or single-item-per-screen formats.

Grouping of rating scale items and presenting them in a grid allows for providing the general introduction on how to complete the items, the general question request, and respective response options just once, whereby the actual question stimulus can be kept relatively short (Alwin, 2007, p. 204; Couper, et al., 2001). Hence, grid formats are space-saving by enabling

several items to be presented on one Web page which can primarily be considered an argument from a survey researcher's point of view. Presenting rating scale items in a grid format enables a reduction in the number of Web pages, avoids arbitrary page breaks, and allows for an efficient arrangement of items on a single screen without the need for excessive clicking or scrolling (Dillman, 2000, p. 100; Schonlau, et al., 2002; Toepoel, Das, & Van Soest, 2008; Toepoel, et al., 2009b).

From a respondent's point of view, item arrangement in grid formats has several implications for the navigational and cognitive effort respondents have to expend and therefore, for the extent of perceived and actual respondent burden that accompanies the processing of a set of rating scale items (Couper, et al., 2013; Dillman, et al., 2009, pp. 179-181; Toepoel, et al., 2009b). Consequently, besides the advantage of being space-saving, grid formats are also time-saving. The extent of time required to complete a number of related items is significantly reduced when a set of rating scale items is presented in a grid format as compared to presenting fewer items together or even each item separately on a single screen. For example, Toepoel and colleagues (2009b) showed a monotonically decreasing effect of the number of items per screen on the time required to complete the whole rating scale: The longest completion time was found in the single-item-per-screen condition (384 seconds), followed by the 4-items-per-screen and 10-items-per-screen condition (316 and 303 seconds, respectively), whereas the 40-items-per-screen condition (302 seconds) had the shortest completion time. These findings are in accordance with the results of Tourangeau and colleagues (2004) who showed that answering 8 items in a single grid took least time (60 seconds), followed by two grids on two single screens (65 seconds) and eight single screens (99 seconds). Couper and colleagues (2001) also found significantly shorter completion times for 11 rating scale items when they were presented altogether in a grid format (114 seconds) as compared to presenting them in a single-item-per-screen format (128 seconds). Other studies also found shorter completion times in grid formats as compared to single-item-per-screen formats but failed to prove statistically significant differences between the two formats (Bradlow & Fitzsimons, 2001; Callegaro, Shand-Lubbers, et al., 2009).

Such evidence of time savings is mostly deemed an indicator of allowing respondents to complete a rating scale easily and efficiently. In this regard, completion time is considered a proxy measure of the extent of task

difficulty and the level of respondent burden. As it has already been pointed out, presenting rating scale items in a grid format enables the survey researcher to provide the general introduction on how to complete the items, the general question request, and respective response options only once, whereby the respective question stimulus can be kept relatively short (Alwin, 2007, p. 204; Couper, et al., 2001). As a result, respondent effort necessary for screen orientation can be reduced. After the initial orientation on a Web page, respondents get used to the question-answer format rather quickly which obviates the need for recurrent orientation to the question request and respective response options with each additional item or with every new Web page (Couper, et al., 2001; Toepoel, et al., 2009b). Thus, the extent of cognitive effort can be reduced by rendering repeated processing of unchanged question components redundant and time savings can be yielded (Andrews, 1984; Couper, et al., 2001; Toepoel, et al., 2009b; Tourangeau, et al., 2004). Furthermore, the extent of navigational effort is reduced in terms of a decreased number of physical mouse movements and clicks required to answer a set of several rating scale items and to navigate to the next Web page, which in turn may also be reflected in time savings (Toepoel, et al., 2008, 2009b). Thus, the extent of cognitive and navigational effort can be reduced, while smooth navigation through a set of rating scale items is allowed.

Besides these reasons that are primarily aimed at streamlining the questionnaire and economizing respondent handling, there are further substantive reasons for grouping related rating scale items in a grid format. First, presenting a set of rating scale items on a single Web page implies that there are no interruptions within the cognitive question-answer processing repeatedly arising with each page break. This in turn allows respondents to fluently answer a battery of items without losing the common thread (Dillman & Bowker, 2001; Dillman, et al., 1998). Second, presenting a set of rating scale items that refer to the same or related issues closely together preserves the necessary question context. When related items are answered in close succession, relevant information has already been retrieved through previous items. This increased temporal accessibility of relevant information is likely to reduce the extent of cognitive effort required for the execution of the four steps of the question-answer process and the time needed for overall rating scale completion (see section 4.3.1) (Couper, et al., 2001; Knowles, 1988; Podsakoff, et al., 2003; Schwarz, Strack, & Mai, 1991; Tourangeau, 1992;

Tourangeau, et al., 1991). This facilitation of the response task coupled with a decrease in the actual respondent burden may be even more pronounced when sets of rating scale items are arranged in a grid format (Couper, et al., 2001; Grandmont, Graff, Goetzing, & Dorbecker, 2010; Tourangeau, et al., 2004). Hence, grid formats allow for rapid and continuous cognitive processing of a set of rating scale items, while respondent burden can be kept within a limit.

In sum, decreased response times needed to complete a set of rating scale items presented in a grid format are commonly attributed to efficiency gains with respect to the reduced navigational and cognitive effort (Couper, et al., 2001; Knowles, 1988; Schwarz, Strack, & Mai, 1991; Toepoel, et al., 2009b). However, a reduction in response times is not necessarily an indication of a reduced navigational and cognitive effort respondents have to expend and thus, no indication of a reduced actual and perceived respondent burden. An alternative explanation for shorter response times in grid questions is a respondent's tendency to fall back on cognitive shortcuts when answering a set of several rating scale items arranged in a grid format rather than attentively and carefully undergoing the four steps of the question-answer process for each item again and again (Andrews, 1984; Callegaro, Shand-Lubbers, et al., 2009; Knowles, et al., 1992; Stieger & Reips, 2010; Tourangeau, et al., 2004).

Two alternative reasons for a respondent's increased susceptibility to cognitive shortcuts in grid questions are presented here. On the one hand, grid formats carry an increased risk of low and probably too low cognitive load, which is why "respondents become fatigued or bored with repeatedly considering similar questions and simplify the strategy that they employ to formulate answers" (Knowles, et al., 1992, pp. 235-236). On the other hand, grid formats usually comprise a considerable amount of information presented simultaneously, and additionally, entail the necessity to combine information from both the rows and the columns. This in turn implies higher sensorimotor effort required for necessary hand-eye-coordination and an increased risk of high cognitive load (Couper, et al., 2013; Couper, et al., 2001; Dillman, et al., 2009, p. 179; Kaczmirek, 2010). Based on these two explanatory approaches, both a very low and a rather high level of cognitive load may actually increase respondent burden, which in turn is likely to have negative effects on the respondents' motivation and restrain them from providing complete and accurate responses. In this regard, the level of cognitive load is primarily attributable to the level of task difficulty in grid questions: Whereas the first

approach foregrounds a level of task difficulty that is too low, the second approach proceeds on the assumption of high task difficulty. Both affect the extent of respondent burden and the respondent motivation negatively. Thus, both alternative explanatory approaches argue within the framework of the satisficing theory and rely on the assumption that respondents need less time to complete rating scale items when they are presented in a grid format because respondents are more likely to rush through a battery of items. This is due to the higher respondent burden and thus, the lowered respondent motivation to engage in attentive and careful question-answer processing and to expend the effort required for complete and accurate responses (Callegaro, Yang, Bhola, & Dillman, 2005; Stieger & Reips, 2010; Tourangeau, et al., 2009).

Low Cognitive Load and Respondent Fatigue

Concerning the first approach, assuming that a low level of cognitive load is responsible for a respondent's susceptibility to cognitive shortcuts in grid questions, grid questions may carry an increased risk of respondent fatigue, and consequently, entail higher respondent burden than other question formats. In general, rating scales are likely to provoke the respondents' boredom and inattention towards the response task which is mainly due to the monotony of repetitive rating scale items referring to the same content and the same response options (Alwin & Krosnick, 1985; Gräf, 2002). As a consequence, "respondents may become fatigued or bored with repeatedly considering similar questions and simplify the strategy that they employ to formulate answers" (Knowles, et al., 1992, p. 235). More precisely, Gräf (2002) argued that when monotonously and uniformly asking a set of rating scale items, respondents get bored and increasingly inattentive and careless towards the response task because they already have an answer ready after they have answered the first few items, resulting in a disappearance of any cognitive stimulation. And, as Malhotra (2009) argued, "people are more motivated to persist in completing tasks when they are intricate, challenging, and enriching" (p. 180). Therefore, making the response task easier for respondents may compromise data accuracy because of a decline in respondent motivation to carefully process all relevant components of a rating scale and to make sufficiently precise distinctions between several rating scale items (Alwin & Krosnick, 1985; Callegaro, et al., 2005; Gräf, 2002; Hui & Triadis, 1985; Knowles, et al., 1992; Malhotra, 2009). Thus, a decrease in

response times for rating scale items arranged in a grid format can be the result of respondent boredom and respondent fatigue due to a low level of cognitive load falling below a critical threshold, with a certain level of cognitive load being necessary for the respondent's cognitive stimulation and involvement. The risk of respondent fatigue is likely to further increase with larger numbers of rating scale items presented in a grid format (Andrews, 1984; Drolet & Morrison, 2001).

Presenting rating scale items in a grid format may further increase respondent fatigue because of the mere visual structure of grid questions and the visual proximity of the items. According to the processing of visual questionnaire features, the heuristics 'like means close' and 'near means related' have impact on the interpretation and answering of rating scale items arranged in a grid format (see also section 2.2.3). Based on the heuristic 'like means close' which is attributable to the principle of similarity, visually similar elements will be seen as conceptually closely linked. Furthermore, according to the heuristic 'near means related' which is attributable to the principle of proximity, grouping items together on a Web page makes their similarity more salient whereby the likelihood is increased that respondents regard several items as conceptually more closely related. Thus, visual grouping of rating scale items is assumed to encourage respondents to consider the items as related entities (Tourangeau, et al., 2004). By implication, proximity effects (or grouping effects) occur when rating scale items are considered more related in content and thus, are rated more similar in grid questions simply because they are presented in close proximity rather than in greater distance (Tourangeau, et al., 2004; Viswanathan, 2005, p. 141; Weijters, et al., 2009). In interactive Web surveys where individual questions are predominantly presented page-by-page, item grouping in terms of presenting the rating scale items in a grid format while remaining survey questions are largely presented separately on several screens is expected to further increase the proximity effect (Couper, et al., 2001; Toepoel, et al., 2009b).

It is generally assumed that applying these grouping principles when designing a Web survey can help respondents with element organization, whereby question comprehension, information retrieval, and judgment is facilitated and overall question processing is fastened (Dillman, et al., 2009, p. 167). In principal, this effect is quite intentional. However, one needs to keep in mind that when items are presented together on the same screen,

compared to placing them on several screens, the respondents' cognitive question-answer processing and measurement properties of rating scale items may be affected. Among others, the use of grid formats is likely to increase the internal consistency of rating scale measures compared to single-item-per-screen formats simply because of the visual proximity of the rating scale items (Andrews, 1984; Callegaro, Shand-Lubbers, et al., 2009; Couper, et al., 2001; S. S. Fricker, et al., 2005; Jenkins & Dillman, 1997; Toepoel, et al., 2009b; Toepoel & Dillman, 2010; Tourangeau, 1992; Tourangeau, et al., 2004). So, instead of processing each of the rating scale items with sufficient attentiveness and carefulness and providing an accurate response to each single item, respondents may rather rely on the visual proximity of the items, which in turn results in less sensitivity to individual item characteristics and less differentiated responses to a grid compared to a single-item-per-screen format (Alwin, 2010; Tourangeau, et al., 2004; Weijters, Baumgartner, & Schillewaert, 2013).

High Cognitive Load and Respondent Frustration

With respect to the second approach suggesting that a high cognitive load is responsible for a respondent's susceptibility to cognitive shortcuts in grid questions, it is assumed that grid questions require more cognitive as well as navigational effort in providing an answer and thus, imply higher respondent burden than other question formats. The increased respondent burden is likely to produce adverse effects in terms of a decreased respondent motivation to engage in attentive and careful processing and an increased susceptibility to cognitive shortcuts. This in turn is reflected in various systematic response tendencies as well as in an increased risk of item missing data (Couper, et al., 2013; Dillman, et al., 2009, p. 179; Gräf, 2002; Kaczmirek, 2010; Krosnick, 1999; Peytchev, 2009; Toepoel, et al., 2009b).

The increased risk of relying on cognitive shortcuts in grid questions is due to several reasons. First, grid questions may carry an increased risk of high cognitive load because of the considerable amount of information presented simultaneously. Besides the sheer amount of information presented on a single screen, respondents have to link the information given in rows with the information presented in columns (Callegaro, Yang, Bhola, Dillman, & Chin, 2009; Couper, et al., 2013; Couper, et al., 2001; Dillman, et al., 2009, p. 179; Gräf, 2002; Kaczmirek, 2010; Tourangeau, et al., 2004). This poses a considerable sensorimotor challenge due to a respondent's limited field of

sharp vision that is restricted to about 2 degrees (Kahneman, 1973, p. 50). Relating to the processing of grid questions, this means that if respondents focus on one of the items arranged in a row, response option labels presented in the topmost cells will be outside their field of sharp vision and vice versa (Gräf, 2002). Thus, respondents have to split their attention to one rating scale component at a time, whereby careful processing of both the items and the response options requires increased sensorimotor effort. This effort is even more pronounced with an increasing number of rating scale items presented in a grid (Dillman, et al., 2009, p. 180; Gräf, 2002; Jenkins & Dillman, 1997).

However, this sensorimotor effort—engendered by the additional eye movements required for repeated matching of the item content and respective response option meanings—may increase the extent of the actual respondent burden and is likely to reduce respondent motivation to expend the effort required for attentive and careful question-answer processing (Couper, et al., 2013; Couper, et al., 2001; Dillman, 2000, p. 105; Kaczmirek, 2010). Respondents will be even more challenged and probably get confused, if the item content is varied in its polarity, i.e., if (counter)balanced rating scales are deployed and the meaning of some rating scale items is reversed (Callegaro, Shand-Lubbers, et al., 2009, p. 5887). Thus, in case of satisfying rather than optimal responding to grid questions, the respondents' attention to one question component may be at the expense of their attention to the other. As a consequence, the likelihood of relying on cognitive shortcuts is increased, which in turn is at the expense of complete and accurate responses to rating scale items arranged in a grid format. For instance, respondents are tempted to align their current answers with previous answers by simply orienting themselves towards the radio buttons ticked previously. Such an 'anchoring and adjusting' response behavior is associated with an increased risk of respondents losing track of related rows and columns and missing fine distinctions between single rating scale items. Thus, various kinds of systematic response tendencies as well as unintentional item nonresponse is more likely in grid questions than in less complex and burdensome question formats (Barge & Gehlbach, 2012; Callegaro, Shand-Lubbers, et al., 2009; Couper, et al., 2013; Gräf, 2002; Kaczmirek, 2010; Toepoel, et al., 2009b; Tourangeau, et al., 2004).

Second, the risk of relying on cognitive shortcuts due to an increased actual respondent burden may be further aggravated by an increased perceived respondent burden accompanied with the grid format per se.

Peytchev (2006) argued in view of Internet users being by now more versed in clicking and scrolling that satisficing behaviors are rather due to the increased perceived burden accompanied with the “daunting image” (p. 3) of a grid question instead of the actual burden resulting from excessive clicking or scrolling. Thus, respondents get the impression of enormous burden already at the initial stage of visual perception which may restrain them from optimizing in grids. Again, this perceived respondent burden is expected to be further increased with an increasing length of the rating scale (Dillman, et al., 2009, p. 180; Peytchev, 2006). Consequently, merely the visual complexity of a grid question and the initial impression of a burdensome task may decrease the respondent motivation to provide complete and accurate responses to the rating scale items presented in a grid format, even before starting the cognitive question-answer processing (Beatty & Herrmann, 2002; Couper, et al., 2013; Peytchev, 2006). This assumption is supported by the fact that—besides the increased risk of systematic response tendencies—grid questions actually suffer from higher rates of item nonresponse and survey breakoff compared to question formats with reduced visual complexity (Couper, et al., 2013; Knapp & Heidingsfelder, 1999; Peytchev, 2009; Toepoel, et al., 2009b).

In sum, the aforementioned considerations show that presenting rating scale items in close proximity by grouping them on the same screen can have a major impact on the cognitive and navigational processing of rating scale items in Web surveys. Besides the facilitation of the cognitive processing, grouping of rating scale items and presenting them in a grid format may also tempt respondents to process the rating scale items inattentively and carelessly. In this regard, it has not yet been clarified whether the increased risk of incomplete and inaccurate responses in grid questions compared to less burdensome question formats is either provoked by the respondent’s fatigue or the respondent’s frustration. To gain a better understanding of how grid questions facilitate the cognitive question-answer processing and, at the same time, also promote item missing data and systematic response tendencies, issues of item-order and response-order effects are discussed next as key aspects of context-related characteristics of rating scales.

4.3 Context-Related Characteristics of Rating Scales

Differences in the cognitive and navigational processing of rating scales depending on variations in the verbal and visual questionnaire features are often examined in terms of item-order and response-order effects. In general, order effects (or context effects) describe the fact that the order in which survey questions and respective response alternatives are presented has differing effects on the respondents' answers to the survey questions. More specifically, in the context of rating scales, item-order effects reveal the fact that the answer to a current item is affected by preceding items, whereas the response-order effects suggest that the order in which the response alternatives are presented affects their selection (Krosnick & Presser, 2010). In the next two sections, the main findings on item-order and response-order effects in rating scales are discussed with special focus on their implications for data accuracy. Furthermore, a better understanding of the item-order and response-order effects will help explain the potential differences in the cognitive and navigational processing of rating scales, arising as a function of presenting the items in a single-item-per-screen or multiple-item-per-screen design.

4.3.1 *Item-Order Effects*

Survey questions asking for attitudes and self-descriptions are particularly sensitive to external influences of the respective situation (Tourangeau, 1992). Thus, even minor changes in the order in which the items are presented can alter the responses to survey questions (Alwin, 2007, p. 124; Krosnick & Presser, 2010; Tourangeau & Rasinski, 1988; Tourangeau, et al., 1991). In general, item-order effects refer to the fact that preceding items affect responses to latter items by influencing one or more stages of the question-answer process. Consequently, changing the position of an item does not only change the sequence in which the items are processed, it also alters the overall context in which the items are processed (Knowles, 1988; Podsakoff, et al., 2003; Siminski, 2008; Strack, 1992; Tourangeau & Rasinski, 1988). Krosnick and Presser (2010) further distinguished serial-order effects from semantic-order effects. While serial-order effects refer to the location of an item within a sequence of items, semantic-order effects relate to the location of an item within a sequence of meanings. Both kinds of item-order effects affect the

question-answer processing in surveys, with serial-order effects potentially inducing learning and fatigue effects, whereas serial-order effects primarily lead to facilitation effects.

Serial-Order Effects

Serial-order effects can occur at the macro level of a questionnaire as well as at the micro level of single questions, with aspects of the respondent learning and the respondent fatigue being of special significance. Answering a number of survey questions, or in the present context, a set of several rating scale items may promote the learning effects since the respondents will become more practiced in conceiving the question format and the respective response options. This increased task familiarity may facilitate the response task, may shorten response times, and may decrease respondent burden (Callegaro, et al., 2005; Krosnick & Presser, 2010; Scherpenzeel & Saris, 1997). Furthermore, data accuracy is likely to increase because of the respondents' learning process. For instance, Couper and colleagues (2001) showed that with repeated exposure to constant sum questions, the respondents got more experienced with the question format. Accordingly, the increased experience in dealing with constant sum questions resulted in higher proportions of correct summations (Couper, et al., 2001). However, serial-order effects can also appear in terms of fatigue effects. In case respondents get increasingly tired or bored over the course of survey completion, or in the present context, over the course of rating scale completion, their inattentiveness and carelessness towards specific requirements of the response task is likely to increase (Callegaro, et al., 2005; Krosnick & Presser, 2010). Thus, the respondent fatigue potentially occurring during answering a set of rating scale items is closely related to the risk of systematic response tendencies such as speeding or nondifferentiation, and the increase in item missing data with an increasing number of questions already being answered (Alwin & Krosnick, 1985; Callegaro, et al., 2005; Galesic & Bosnjak, 2009; Gräf, 2002; Knowles, et al., 1992; Krosnick & Presser, 2010; Malhotra, 2009). The interplay between learning effects and fatigue effects has been illustrated by findings of Andrews (1984), revealing that poorer data accuracy came from survey questions placed at the beginning of a questionnaire (within the first 25 items) since the respondents had not yet warmed up to the response task, or at the end of a questionnaire (from 100 items on) as respondent fatigue and carelessness increased. Thus, both respondent learning and respondent fatigue

may occur at different times during the completion of a questionnaire, with different impacts on data accuracy.

At the macro level of the questionnaire, the overall length of a questionnaire and the respective position of a survey question within it may affect respondent burden and the respondents' susceptibility to cognitive shortcuts in answering survey questions later on in a questionnaire (Galesic, 2006; Herzog & Bachman, 1981; Krosnick & Presser, 2010; Peytchev, 2009). At the micro level of individual survey questions, the overall length of a rating scale may also decisively influence the risk of fatigue effects and the respondents' susceptibility to cognitive shortcuts (Andrews, 1984; Couper, et al., 2013; Dillman, et al., 2009, p. 179; Drolet & Morrison, 2001; Hinkin, 1995; Toepoel, et al., 2008). Thus, the increase in the perceived and actual respondent burden as a function of the overall length of rating scales is likely to enhance the risk of fatigue effects, which in turn increases the risk of cognitive shortcuts in answering rating scales. In this regard, Andrews (1984) showed that the overall length of a rating scale (not distinguishing between series and batteries) is a more important factor affecting data accuracy than the position of an item within a rating scale. The author concluded that "respondents [...] recognize the 'production line' character of this survey strategy and that it promotes carelessness in the way questions are asked and answered" (Andrews, 1984, p. 431).

Semantic-Order Effects

In attitude and personality measurement, different types of semantic-order effects can arise within each stage of the question-answer process described as carryover effects or backfire effects². The former effect results in responses that are highly similar to each other, whereas the latter effect evokes responses quite different from each other (Groves, 1989, p. 478; Schwarz & Bless, 1992; Strack, 1992; Tourangeau & Rasinski, 1988). Within the first step of question comprehension, framing effects may occur. Preceding items provide either a common framework for interpreting later items, whereby respondents tend to give responses similar to the preceding items, or there is a clear demarcation as the respondents' current responses differ from previous

² Also, the terms consistency and assimilation effect are common to address carryover effects, whereas the terms inconsistency and contrast effect refer to the same as backfire effects (Tourangeau, 1992).

ones in order to avoid redundancies with earlier answers. Carryover effects often occur with items dealing with unfamiliar issues, being ambiguous in content, or asking for inaccessible information (Tourangeau & Rasinski, 1988). Within the second stage of information retrieval, preceding items may influence later items in terms of priming effects. Irrespective of whether it relates to a conscious or automatic process, information is more accessible and likely to be used in answering the current item when information has already been retrieved through preceding items and deemed relevant to the present one. Otherwise, if the respondents consider the accessible information to be invalid or irrelevant for the present item, this information simply has no effect on the current response or provokes contrasting responses (Strack, 1992; Tourangeau, 1992). Within the third stage of judgment, preceding items may serve either as a joint or contrasting standard of comparison for the current judgment, which implies similar answers in the former case and deviating answers in the latter case. Thus, such anchoring effects can alter the respondents' mental representation in both directions as judgments on later items are likely to be influenced by preceding items, and at the same time, change the nature of the standard used to answer later items. In general, the respondents' attempts to avoid response inconsistencies promote carryover effects if preceding items are similar in content, whereas backfire effects are particularly likely if preceding items constitute an extreme point of comparison or are considered dissimilar (Krosnick & Presser, 2010; Tourangeau & Rasinski, 1988). Within the fourth stage of reporting, respondents are required to format their responses by mapping their judgment onto one of the predefined response options, and if deemed necessary, to edit their responses with respect to social desirability and consistency. A respondent's need for consistency plays a crucial role in evoking consistency effects and thus, responses that are highly similar to each other (Tourangeau, 1992; Tourangeau & Rasinski, 1988). Tourangeau and Rasinski (1988) noted that "the consistency effect should be greatest when a close, even logical, relation exists among the items, when this relation has been made salient, and when respondents are sufficiently involved in the issue to care about being consistent" (p. 308). Thus, grid questions presenting related rating scale items in close proximity may principally promote the occurrence of consistency effects at this final stage, provided that the respondents were sufficiently involved in the cognitive processing on previous stages and that they are actually concerned about providing consistent answers.

Consequently, preceding items may facilitate the interpretation of later items, the retrieval of relevant information and the judgment on later items by specifying the meaning of a question more clearly, by providing a common understanding of the item content, and by increasing temporary accessibility of information and judgments on the issues in question, all of which contribute to a decrease in the cognitive load and overall respondent burden in several respects (Knowles, 1988; Krosnick & Presser, 2010; Schwarz, Strack, & Mai, 1991; Sudman, et al., 1996, p. 70; Tourangeau & Rasinski, 1988; Tourangeau, et al., 1991). These facilitation effects related to the ease of the cognitive question-answer processing are reflected in decreased response times and the provision of more accurate answers to later survey questions (Callegaro, et al., 2005; Couper, et al., 2001; Knowles, 1988; Knowles & Byers, 1996; Tourangeau, 1992; Tourangeau & Rasinski, 1988; Tourangeau, et al., 1991). Knowles (1988) and colleagues (1996; 1992), for example, showed that although the item means remained unaffected, the item-total correlations among rating scale items were influenced by item order; the later an item appeared in a rating scale, the stronger the respondent's answers to single items were correlated with the total score. Based on their findings, Knowles and Byers (1996) summarized that "early questions clarify the meaning of a measure and improve the reliability of later answers" (p. 1080). The explanation the authors provided is that by answering several items belonging to the same rating scale, a specific meaning or interpretation of the items becomes salient which is equated with an enhanced understanding of the rating scale. This enhanced understanding of the theoretical construct underlying the rating scale may in turn facilitate succeeding stages of the question-answer process, particularly by making information retrieval and judgment more focused. Consequently, respondents are better able to hold a precise understanding of individual items, and beyond, to provide their responses more quickly (Knowles & Byers, 1996).

The facilitation of the response task due to semantic-order effects may be even more pronounced when sets of rating scale items are arranged in close proximity as they are in grid formats. Preceding items can be used more easily to answer later items as repeated interruptions in the respondent's concentration and annoyance are absent, occurring when the respondents otherwise try to remember previous items and responses, in order to take them into account when answering the current item (Dillman & Bowker, 2001; Dillman, et al., 1998; Krosnick & Presser, 2010; Podsakoff, et al., 2003;

Schwarz, Strack, & Mai, 1991; Tourangeau, 1992; Tourangeau, et al., 1991). Thus, the likelihood of semantic-order effects being noted is decisively affected by the grouping of rating scale items. However, since the respondents are expected to be involved in an optimal processing of survey questions rather rarely because of, among others, their limited motivation or time constraints, the impact of item order is even more pronounced which could also be at the expense of data accuracy: As respondents may simply rely on prior items to derive the meaning of later items instead of consciously processing the content of each single item, the respondents' answers to rating scale items are likely to assimilate in terms of "prior items [...] influencing the respondent's view of what issue the later item is supposed to be about" (Tourangeau & Rasinski, 1988, pp. 301-302). In addition, instead of endlessly searching for relevant information, respondents may rather draw on information that is easily accessible. Thus, since the preceding items increase the temporary accessibility of certain information, the salience of the same or closely related information is enhanced for later items. This may even be the case if the information is not particularly relevant, neither for preceding nor for later items (Schwarz & Bless, 1992; Tourangeau, 1992; Tourangeau & Rasinski, 1988). And, instead of repeatedly re-evaluating a judgment in an infinite loop of complex comparisons, the respondents are inclined to use salient anchors for their judgments and thus, are likely to use previous judgments as anchors for subsequent responses. Therefore, later judgments are assimilated in the direction of earlier ones (Harrison & McLaughlin, 1993; Schwarz & Bless, 1992; Tourangeau, 1992). In this regard, proximity effects can also be considered the result of cognitive shortcuts in survey responding according to which "respondents [...] may be less thoughtful about a question's meaning, they may search their memories less thoroughly, they may integrate retrieved information more carelessly, and/or they may select a response choice more haphazardly" (Krosnick, et al., 1996, p. 31).

In sum, the serial-order effects in terms of learning effects, and the semantic-order effects in terms of facilitation effects, may both unburden the respondents by facilitating the navigational and cognitive aspects of the response task. However, facilitation may in turn induce fatigue effects since respondents easily lose their motivation to spend the effort required for attentive and careful question-answer processing. By contrast, if learning effects fail to appear impeding the smooth navigation within a rating scale,

and/or the necessary question context is not provided impeding a comprehensive understanding of the underlying concept of a rating scale, respondents are potentially discouraged or distracted from accomplishing the actual response task. This higher level of task difficulty and increased respondent burden is likely to result in the respondents' frustration and susceptibility to cognitive shortcuts in answering rating scales.

4.3.2 Response-Order Effects

Previous research showed that the order in which the response alternatives are presented to respondents can affect their selection (Krosnick & Alwin, 1987; Krosnick & Presser, 2010; Krosnick & Schuman, 1988; Schuman & Presser, 1996, p. 56). Thus, when designing rating scales in Web surveys, survey researchers have to decide how to align the response alternatives in terms of their spatial and categorial arrangement with the objective of preventing or attenuating response-order effects.

Spatial Arrangement

Concerning the spatial arrangement of rating scales, response options can be aligned either vertically or horizontally. Whereas a vertical orientation of response options requires more space and extends the physical length of a rating scale, a horizontal orientation saves space and is more appropriate to convey the sense of continuous response options (Best & Krueger, 2004, pp. 58-60; Jenkins & Dillman, 1997). In addition, a horizontal orientation of response options allows for grouping a list of several items sharing the same response options together on one screen, either in a series or a grid format. Thus, in multiple-item-per-screen designs such as series or grids, response options are commonly presented in a horizontal orientation, whereas in single-item-per-screen designs both vertical and horizontal orientation is conceivable (Dillman, et al., 2009, p. 145). However, otherwise identical rating scales solely differing in the horizontal and vertical order of response options may not necessarily elicit the same answers: In rating scales with horizontally arranged response options, it can be assumed that more hand-eye movements between the items and the response options provided on the outmost right become necessary. Therefore, a shift in mean responses towards the left side of the response scale is likely to be observed (Toepoel, et al., 2009a). Previous findings concerning a comparison between horizontally and

vertically aligned response options are rare but largely consistent. Contrary to expectations, no differences in mean responses or response distributions were found as a function of presenting rating scale response options either in a vertical or horizontal arrangement (Callegaro, Shand-Lubbers, et al., 2009; Funke & Reips, 2009; Funke, et al., 2011; Toepoel, et al., 2009a).

Categorical Arrangement

Besides variations in the vertical or horizontal arrangement, verbally labeled response options in rating scales can be presented either in a positive-to-negative or negative-to-positive scale arrangement. The interpretive heuristics ‘left and top mean first’ and ‘up means good’ may provide advice concerning the categorical scale arrangement (see also section 2.2.3). The heuristic ‘left and top mean first’ refers to a respondent’s reading sequence of the German language as well as of most Western languages according to which respondents start reading from top to bottom and from left to right. This heuristic provides an important implication for the design of verbally labeled rating scales. When presenting a number of ordered response options, respondents expect both ends of the scale to represent opposing endpoints with the remaining response options being arrayed in a logical sequence (Jenkins & Dillman, 1997; Tourangeau, et al., 2004). Furthermore, the interpretive heuristic ‘up means good’ as a variant of the former heuristic implies that respondents expect the affirmative response option to be presented first in both a vertically or horizontally aligned list of ordered response options (Holbrook, Krosnick, Carson, & Mitchell, 2000; Tourangeau, et al., 2004). Both heuristics affect a researcher’s decision concerning the categorical arrangement of verbally labeled rating scales by recommending aligning response options in a logical sequence and starting with the positive end of the scale at the very top in a rating scale with vertically aligned response options, or leftmost in a rating scale with horizontally aligned response options. Previous findings showed that when rating scales conform to the ‘left and top mean first’ and ‘up means good’ heuristics, i.e., when response options are logically ordered from the most positive to the most negative response option, respondents are able to provide their responses more quickly and more accurately, while any deviation from these heuristics may lead to a respondent’s confusion being reflected in longer response times and less accurate answers (Christian & Dillman, 2004; Christian, et al., 2009; Holbrook, et al., 2000; Toepoel, et al., 2008;

Tourangeau, et al., 2004). Therefore, presenting response options in the unconventional negative-to-positive order can result in the respondents' distraction from the actual response task and superficial answering rather than enhanced attention and more thoughtful answering. This is in line with the findings of Holbrook and colleagues (2000) who explicitly asked respondents to list all their thoughts that came to their mind while thinking about a survey question. In fact, they found that a negative-to-positive response option order resulted in more irrelevant thoughts distracting respondents from thinking about the actual survey question. Additionally, Salzberger and Koller (2013) demonstrated that a positive-to-negative response option order allowed for a better measurement of latent constructs than the reverse scale arrangement. Toepoel and colleagues (2009a), however, found no significant differences in response times depending on the categorial scale arrangement, and consequently, no evidence of significantly longer response times with the negative-to-positive as compared to the positive-to-negative response option order.

Up to this point, previous studies indicated that a positive-to-negative response option order facilitates the response task and helps respondents fully concentrate on the content of respective rating scale items. However, one needs to keep in mind that the alteration of the categorial arrangement of verbally labeled rating scales impacts the order in which respondents will read and process the response option labels. Furthermore, by reversing the scale arrangement, the relative position of a response option is changed as well as the context in which the response options are processed (Chan, 1991; Toepoel & Dillman, 2010). Hence, the categorial arrangement of rating scale response options, i.e., whether the positive or negative end of the scale is presented first, may also have an effect on the occurrence of response-order effects (Christian, et al., 2008; Toepoel & Dillman, 2010). In general, response-order effects—occurring when response distributions are affected by the order in which the response options are presented—can be explained by the extent of cognitive processing varying depending on the mode of presenting the response options. In interviewer-administered surveys, the response options are presented verbally and recency effects predominate. In self-administered surveys, however, the response options are commonly presented visually and primacy effects are to be expected (Krosnick, 1991; Krosnick & Alwin, 1987). This general distinction applies to survey questions with unordered response alternatives. In rating scales with ordered response options, primacy

effects tend to prevail regardless of whether response options are presented verbally or visually (Krosnick, et al., 2005; Sudman, et al., 1996, p. 157f). Thus, response options at the positive end of a conventional positive-to-negative scale arrangement or at the negative end of a negative-to-positive scale arrangement have an increased probability of being selected, irrespective of the mode of data collection.

Despite certain inconsistencies in previous findings on primacy effects in Web surveys, the vast majority of studies provide evidence of primacy effects in terms of an increased endorsement of response options at the beginning of a rating scale. Key findings of previous studies are described in brief. Callegaro and colleagues (2008) examined the effect of reversing the categorical scale arrangement of 7 rating scale items presented on single screens with vertically aligned response options. The authors found significant primacy effects in response distributions of 4 out of 7 items: The most positive response option 'very satisfied' was selected significantly more often when presented first than when presented last, and conversely, significantly more respondents selected the most negative response option 'very unsatisfied' when presented first rather than when arranged last. Similarly, Toepoel and colleagues (2009a) found significant primacy effects for two rating scale items vertically aligned on single screens. Malhotra (2009) demonstrated that for 6 out of 7 vertically arranged rating scale items, respondents were significantly more likely to select response options listed near the top rather than response options listed toward the bottom. Primacy effects were more pronounced among lower educated respondents. Keusch (2012) reported three different studies examining the extent of primacy effects in rating scales depending on the categorical scale arrangement: In a first online panel survey, respondents showed significantly more agreement when they were confronted with a positive-to-negative scale arrangement compared to those answering the same list of 11 rating scale items in a grid format on a negative-to-positive scale. In a second study conducted among general students, the author found significant primacy effects for each of the 4 rating scale items presented in a grid format. Separate analyses depending on the respondents' prior Web survey experience revealed a significant interaction effect indicating that primacy effects were fully attributable to respondents who took part in at least one Web survey before, whereas respondents who have never participated in a Web survey showed no significant primacy effects. In a third study conducted among small business owners, significant

primacy effects were found only for 5 out of 16 items (Keusch, 2012). Keusch (2012) concluded that rating scales are generally prone to primacy effects, and that respondents with certain Web survey experience are specifically prone to primacy effects, resulting in a higher probability of choosing response options presented at the beginning of a horizontally aligned rating scale. Similarly, for vertically aligned rating scales in Web surveys, Toepoel and colleagues (2008) found an overall effect of respondents selecting the positive response option 'very good' more often when items were presented in the second position of a positive-to-negative scale than when presented in the fourth position of a negative-to-positive scale. Furthermore, they found a significant interaction effect with prior survey experience according to which trained respondents were more prone to primacy effects than fresh respondents.

The studies presented revealed clear evidence of primacy effects in rating scales, irrespective of a vertical or horizontal scale arrangement and irrespective of whether the items are presented in a single-item-per-screen or multiple-item-per-screen format. However, there are other studies which found no significant effect of variations in the categorial arrangement of rating scales on the response distributions, scale means and other scale characteristics such as inter-item correlation, covariance, and factor structure, neither in horizontally nor in vertically aligned rating scales (Christian, et al., 2008; Christian, et al., 2009; Hofmans, et al., 2007; Rammstedt & Krebs, 2007; Toepoel & Dillman, 2008; Weng & Cheng, 2000). Studies systematically comparing variations in the categorial scale arrangement and their effect on the prevalence of primacy effects in a single-item-per-screen versus a multiple-item-per-screen format are not known.

In summary, especially in view of the inconsistencies in previous findings regarding the prevalence of primacy effects in rating scales, the convention to offer the positive response alternative first has been established in questionnaire design. Due to the fact that respondents expect to start with the affirmative response option, response options in rating scales are conventionally aligned in a positive-to-negative order. Disregarding this convention is a potential source of measurement error due to the respondents' confusion and interruptions in the question-answer process by distracting their attention away from simply answering the question (Christian & Dillman, 2004; Christian, et al., 2009; Holbrook, 2008; Holbrook, et al., 2000; Tourangeau, et al., 2004).

In view of the above, there are several advantages of applying a rating scale and grouping the rating scale items in a grid format. By presenting related rating scale items in close proximity, the question context necessary for a complete understanding of the underlying theoretical construct is preserved and fluent cognitive and navigational processing of the items is enabled. Overall, this is likely to facilitate the question-answer processing and unburden the completion of a set of rating scale items. Furthermore, compliance with the respondents' expectations by presenting the response options in a positive order may facilitate the response task. At the same time, however, grouping of rating scale items and presenting them in a grid format may also encourage respondents to fall back on cognitive shortcuts and to further reduce the effort expended on attentive and careful processing of the rating scale. The implications for data accuracy need to be carefully assessed, which in turn requires a set of data accuracy indicators outlined in the next sections.

5. ASSESSING DATA ACCURACY IN RATING SCALES

It is recommended that survey questions in general and rating scale items in particular are grouped together provided that they are about the same or similar issues (Krosnick & Presser, 2010). Beyond practical considerations such as the space-saving implementation and time-saving processing of rating scales, presenting rating scales in a multiple-item-per-screen design such as a grid format is assumed to enable a complete understanding of the item set as a whole (Couper, et al., 2001; Schwarz, Strack, & Mai, 1991; Toepoel, et al., 2009b). However, presenting rating scale items in close proximity may also induce respondents to see the rating scale items more related as if the same items are presented separately. This in turn may have differing implications for the cognitive processing of rating scales (Couper, et al., 2001; Schwarz, Strack, & Mai, 1991; Tourangeau, et al., 2004).

As Couper and colleagues (2001) mentioned, higher correlations between rating scale items as a result of grouping related items on a single screen or in a grid are considered neither desirable nor problematic at first glance. On the one hand, increased internal consistency may be interpreted as higher scale reliability. On the other hand, however, it either may simply reflect semantic redundancy and lower between-item discrimination of rating scale items, or may be an indication of systematic measurement error (Alwin, 2010; Callegaro, Shand-Lubbers, et al., 2009; Diamantopoulos, et al., 2012; Drolet & Morrison, 2001; Peytchev, 2006; Viswanathan, 2005, p. 110). Thus, with respect to the latter issue, higher correlations among rating scale items are not necessarily tantamount to more accurate responses because higher correlations can also be due to the respondents' reliance on cognitive shortcuts, resulting in their failure to perceive distinctions in item meanings and to sufficiently differentiate between the items (Callegaro, Shand-Lubbers, et al., 2009; Peytchev, 2006; Tourangeau, et al., 2004). If respondents are not sufficiently involved in attentive and careful processing but rather rely on superficial processing, "respondents seem[ed] to use the proximity of the items as a cue to their meaning, perhaps at the expense of reading each item carefully" (Tourangeau, et al., 2004, p. 390). Likewise, Alwin (2010) annotated that "similarity of question content and response format may distract respondents from fully considering what information is being

requested, making them less attentive to the specificity of questions” and making them more prone “to ‘streamline’ their answers when investigators ‘streamline’ questionnaires” (p. 424). Thus, respondents may simply rely on the proximity of rating scale items presented in a grid format without processing the items and their response options with sufficient attentiveness and carefulness, resulting in less sensitivity to individual item characteristics and less differentiated responses (Alwin, 2010; Tourangeau, et al., 2004; Weijters, et al., 2013). As a result, the internal consistency of the respondents’ answers to grid questions is likely to be overestimated, while the respondents’ true variation in attitudes and self-descriptions will be underestimated (Barge & Gehlbach, 2012; Callegaro, Shand-Lubbers, et al., 2009; Harrison & McLaughlin, 1993; Knowles & Byers, 1996; Peytchev, 2006; Tourangeau & Rasinski, 1988).

After outlining key findings regarding the grouping hypothesis and potential increases of the internal consistency of rating scale measures (see section 5.1), various kinds of cognitive shortcuts typically encountered with rating scales in general and with grid questions in particular are discussed (see section 5.2.1 to section 5.2.5). As already mentioned in section 2.1.1, cognitive shortcuts in the sense of merely satisfying rather than optimal survey responding cannot be observed directly. Hence, several systematic response tendencies are distinguished that are supposed to reflect a respondent’s superficial processing of survey questions and affect data accuracy in rating scales negatively. In addition, the extent of item missing data can be used as a rather indirect indicator of data accuracy in rating scales, with item nonresponse and survey breakoff being commonly considered the result of a respondent’s difficulties in processing a survey question with difficulties potentially occurring on one or more stages of the question-answer process (see section 5.3). Thus, what was yet referred to as completeness and accuracy of survey responses, is now explored in detail in terms of more concrete indicators of data accuracy. Besides the completeness of survey responses referring to the occurrence of item nonresponse and survey breakoff, the accuracy of survey responses is assessed in terms of various kinds of systematic response tendencies including reversed-item bias, nondifferentiation, acquiescence, extremity, and primacy effects. Lastly, response times are deemed to reflect a respondent’s difficulties in processing a survey question, whereas at the same time, it may also be indicative of a respondent’s susceptibility to cognitive shortcuts within the question-answer

process. Thus, it can be used as another indirect indicator of data accuracy in rating scales (see section 5.4). The following remarks on differing manifestations of cognitive shortcuts relate to both the theoretical basics of their underlying causes and the empirical findings that are available at present.

5.1 Internal Consistency

Whereas the rating scale format, i.e., presenting several rating scale items either in a single-item-per-screen format or in a multiple-item-per-screen format such as a grid format obviously had no effect on mean responses (Bradlow & Fitzsimons, 2001; Callegaro, Shand-Lubbers, et al., 2009; Peytchev, 2006, 2007; Toepoel, et al., 2008, 2009b; Yan, 2005), previous findings concerning the grouping hypothesis and the associated increase in the internal consistency of rating scale measures in Web surveys were mixed. Tourangeau and colleagues (2004) found significantly higher correlations between rating scale items when the items were presented in a grid format, indicating that responses to rating scale items are more consistent when all items are presented in close proximity: Based on 8 rating scale items measured on an endpoint labeled 7-point agree-disagree scale with 2 of the items being reversed in content, Cronbach's alpha was significantly higher when items were presented in a single grid (.621) compared to when they were separated into two grids with 4 items each (.562), or when each item was presented on a single screen (.511). Based on a reanalysis of Tourangeau and colleagues (2004) findings, Peytchev (2006) found a higher internal consistency in grid questions in terms of increased covariances between the last item and each of the other items when all items were presented in a single grid compared to when the items were presented in two grids on separate screens or each item on a single screen. Based on an unbalanced rating scale with 11 items being measured on a 5-point Likert-type scale, Couper and colleagues (2001) presented a set of rating scale items in a single-item-per-screen design or in several grids with either 4, 4, 3, or 5 items, respectively. They found consistently higher Cronbach's alpha coefficients among the items when they were grouped together on a single screen (.640) than when the items were spread across several screens (.610). However, these differences were not statistically significant. Based on a balanced rating scale

with 40 items measured on a 5-point Likert-type scale, Toepoel and colleagues (2009b) also reported a slightly higher Cronbach's alpha with grids of 10 items (.887), followed by grids of 4 items (.885) and the single-item-per-screen format (.880), whereas presenting every 40 items in a single grid yielded the lowest inter-item correlations (.879). However, these differences also failed to reach statistical significance. In a 2 x 2 x 2 between-subjects factorial design, Toepoel and Dillman (2008) varied the format of an unbalanced rating scale consisting of 5 items that were measured on a 5-point Likert-type scale. The rating scale items were presented either in a grid format or on 5 separate screens. Additionally, verbal labels were used with the response options being either endpoint labeled or fully labeled. And last, numbers were added to each of the response options. The authors found a significantly higher Cronbach's alpha when items with endpoint labeled response options were presented in a grid format (.631) compared to a single-item-per-screen format (.483). This difference got smaller when numbers were added to the response options (.640 and .557, respectively), and disappeared completely when verbal labels were added to each of the response options (.567 in the grid format and .567 in the single-item-per-screen format when the scale was fully labeled; .605 in the grid format and .634 in the single-item-per-screen format when the scale was fully labeled and numbers were added). Based on these findings, Toepoel and Dillman (2008) concluded that "respondents are more likely to use visual language when verbal and numerical labels provide minimal support" (p. 18). This interpretation is in line with Tourangeau and colleagues (2007a) who proceed on the assumption that verbal labels, numerical labels, and purely visual cues can be ranked in descending order in terms of their impact on response decisions made during the question-answer process. This explains why proximity effects according to the 'near means related' heuristic diminished in rating scale formats with fully labeled response options. Based on this conclusion, Toepoel and Dillman (2008) were able to explain the mixed findings of previous studies since Couper and colleagues (2001) and Toepoel and colleagues (2009b) used fully labeled rating scales and failed to prove significant differences between grid formats and single-item-per-screen formats, whereas Tourangeau and colleagues (2004) used endpoint labeled rating scales and found significant differences between the two formats.

Thus, previous findings concerning the internal consistency of rating scale measures depending on a single-item-per-screen or multiple-item-per-

screen design in terms of a grid format remain inconclusive. Most of the studies found that the more rating scale items are presented in a grid, the more likely to result in increased internal consistency of rating scale measures. However, most studies found no significant differences between differing rating scale designs. Despite the mixed findings concerning a proximity effect in terms of an increased internal consistency of rating scale measures when the items are presented in close proximity, the prevailing view is still that respondents are more likely to regard items as belonging together when they are presented in a grid format (Callegaro, Shand-Lubbers, et al., 2009; Couper, et al., 2001; Peytchev, 2006; Toepoel, et al., 2008; Toepoel & Dillman, 2008; Tourangeau, et al., 2004).

Nevertheless, variations in the internal consistency of rating scale measures are considered an inadequate indicator of a respondent's susceptibility to cognitive shortcuts and overall data accuracy of rating scale measures. First, one of the most severe shortcomings is the fact that the internal consistency can be inflated because of the correlated nonrandom sources of measurement error. For instance, because of a respondent's susceptibility to cognitive shortcuts in terms of simply selecting the same response option for all the rating scale items, the internal consistency will be spuriously increased. Moreover, the internal consistency is further increased because of a respondent's need for consistency. Second, there are usually no objective standards to determine a critical threshold in rating scale measures indicating a spuriously inflated internal consistency. Furthermore, besides a lack of significant differences between varying rating scale formats, another shortcoming of the studies presented is that they disregarded whether unbalanced or (counter)balanced rating scales were used, with the implications for the interpretation of the internal consistency of rating scales being discussed in the next section.

In the following, the focus is on two specific indicators of reduced data accuracy in rating scales: reversed-item bias and nondifferentiation. Both are regarded as systematic response tendencies arising from the proximity of rating scale items when arranging them in a grid format. Furthermore, acquiescence and extremity are discussed as to systematic response tendencies frequently encountered with rating scales in general. Primacy effects emerge in all kinds of closed questions presented in self-administered surveys and thus, can also be found in rating scales presented in a Web survey.

5.2 Systematic Response Tendencies

5.2.1 *Reversed-Item Bias*

In general, (counter)balanced rating scales containing positively and negatively worded items that measure the same underlying theoretical construct typically yield a lower internal consistency because of the weaker correlations between the original and reversed items, which is also referred to as reversed-item bias (Barnette, 2000; Harrison & McLaughlin, 1993; Weijters, et al., 2009). The occurrence of reversed-item biases is commonly attributed to careless responding which is described as a respondent's inattentiveness or carelessness towards the reverse wording of rating scale items: Instead of consciously processing the content of each single rating scale item, respondents rather rely on the context being created by previously answered items. More precisely, the meaning of the current item is likely to be derived from the item meanings of previous items, i.e., from a respondent's established expectation of what construct the rating scale is intended to measure (Podsakoff, et al., 2003; Tourangeau & Rasinski, 1988; Weijters, et al., 2013). As a consequence, a respondent's attention to item specifics may be reduced, which in turn increases the risk of missing the reverse wording of rating scale items. Although reversed-item bias as a consequence of careless responding is not explicitly listed among the systematic response tendencies explained by the satisficing theory, careless responding can nevertheless be regarded as a form of strong satisficing, and probably even one of the strongest forms of satisficing because a respondent is not even willing to read or interpret the content of rating scale items as deemed necessary within the first stage of the question-answer process to gain a thorough understanding of the items. In order to minimize the cognitive effort necessary to answer the rating scale, respondents may even skip this first step of question comprehension and rather rely on the surrounding information provided by previously answered items to answer the current item.

In grid questions, a lot of information is presented simultaneously in rows and columns. This may result in a higher risk of respondent confusion about the respective item meaning and the appropriate linkage of information presented in rows and columns. Furthermore, grid questions may carry a higher risk of processing each item and the respective response options less carefully, whereas in a single-item-per-screen format respondents are expected to pay more attention to each single item (see also section 4.2)

(Callegaro, Shand-Lubbers, et al., 2009; Couper, et al., 2013; Gräf, 2002; Tourangeau, et al., 2004). Hence, if grid formats actually alter the respondents' focus on individual items and restrain them from processing the meaning of each single item attentively and carefully, while single-item-per-screen formats promote higher focus on item idiosyncrasies, responses to (counter)balanced rating scales presented in a grid format will be more inconsistent, and thus, more prone to reversed-item biases than in a single-item-per-screen design (Callegaro, Shand-Lubbers, et al., 2009; Grandmont, et al., 2010; Toepoel, et al., 2008; Tourangeau, et al., 2004). Simply put, in unbalanced rating scales, a higher internal consistency is to be expected in grid formats because of proximity effects. In (counter)balanced rating scales, however, lower correlations between the rating scale items are expected in grid formats because of the higher risk of response inconsistencies in the respondents' answers to original and reversed items (Callegaro, Shand-Lubbers, et al., 2009).

Previous findings on the differences in reversed-item biases in (counter)balanced rating scales as a function of different rating scale formats are outlined briefly below. Based on two different counterbalanced rating scales, Callegaro and colleagues (2009) found—as it was expected—a higher Cronbach's alpha (.851 and .868, respectively) and higher correlations between the original and reversed items after recoding the reversed ones (.593 and .652, respectively) when the rating scale items were presented in a single-item-per-screen format as compared to a grid format (Cronbach's alpha: .823 and .849, respectively; correlations between the original and reversed items: .535 and .614, respectively). However, differences between the two rating scale formats did not reach statistical significance (Callegaro, Shand-Lubbers, et al., 2009). Nonetheless, Callegaro and colleagues (2009) concluded on the basis of their findings that “on a grid there are more chances to ‘miss’ the meaning of the items, resulting in more inconsistencies than when evaluating one item per screen” (p. 5895). Based on a balanced rating scale consisting of 6 pairs of original and reversed items, Grandmont and colleagues (2010) found consistently lower negative correlations between respective item pairs when presenting them in a grid format as compared to a single-item-per-screen format. The differences for 3 out of 6 item pairs were statistically significant (-.281 and -.433, -.281 and -.437, -.480 and -.639, respectively). The authors concluded that presenting rating scales in a grid format attenuates the respondents' sensitivity to the reverse wording of rating scale items,

compared to presenting them in a single-item-per-screen design. No significant differences were found compared to presenting the items either in several smaller grids or in a series of items on the same screen (Grandmont, et al., 2010). Similarly, Tourangeau and colleagues (2004) found that respondents were less sensitive to the reverse wording of items when they were presented in a grid format, which found its expression in weaker item-total correlations between each of the two reversed items and the overall scale score, when all 8 items were arranged in a single grid (-.331 and -.097, respectively) as compared to two grids (-.395 and -.151, respectively), or a single-item-per-screen format (-.427 and -.187, respectively) (Tourangeau, et al., 2004). However, the statistical significance was not explicitly specified. Furthermore, Toepoel and colleagues (2008) showed that respondents who already acquired certain survey experience were less likely to notice the reverse wording of rating scale/binary items when the items appeared on a single screen compared to respondents who had little or no prior experience. This was reflected in weaker item-total correlations between each of the 5 reversed items and the overall scale score when all 10 rating scale items were arranged in a single grid as compared to when the items were presented in two grids or in a single-item-per-screen format for trained respondents (Toepoel, et al., 2008). Again, statistical significance was not explicitly specified. Based on their findings, Toepoel and colleagues (2008) suggested that experienced respondents who have already developed a certain routine in survey completion process items that are presented in a grid format less attentively than inexperienced respondents which may be at the expense of a respondent's sensitivity to fine distinctions in item meaning. Thus, instead of processing each item carefully, experienced respondents rather rely on the proximity of the items to deduce the meaning of the current item from preceding items whereas inexperienced respondents still pay more attention to each single item (Toepoel, et al., 2008). The findings of Weijters and colleagues (2013) concerning their expectation that careless responding is more pronounced when related items are grouped together instead of being dispersed across the questionnaire, remained inconclusive since in one study expectations could be confirmed, whereas in another study no significant differences were found. These mixed findings were explained by the generally low incidence of careless responding in the dataset (Weijters, et al., 2013).

In sum, previous evidence of artificially increased internal consistency in grid questions due to providing rating scale items in close proximity is rather inconclusive. Therefore, internal consistency appears as an inadequate indicator of a respondent's increased susceptibility to cognitive shortcuts and reduced response accuracy in rating scales. Although previous findings on response inconsistencies between the original and reversed items are rare and partly fail to reveal large differences between various rating scale formats, decreased correlations between the original and reversed items in (counter)balanced rating scales are considered here to be a more adequate indicator of a respondent's susceptibility to cognitive shortcuts and decreased data accuracy in rating scales, compared to the internal consistency of rating scales. In line with the theoretical assumptions set out above, the findings of previous studies provide initial evidence that respondents are more prone to inattentive or careless processing of rating scale items and thus, less likely to take note of the reverse wording of rating scale items, when the items are presented in close proximity in terms of a grid format as compared to presenting them on separate screens (Barnette, 2000; Callegaro, Shand-Lubbers, et al., 2009; Harrison & McLaughlin, 1993; Weijters, et al., 2009). Hence, the risk of reversed-item biases is increased, which in turn may have adverse effects on the internal consistency and factor structures of rating scale measures (Barnette, 2000; Harrison & McLaughlin, 1993; Weijters, et al., 2009).

5.2.2 Nondifferentiation

In general, it is assumed that a high level of respondent burden in grid questions lowers respondent motivation (see also section 4.2.2). "Given their lowered motivation, respondents are then likely to look for easier ways of responding [...] [whereby] long item sets provide an inviting setting for adopting a uniform and therefore less taxing response strategy" (Herzog & Bachman, 1981, p. 558). Such uniform and therefore, less taxing responding in grid questions is also described as nondifferentiation. In general, nondifferentiation among rating scale items describes a respondent's tendency to use the same or nearly the same response options to answer a set of several rating scale items instead of making use of the full range of response options (Krosnick, 1991; Krosnick & Alwin, 1988; McCarty & Shrum, 2000). Herzog and Bachman (1981) first mentioned this kind of systematic response

tendency in terms of ‘straight-line responding’ referring to respondents answering a number of consecutive rating scale items with the identical response option as distinct from ‘almost-straight-line responding’ according to which respondents select nearly the same response option for each of the rating scale items. In the following, the term nondifferentiation will be used inclusively without distinguishing between straight-line and almost straight-line responding unless indicated explicitly.

According to the theory of survey satisficing, nondifferentiation is explained by a respondent’s inability or unwillingness to conscientiously process the four steps of the question-answer process with the stage two and three of information retrieval and judgment being affected in particular. Nondifferentiated responding as a respondent’s failure to sufficiently differentiate among a set of rating scale items is considered a form of strong satisficing because respondents simply answer the first item by selecting a response option that seems reasonable and adjust subsequent answer choices to this response option (Krosnick, 1999). However, Herzog and Bachman (1981) noted that “respondents did not stop reading altogether; rather, when they found a large set of relatively less interesting items they tended to slip into a comfortable ‘groove’ that allowed them, in effect, to skip on to the next questions” (p. 554). More specifically in the context of Web surveys, Gräf (2002) presented a further navigation-based rather than cognitive-based argument by stating that “the effort involved in positioning the cursor anew in each row is more costly than continuing in the same or neighboring column for each row” (Gräf, 2002, p. 56).

Since a certain degree of differentiation among a set of rating scale items requires more cognitive effort than simply rating all items equally or at least similarly, respondents with a low level of cognitive ability are deemed susceptible to nondifferentiation (Kaminska, et al., 2010; Krosnick & Alwin, 1988; McCarty & Shrum, 2000). In fact, respondents with lower cognitive ability and lower cognitive sophistication are more prone to nondifferentiation when using the level of educational attainment as a surrogate measure for cognitive ability and cognitive sophistication (Kaminska, et al., 2010; Krosnick & Alwin, 1988; Krosnick, et al., 1996; McCarty & Shrum, 2000). Aside from the respondent’s ability to sufficiently differentiate among several rating scale items, respondent motivation is considered a decisive factor for the risk of nondifferentiation since respondents with low motivation may strive to minimize their effort required to answer a question by simply rating

all items more or less equally (Herzog & Bachman, 1981; Krosnick, 1999; Krosnick & Alwin, 1988; McCarty & Shrum, 2000). Accordingly, findings of Herzog and Bachman (1981) indicated that while a decrease in respondent motivation over the course of the completion of long questionnaires (about 2 hours) was observed, the extent of nondifferentiated responding increased in a self-administered paper-based questionnaire. By manipulating the length of the questionnaire (10, 20, and 30 minutes) and the position of the grid questions within the questionnaire, Galesic and Bosnjak (2009) also showed that the average variance of answers to grid questions progressively reduced, the later the grids were positioned within a Web survey. Similarly, Cole and colleagues (2012) and Taylor (2006) showed that nondifferentiated responding occurred in particular for rating scale items presented later in a Web questionnaire. These findings of an increasing risk of nondifferentiated responding with a growing number of questions already answered, support the “researchers’ views of nondifferentiating respondents’ answers to rating questions as invalid and as reflecting lack of motivation” (Krosnick & Alwin, 1988, p. 536). However, Herzog and Bachman (1981) have also shown that the negative effect of decreased respondent motivation in long questionnaires on the extent of scale differentiation can be moderated by the respondent’s interest in survey topic and the personal relevance of question content. More generally, Krosnick and colleagues (1996) showed that a higher perceived value of the survey is positively related to scale differentiation in a self-administered paper-based questionnaire. Chang and Krosnick (2009) showed as well that a respondent’s interest in the survey topic may increase the extent of scale differentiation in a Web survey.

Aside from the respondent ability and respondent motivation, nondifferentiation in rating scales is decisively affected by task difficulty because respondents may just rate all items equally or nearly equally in order to reduce the complexity of the response task (Krosnick & Alwin, 1988). By implication, from a survey researcher’s perspective, another strategy to reduce the risk of satisficing would be to simplify the response task (Couper, et al., 2013; Krosnick & Alwin, 1987; McCarty & Shrum, 2000). For example, McCarty and Shrum (1997, 2000) showed that nondifferentiation can be reduced by simplifying the rating task in terms of splitting it up in a two-step process: Respondents were asked to rank the items first by selecting the most and least important items before they were requested to rate the remaining items. Using an attitude scale comprising 9 items with response options

ranging from 1 ('very unimportant') to 10 ('very important') in a paper-based survey, the degree of differentiation was significantly increased in the 'most-least' procedure compared to the conventional 'rate-only' method in a student sample (.67 and .56, respectively), as it was also the case for a study among the general population (.70 and .60, respectively) (McCarty & Shrum, 1997, 2000). Similarly, Couper and colleagues (2013) argued that "reducing cognitive load may decrease satisficing by increasing attention to the task" (p. 326). However, with regard to the risk of nondifferentiation, this general assumption could not be confirmed since (a) splitting grid questions comprising behavioral-frequency items into component parts to reduce the complexity of the response task, and (b) providing visual feedback to ease navigation within a grid question, had no effects on the risk of straight-lining among opt-in panel members in a Web survey (Couper, et al., 2013). Based on these findings, it remains unclear whether splitting grid questions or providing visual feedback actually has any effect on the extent of straight-lining, or whether a lack of significant differences is rather due to the use of behavioral-frequency questions and their comparatively low susceptibility to straight-lining. At least in respect of attitude measures, Alwin and Krosnick (1985) hold another view by arguing that "making the task easier may also reduce respondents' willingness to make more precise distinctions about the relative importance of valued qualities" (p. 537). According to this, low levels of task difficulty may be the reason why rating scales suffer from systematic response tendencies such as nondifferentiation.

Previous findings systematically examining the extent of nondifferentiation in rating scales, depending on presenting the items either in a multiple-item-per-screen design in the form of a grid question or in a single-item-per-screen design, are rare because nondifferentiation is considered a systematic response tendency that is explicitly inherent to the nature of the grid format. Results from the very few studies are mixed. Tourangeau and colleagues (2004) revealed that the respondents' answers to an attitude measure consisting of 8 rating scale items with 2 of the items being reversed in content, featured the highest degree of nondifferentiation when all 8 items were presented in a single grid as compared to separating the items into two grids with 4 items each or compared to presenting each item on a single screen. The differences between all three rating scale formats were statistically significant (.436, .422, and .412, respectively, with higher values indicating a higher degree of nondifferentiation) (Tourangeau, et al., 2004).

Based on a balanced rating scale consisting of 12 items, Grandmont and colleagues (2010) examined variations in the extent of straight-lining as a function of the rating scale format and distinguished between a single-item-per-screen design and three different multiple-item-per-screen designs (a standard grid, several individual grids with less items, and a series). As expected, a single-item-per-screen design entailed the lowest share of straight-lining respondents (5% of all 66 straight-lining respondents). Among the multiple-item-per-screen designs, however, the standard grid induced less straight-lining (9%) than a series (35%) or partitioned grids (52%). Whereas the single-item-per-screen format and standard grid format did not differ significantly, both formats yielded a significantly lower proportion of straight-lining respondents compared to the series and partitioned grid format. Grandmont and colleagues (2010) concluded that if respondents are confronted with a longer grid question, they make a conscious effort “to not give the ‘appearance’ of entering straight-lined data” (p. 5956). Lastly, Couper and colleagues (2001) found no significant differences in the extent of straight-lining when the 11 items of an attitude measure were presented either in several grids with 4, 4, 3, and 5 items, respectively, or in a single-item-per-screen design.

In sum, the degree of nondifferentiation among rating scale items is a common measure used to assess a respondent’s susceptibility to cognitive shortcuts in grid questions, and is frequently used as an indicator of data accuracy in rating scale measures. In line with the theory of survey satisficing, the risk of nondifferentiation increases with decreased respondent ability and respondent motivation. However, there is still continuing disagreement concerning the extent of task difficulty in grid questions and its impact on the risk of nondifferentiation (see also section 4.2.2). In general, a lack of scale differentiation potentially affects the statistical properties of a rating scale measure and the relationship with other variables: Whereas correlations among items of the same rating scale measure can be spuriously inflated and thus, are likely to be overestimated, correlations with other variables are likely to be underestimated because of low variations in rating scale measures (McCarty & Shrum, 2000).

5.2.3 *Acquiescence*

Acquiescence (or acquiescence response bias) is another form of systematic response behavior and is generally characterized by a respondent's tendency to agree rather than disagree with rating scale items, irrespective of item content (Krosnick, 1999; Paulhus, 1991). Attitude and personality measures are often based on rating scale items that ask respondents to indicate their extent of agreement or disagreement. Thus, the examination of acquiescent response biases in rating scales is of particular importance "since it is intrinsically involved in the method of measurement itself" (Couch & Keniston, 1960, p. 151).

Within the theory of survey satisficing, acquiescence in self-administered Web surveys is considered the result of weak satisficing. In view of the fact that most respondents tend to start with the retrieval of arguments that are in favor of a stated item, premature termination of cognitive processing may result in an enhanced likelihood of affirmative responses because respondents already reached a satisfactory answer before considering arguments against the stated item (Krosnick, 1999). This assumption is supported by Knowles and Condon (1999) who showed that while acquiescent, disacquiescent and appropriate respondents did not differ significantly in the response times spent on disagreement with an item, acquiescent respondents took significantly less time to agree with an item compared to the other two groups of respondents. By contrast, disacquiescent responding seems to be as cognitively demanding as appropriate responding (Knowles & Condon, 1999). These findings on shorter response times with acquiescent responding give evidence of reduced cognitive effort due to a premature termination of the question-answer process. Moreover, acquiescence can also be considered a kind of strong satisficing since respondents may simply rely on the "social convention to be polite" rather than referring to internal cues relevant to the issue in question (Krosnick, 1999, p. 554). This is the case if respondents are not able or not willing to spend the cognitive effort to pass through all four stages of the question-answer process but instead simply agree with each of the statements, possibly even without having read it (Krosnick, 1991, 1999). Thus, shorter response times with acquiescent responding may also indicate that respondents skip the stages of information retrieval and judgment or do not even start the cognitive processing.

In most previous research, acquiescent responding is considered a systematic response tendency that is relatively stable over time because of its correlation with personal characteristics and personality traits (Couch & Keniston, 1960; Jackson & Pacine, 1961; Paulhus, 1991; Schuman & Presser, 1996, p. 204), whereas others supposed that acquiescence is best understood as a joint function of various factors related to the respondent and the method (Cronbach, 1946; Jackson, 1967; Peabody, 1961). As already proved by early studies, the likelihood of acquiescent responding is mainly a function of respondent characteristics and characteristics related to the content of rating scale items. For example, respondents with low cognitive ability and low cognitive sophistication are more prone to acquiescent responding (Jackson, 1967; Narayan & Krosnick, 1996), just as items that are ambiguous or vague in content (Cronbach, 1946; Jackson, 1967; Peabody, 1961), or items of little or no personal relevance for the respondent (Cronbach, 1946; Peabody, 1961). Baumgartner and Steenkamp (2001, 2006) provide a comprehensive overview of the theoretical explanations of acquiescent responding which are mainly related to characteristics of the respondent. Form-related characteristics of rating scales in terms of whether rating scale items are presented in a single-item-per-screen or in a multiple-item-per-screen design such as a grid format have not yet been systematically examined in previous studies. Among the studies reported above, only Peytchev (2006) mentioned parenthetically that no differences in the extent of acquiescent responding were found as a function of different rating scale formats (one grid with 8 items, two grids with 4 items each, and 8 screens with one item each).

In unbalanced rating scales, a high total item score indicates either true endorsement of the item content or merely a respondent's tendency to acquiescent responding. In case of acquiescent responding, however, the observed means are inflated (or deflated) irrespective of the item content. Even if the content of rating scale items is heterogeneous and largely uncorrelated, acquiescent responding may spuriously inflate the internal consistency of rating scale measures. Therefore, the use of balanced scales is the best practice to identify and control acquiescence. A respondent who agrees with an original item may not necessarily disagree with its reverse equivalent, which in turn increases the risk of inconsistent responses to item pairs. Thus, a balanced scale is suited for an implicit, rather than explicit measure of acquiescence with an integrated control because a high total item score requires true endorsement of item content rather than simple agreement

to each item irrespective of its content. Hence, balanced scales prevent a cumulative effect of acquiescence on total item scores which otherwise occurs in unbalanced scales (Baumgartner & Steenkamp, 2001; Jackson, 1967; Knowles & Condon, 1999; Messick, 1967; Paulhus, 1991; Van Vaerenbergh & Thomas, 2012). Although the use of balanced scales prevents confounding of acquiescent responding with the total item score, the negative effect of acquiescence on accurate measurement is not eliminated (Jackson, 1967; Knowles & Condon, 1999). This is aggravated by the fact that in balanced scales, the simple agreement to each item irrespective of its content is likely to result in reversed-item biases. Thus, although the underlying mechanisms differ, both the respondents' tendency to acquiescent responding and their susceptibility to careless responding result in an increased risk of inconsistencies between the original and reversed rating scale items (Paulhus, 1991; Weijters, et al., 2013).

5.2.4 *Extremity*

Extremity (or extremity response bias) is described as a tendency to choose the most extreme response options of a rating scale while excluding intermediate response options, irrespective of the content intended to be measured by the item (Greenleaf, 1992; Paulhus, 1991). Extreme responding is not explicitly included among the systematic response tendencies explained by the satisficing theory. However, in accordance with the assumptions of the satisficing theory, previous findings indicate that extreme responding increases with decreasing cognitive abilities (Greenleaf, 1992; Kaminska, et al., 2010; Meisenberg & Williams, 2008; Shulman, 1973). Closely related to this, respondents with less differentiated cognitive structures are considered more prone to extreme responding (Shulman, 1973). Baumgartner and Steenkamp (2001, 2006) provide a comprehensive overview of the theoretical explanations of extreme responding which are mainly related to respondent characteristics. Although previous research showed that extreme responding is strongly determined by respondent-related characteristics and thus, highly stable over time, it may also be affected by situational method-related characteristics (Baumgartner & Steenkamp, 2001; Paulhus, 1991). One reason for extreme responding ascribed to the content of a survey question is that respondents may think in extreme positive and negative categories and select the most positive or most negative response option in order to cope with the

ambiguity and complexity of the response task (De Jong, Steenkamp, Fox, & Baumgartner, 2008; Shulman, 1973). Thus, the likelihood of extreme responding is affected by the task difficulty and can be considered another systematic response tendency which is applied to keep the cognitive effort within manageable limits. Just as for acquiescent responding, however, form-related characteristics of rating scales in terms of whether rating scale items are presented in a single-item-per-screen or a multiple-item-per-screen design such as a grid format, have not yet been systematically examined with respect to the likelihood of extreme responding. None of the studies reported above gives any indication concerning this matter.

As is generally the case with systematic response tendencies, there is a clarification needed whether extreme responses actually reflect a respondent's true opinion or, rather his or her tendency to use the extremes of a rating scale. In order to increase the accuracy of extremity assessment, rating scale items should generally be heterogeneous in content with low inter-item correlations. This decreases the risk that extreme responses are mistakenly attributed to a systematic response tendency instead of reflecting a respondent's actual mental representation. Furthermore, item means should be close to the midpoint of the scale in order to not confuse extreme responses with strong general opinions (Baumgartner & Steenkamp, 2006; De Jong, et al., 2008). And lastly, the items should have approximately equal extreme response proportions to increase reliability of extremity assessment (Greenleaf, 1992). Similar to nondifferentiation and acquiescence, extreme responding also results in ratings that are limited to only a small number of potential response options. Moreover, extreme responding causes spurious inflation of correlations among otherwise unrelated items (Paulhus, 1991).

5.2.5 Primacy Effects

Primacy effects in rating scales describe a respondent's tendency to select response options at the positive (or negative) end of a rating scale more frequently when the response options are arranged in a positive-to-negative (or negative-to-positive) response option order (Krosnick, 1991; Krosnick & Presser, 2010). This systematic response tendency comprises both a systematic shift towards the left side of a horizontally arranged rating scale ('left-side bias') as well as towards the top end of a vertically arranged rating scale ('top-end bias').

According to the satisficing theory, primacy effects are due to respondents considering themselves satisfied with an acceptable instead of an optimal answer in order to minimize their cognitive effort. This in turn finds its expression in the selection of the first acceptable response option, while the remaining response options are considered to a lesser extent or they are even ignored by the respondents (Krosnick, 1999; Krosnick & Alwin, 1987; Malhotra, 2009; Tourangeau, 1984). In this regard, respondent fatigue is deemed to make a substantial contribution to the occurrence of primacy effects (Krosnick & Presser, 2010). Empirical evidence for this satisficing explanation of primacy effects in Web surveys was provided by Galesic and colleagues (2008) who identified—on the basis of eye tracking data—two response patterns being responsible for primacy effects. First, respondents actually spent more time on processing the response options listed first, compared to response options placed near the end of a list of several unordered response options. Second, some respondents even skipped latter response options completely (Galesic, et al., 2008). Thus, in line with the satisficing theory, the occurrence of primacy effects is attributable to the fact that early response options are processed more deeply compared to response options presented towards the end of a list. In some instances, response options listed later are even completely ignored as respondents stop cognitive processing once they have come to a satisfying answer (Galesic, et al., 2008; Krosnick, 1991).

According to the satisficing theory, primacy effects are regarded as a kind of weak satisficing. Respondents are tired of “carefully assessing the appropriateness of each of the offered response alternatives before selecting one” (Krosnick & Presser, 2010, p. 278). Satisficing explanations of primacy effects are supported by the enhanced prevalence of primacy effects among respondents with lower cognitive abilities. In general, it is assumed that providing an optimal answer implies accelerated effort for respondents with lower cognitive abilities. Therefore, these respondents are more likely to take cognitive shortcuts (Krosnick, 1991; Krosnick & Alwin, 1987). In fact, primacy effects are more common among respondents with a lower level of education and, by implication, with lower cognitive abilities compared to respondents with higher cognitive abilities (Krosnick, 1999; Krosnick & Alwin, 1987; Malhotra, 2009). According to the satisficing theory, increasing task difficulty should have a reinforcing effect on the occurrence of primacy effects. Difficult response tasks imply increased cognitive effort, which in

turn may encourage respondents to apply some kind of cognitive shortcuts in order to reduce this effort (Krosnick, 1991). By contrast, as already discussed in section 4.2.2, Malhotra (2009) argued that “people are more motivated to complete tasks when they are intricate, challenging, and enriching” (p. 182). Consequently, the simplicity of a response task may further promote the respondents’ boredom and fatigue, resulting in less attention towards the response task, whereas the complexity of a response task may encourage the attentive and careful processing of the rating scale items and their response options because respondents are likely to invest more effort, simply because they have to (Malhotra, 2009). Accordingly, Malhotra (2009) showed that complex ranking tasks were generally unaffected by primacy effects, whereas simple rating tasks resulted in significant primacy effects, particularly among respondents with low education. These findings contradict the assumption of a monotonic increase in the risk of cognitive shortcuts with higher task difficulty. Quite the contrary, “task difficulty may in fact encourage closer attention to the questionnaire because respondents are forced to read every response option” (Malhotra, 2009, p. 194).

The prevalence of primacy effects in terms of a left-side or top-end bias occurring in rating scales with ordered response options is related to the fact that respondents may infer the dimension of evaluation (e.g., agreement, importance, satisfaction) and deduce the remaining response options already after reading or hearing the first few ones (Malhotra, 2009). In addition to this, Gräf (2002) argued within the context of Web surveys that when respondents use the mouse cursor to navigate and enter their responses, “respondents might possibly choose those answers that are easiest to reach with the mouse” (Gräf, 2002, p. 62). Thus, in order to minimize their navigational effort, respondents provide their response as soon as they have reached a reasonable response option with their mouse cursor. This explains a respondent’s tendency to select one of the first response options when answering a list of rating scale items, both in terms of a left-side bias with horizontally arranged response options and a top-end bias with vertically arranged response options. To stick with the language of the satisficing theory, a respondent’s preference for nearest response options opted for solely in order to reduce the extent of navigational effort can be considered a kind of strong satisficing. Presumably, this explanation is more appropriate to explain the occurrence of primacy effects in rating scales arranged in a grid format.

As already described in section 4.3.2, based on previous findings it remains unclear whether primacy effects in rating scales actually exist or not. Previous studies that systematically examined the occurrence of primacy effects in multiple-item-per-screen designs such as grid formats versus single-item-per-screen designs by varying the categorial arrangement of a rating scale are not known to date. Previous findings concerning the effect of the categorial arrangement of rating scale response options on other systematic response tendencies such as acquiescence, extremity, or nondifferentiation are extremely rare. Only three studies reported by Keusch (2012) address this issue. Concerning the extent of (dis)acquiescent responding, the findings consistently indicated a significantly higher acquiescence response bias in the positive-to-negative, and a higher disacquiescence response bias in the negative-to-positive response option order. Findings on nondifferentiation were inconclusive: In all three studies conducted among panel members, students, and small business owners, the degree of scale differentiation was higher in the negative-to-positive response option order; however, the differences only reached statistical significance among panel members. The results on extreme responding showed more extremity in the positive-to-negative response option order, whereby solely in the student sample, differences reached statistical significance (Keusch, 2012).

In summary, when complying with the conventional categorial positive-to-negative order of rating scales, agreement with rating scale items may be the result of respondents being more inclined to select one of the first response options. Rating scales with a positive-to-negative response option order may tempt respondents to rush through a set of items at a faster pace, which, at the same time, enhances the respondents' reluctance to process all response options consciously. A respondent's tendency to prefer the response alternatives that are listed first is considered an indication of respondent fatigue and therefore, an indication of increased respondent burden and reduced respondent motivation. Hence, more respondent agreement in rating scales can be due to primacy effects as another kind of systematic response tendency which otherwise would have been regarded as acquiescence (Schuman & Presser, 1996, p. 74). In (counter)balanced rating scales, primacy effects can also be drawn on to explain the occurrence of reversed-item biases in terms of response inconsistencies between original and reversed items (Schuman & Presser, 1996, p. 74). Furthermore, mean responses are biased towards the response options listed leftmost in a horizontally arranged rating

scale and topmost in a vertically arranged rating scale. Thus, although counterbalancing the order in which response options are presented is one possible approach to handle primacy effects, a more effective approach would be to address the problem at source by reducing the risk of cognitive shortcuts in rating scales (Krosnick & Presser, 2010). In order to minimize the risk of systematic response tendencies such as primacy effects, respondents have to be motivated to spend extra time and effort in carefully processing and answering survey questions.

5.2.6 Summary

Although previous findings regarding the occurrence of systematic response tendencies in rating scales are not always clear, they are nevertheless indicative of the respondents' higher susceptibility to cognitive shortcuts in the processing of rating scales when the items are presented in a grid format as compared to a single-item-per-screen design. According to previous findings, it is quite conceivable that similarity and proximity of rating scale items in grid formats tempt respondents to process the grouped items only superficially, while the respondents are at a higher risk to miss fine distinctions that exist between them (Alwin, 2010; Callegaro, Shand-Lubbers, et al., 2009; Toepoel, et al., 2008; Tourangeau, et al., 2004). Thus, instead of attentively and carefully processing each single rating scale item and constantly reassessing the appropriateness of a response option, respondents may rather rely on the proximity of the items to infer the meaning of the current item from the preceding ones and simply adjust their answers to them. However, when merely relying on the context of rating scales, while disregarding item specifics or even completely ignoring item content, data accuracy is likely to be compromised.

Systematic response tendencies compromise data accuracy in terms of artificially inflating the internal consistency of rating scales when no reversed items are included (Barnette, 2000; Weijters, et al., 2009). This applies in particular to nondifferentiation, acquiescence, and primacy effects, and in the case of a merely one-sided preference for extreme response options, also for extremity. Thus, high internal consistency in unbalanced rating scales "may simply signal mindless and mechanical repetition of responses to items that are minor and redundant variations of the same basic question" or, at least, perceived as such by the respondents (Weijters & Baumgartner, 2012, p. 737).

In (counter)balanced rating scales, however, the respondents' reliance on systematic response tendencies is likely to have adverse effects on the internal consistency and factor structures of rating scale measures because of the higher risk of inconsistent responses to reversed rating scale items in terms of a reversed-item bias (Barnette, 2000; Harrison & McLaughlin, 1993; Weijters, et al., 2009).

Although various systematic response tendencies can lead to similar outcomes, their underlying causes are different. Careless responding in terms of respondent inattention or carelessness towards content-related details of rating scale items resulting in reversed-item biases in balanced rating scales is related to cognitive shortcuts within the first stage of the question-answer process: Instead of attentively and carefully processing the meaning of each single item or reading the items at all, respondents may rather rely on the proximity of the items to infer the meaning of the current item from surrounding ones (Podsakoff, et al., 2003; Tourangeau & Rasinski, 1988; Weijters, et al., 2013). By contrast, nondifferentiated responding is primarily attributable to cognitive shortcuts within the second and third stage of the question-answer process: Even if respondents pay sufficient attention to the content of single rating scale items, their answers will be insufficiently differentiated because information retrieval and judgment is executed only superficially (Herzog & Bachman, 1981; Krosnick, 1991). Similarly, primacy effects refer to cognitive shortcuts within the second and third stage of the question-answer process since the respondents are satisfied with the first available response option, resulting in a systematic shift towards the left side of a horizontally and towards the top end of a vertically arranged rating scale (Galesic, et al., 2008; Krosnick, 1991). Focusing on Web surveys, nondifferentiation and primacy effects can also be explained by a respondent's susceptibility to simply select the response option which is the easiest to be reached with the mouse cursor (Gräf, 2002). By contrast, acquiescent and extreme responding as the most frequently examined systematic response tendencies in rating scale measures are ascribed to cognitive shortcuts within the fourth stage of the question-answer process (Podsakoff, et al., 2003; Shulruf, Hattie, & Dixon, 2008; Weijters, et al., 2013). Acquiescent responding is mostly considered the result of a respondent's "evaluation apprehension" (Podsakoff, et al., 2003, p. 888) and his or her effort to comply with the "social convention to be polite" (Krosnick, 1999, p. 554). Extreme responding occurs once a respondent

reaches an answer and needs to fit this answer into the predefined response options. In this regard, a respondent's tendency to use merely the endpoints of a rating scale is deemed to reflect the individual differences in the interpretation and use of the rating scale. Thus, extreme responding is inherent in a rating scale per se and varies, among others, depending on the number of response options, rather than depending on the visual presentation of these response options (Paulhus, 1991; Shulman, 1973). Hence, acquiescence and extremity are considered the result of mapping the judgments onto the predefined response options per se, rather than a matter of variations in the visual presentation of the response options.

5.3 Item Missing Data

The extent of item missing data in rating scale measures is a more indirect indicator of data accuracy since an increased risk of item missing data in grid questions, compared to other question formats, may reflect an increased actual and perceived respondent burden and decreased respondent motivation to attentively and carefully process the set of rating scale items (Dillman, et al., 2009, p. 179; Grandmont, et al., 2010; Knapp & Heidingsfelder, 1999; Peytchev, 2009; Toepoel, et al., 2009b). In this regard, item missing data may be indicative of a respondent's difficulty in processing a survey question, with difficulties potentially occurring at one or more stages of the question-answer process (Beatty & Herrmann, 2002; De Leeuw, et al., 2003). Moreover, the mere visual appearance of a grid question and its visual complexity may prevent respondents from even starting the question-answer processing, resulting in item nonresponse, or in its extreme form, in survey breakoff (Beatty & Herrmann, 2002; Couper, et al., 2013; Peytchev, 2006).

Toepoel and colleagues (2009b), for example, found that the risk of item nonresponse, in terms of an increased average number of missing items per respondent, significantly increased as the number of rating scale items presented in a grid format was increased as well. They also found an increased probability that a substantive answer to at least one item was missing. One plausible explanation for the latter finding may be that respondents accidentally miss an item because of their confusion about the appropriate correlation between rows and columns and the lack of a prescribed order in which the items in a grid question should be processed.

Following this reasoning, increased item nonresponse rates in grid questions are predominantly caused by the unintentional skipping of items (Dillman, et al., 2009, p. 179; Gräf, 2002). In most instances, however, item nonresponse is considered a conscious decision that is related to the reduced respondent ability and/or reduced respondent motivation to provide a substantive answer (Beatty & Herrmann, 2002). Thus, Toepoel and colleagues' (2009b) findings of increased item nonresponse rates in larger grid questions can be explained by intentional rather than unintentional skipping of items since a higher respondent burden in larger grid questions is likely to induce a respondent's reluctance to answer each single item. By contrast, Callegaro and colleagues (2009) found no significant differences in item nonresponse rates between a grid format and a single-item-per-screen format, whereby item nonresponse rates were very low in either case. Most other studies outlined above did not report item nonresponse rates or included nonsubstantive responses (e.g., an explicit 'don't know' option) in the calculation of their item nonresponse rates which, however, does not correspond with the present definition of item nonresponse.

In general, grid questions are likely to produce higher rates of survey breakoff compared to other question formats. This observation, again, is explained by the higher level of perceived and actual respondent burden induced by the grid questions (Knapp & Heidingsfelder, 1999; Peytchev, 2009). Previous studies systematically examining differences between grid questions as a multiple-item-per-screen format versus a single-item-per-screen format, rarely reported differences in survey breakoff rates. Only two of the studies reviewed above provided information on survey breakoff. Grandmont and colleagues (2010) found a significantly higher breakoff rate when a set of 12 rating scale items was presented in a grid format rather than in a series on one screen or in a single-item-per-screen design. Peytchev (2006), however, found no significant differences in survey break off rates between the three experimental conditions comprising a grid with 8 items, two grids with 4 items each, and 8 screens with one item each.

Besides single-item-per-screen and multiple-item-per-screen designs using radio buttons, new rating scale formats such as visual analogue scales necessitating specific technical requirements and involving new data input methods, entail a higher risk of item nonresponse and survey breakoff compared to conventional radio button scales (see section 6.3). This increase in item missing data can also be explained by a higher level of task difficulty

and increased respondent burden for at least less computer literate respondents in a Web survey (Couper, et al., 2006; Funke, et al., 2011; Peytchev, 2009).

5.4 Response Times

Within the context of Web surveys, the assessment of the time respondents spent on answering a survey question is commonly used to gain deeper insights into the cognitive processes underlying survey responses and the processing of verbal and visual features of particular survey questions. In general, response times are considered a proxy of a respondent's difficulty in processing survey questions. Furthermore, response times are used as an indicator of a respondent's susceptibility to cognitive shortcuts within the question-answer process and therefore, an indicator of data accuracy (Bassili & Scott, 1996; Callegaro, et al., 2005; Christian, et al., 2009; Draisma & Dijkstra, 2004; Heerwegh, 2003; Husser & Fernandez, 2013; Stern, 2008; Stieger & Reips, 2010).

Range of Possible Uses

According to Yan and Tourangeau (2008), possible applications of response time measures in survey research can be grouped into the following three categories: (a) testing of theories, (b) pretesting of survey questions, and (c) investigation of Web survey methodologies. Within the first category of theory testing, response times are used as a proxy measure of the strength and stability of attitudes, whereas a lack of knowledge about the issue in question and the instability of attitudes are considered to result in longer response times (Bassili & Fletcher, 1991; Heerwegh, 2003). Response time measures also reflect the accessibility of attitudes about an issue, whereas shorter response times may be indicative of higher accessibility of relevant information (Draisma & Dijkstra, 2004; Tourangeau, et al., 1991).

In reference to the second category, response times are used for pretesting survey questions at the early stages of questionnaire design in order to identify problematic questions. The basic assumption is that the response time increases with the additional time required for solving problems arising within the question-answer process (Bassili & Scott, 1996; Stieger & Reips, 2010). For example, Bassili and Scott (1996) demonstrated that 'bad'

questions that were negatively worded and double-barreled in content took significantly longer to be answered than their 'good' equivalents. Following Draisma and Dijkstra (2004), the correlation between the strength, stability, and accessibility of attitudes as well as the wording of survey questions on the one hand and response times on the other, is basically attributable to the correlation between the question difficulty, cognitive processing, and response time. The less intense, stable, or accessible an attitude is, or the less comprehensible a survey question, the more cognitive processing is required to arrive at an answer and the more time it takes a respondent to answer a survey question. Simply put, "a difficult question needs more processing and hence results in a long response latency" (Draisma & Dijkstra, 2004, p. 132). Thus, response time is considered an indicator of task difficulty and the cognitive burden primarily imposed by difficult question content (Draisma & Dijkstra, 2004; Lenzner, et al., 2010; Stieger & Reips, 2010).

However, the investigation of Web survey methodologies as the third category of Yan and Tourangeau's (2008) classification is considered more important in the present context of evaluating different rating scale designs in terms of various indicators of data accuracy. Basically, this category refers to the examination of effects of different features of questionnaire design and administration on response times. In this regard, Yan and Tourangeau (2008) revealed various respondent-related and method-related characteristics altering response times in Web surveys. For instance, response times increased with additional and fully labeled response options, and were longer for related survey questions that were presented in isolation compared to those shown in close proximity. Furthermore, response times decreased for survey questions presented later within a questionnaire. These effects persisted even after controlling for other method-related and respondent-related characteristics (i.e., age, education, prior survey experience, Internet experience) (Yan & Tourangeau, 2008).

According to these findings, Callegaro and colleagues (2005) mentioned three reasons that explain a decrease in response times over the course of the completion of 133 Likert-type items with a 5-point response scale horizontally arranged on separate Web screens: First, as a result of using the same response scale throughout the questionnaire, respondents may become increasingly experienced in handling it. Thus, learning effects can help increase the speed of responding. Second, in consequence of answering a set of items related in content, question context may increasingly be enriched.

Hence, facilitation effects can fasten a respondent's cognitive processing. Third, answering a large number of items that have a similar content and are based on the same response options may tempt respondents to speed up the scale completion in order to come to an end at a quicker pace. Consequently, fatigue effects are likely to be responsible for shorter response times (Callegaro, et al., 2005).

Regarding the first two explanations of Callegaro and colleagues (2005), response times provide an indication of the respondents' difficulties becoming manifest on one or more stages of the question-answer process, with difficult response tasks necessitating more cognitive and/or navigational processing and consequently, more time to be completed. Hence, increases in response times may serve as an indicator of respondent's difficulties in interpreting and processing the verbal and visual features of survey questions (Christian, et al., 2009; Toepoel, et al., 2008; Tourangeau, et al., 2004). Conversely, presenting rating scale items together on a single screen, for instance, may facilitate the cognitive processing and navigation whereby survey responding is accelerated (see also section 4.3.1) (Callegaro, et al., 2005; Couper, et al., 2001; Knowles, 1988; Toepoel, et al., 2009b; Tourangeau, et al., 2004). Or, respondents may be faster in responding when the design and administration of survey questions is consistent with their general expectations and prior experiences with surveys, e.g., when rating scales are arranged in a positive-to-negative response option order (see also section 4.3.2) (Christian, et al., 2009; Tourangeau, et al., 2004). And, some data input methods (e.g., radio buttons) are considered easier to use, and thus, less time-consuming than others (e.g., slider scales, drop-down menus) (see section 6.3) (Couper, et al., 2006; Funke, et al., 2011; Healey, 2007; Heerwegh & Lossveldt, 2002).

While increases in response times may serve as an indicator of a respondent's difficulties in cognitive and/or navigational processing in Web surveys, response times may also be "a rough measure of the overall amount of respondent effort and attention spent on completing the Web questionnaire" (Tourangeau, et al., 2009, p. 316). Hence, according to Callegaro and colleagues' (2005) third explanation, comparatively short response times may reflect inattentive or careless processing and increased respondent's susceptibility to rely on systematic response tendencies (Callegaro, et al., 2005; Funke & Reips, 2012; Salzberger & Koller, 2013; Smyth, et al., 2006; Stieger & Reips, 2010; Toepoel, et al., 2008; Tourangeau,

et al., 2009; Tourangeau, et al., 1991). Based on the assumption that high response accuracy requires sufficient consideration of a survey question, Salzberger and Koller (2013) showed that instructions to answer a grid question comprising 12 rating scale items spontaneously, as compared to answer them well-considered, actually reduced response times, which in turn was associated with a worse measurement of the latent construct and a higher risk of respondents relying on systematic response tendencies such as nondifferentiation and extremity (Salzberger & Koller, 2013). Furthermore, Stieger and Reips (2010) showed that short response times may indicate cognitive shortcuts in terms of click-through behaviors which are thought of as “answering without really reading the questions” (p. 1492): 46% of the respondents showed a click-through behavior at least once during the course of survey completion, whereas over 65% of these click-through behaviors occurred in semantic differential scales presented in a grid-like format (Stieger & Reips, 2010). This suggests that grid formats are at a particularly high risk of inducing cognitive shortcuts in terms of click-through behaviors without considering the content of rating scale items with reasonable attentiveness and carefulness.

Thus, response times can either reflect the respondents’ difficulties in cognitive processing, or they are indicative of a respondent’s inattentiveness and carelessness towards the response task. In this regard, response times are also affected by specific characteristics of the respondent, including the respondent’s cognitive capacity in terms of age and education, computer and Internet skills, and prior survey experience. Everything else constant, Yan and Tourangeau (2008) found that older respondents and respondents without a high school degree needed a longer time to answer the questions of a Web survey, whereas respondents with certain expertise in Internet use and Web survey completion were faster in responding. In terms of the satisficing theory, Toepoel and colleagues (2008) found that trained respondents with certain experience in participating in Web surveys were more likely to “go through the questions faster if they recognize the question structure and layout” or, to put it more drastically, “may ‘speed’ through the survey to reduce the burden of their task” (p. 988), resulting in significantly shorter times for survey completion. Thus, as already mentioned in section 2.1.2, respondents with certain computer and Internet knowledge and prior survey experience may need less time either because they are more practiced or because they are less careful and attentive towards the response task. By

contrast, however, Coen, Lorch and Piekarski (2005) found no evidence that experienced respondents tend to speed through a survey.

To summarize, response times comprise the time spent on different stages of the question-answer process, namely: time for reading and comprehending a question, time for retrieving information and judging, and time for formatting a response. Apart from that, differences in response times suggest two different interpretations: On the one hand, they may reflect difficulties in interpreting and processing the verbal and visual features of survey questions and on the other hand, they may indicate the amount of effort spent on interpreting and processing the verbal and visual features of survey questions. Thus, it remains still unclear whether increases in response times are indicative of increased or decreased data accuracy, which is why response times are best interpreted in conjunction with further indicators of data accuracy.

Aspects of Technical Implementation

Web surveys present the advantage that paradata can be gathered automatically during the data collection process. In one of the questionnaire design guidelines for Web surveys, Dillman and colleagues (2009) recommend the collection of paradata whenever it is possible because paradata “provide[s] feedback on how the respondent interacts with the questionnaire” (p. 216). In general terms, paradata can be used to describe and evaluate the respondent’s question-answer processing. More specifically, the analysis of paradata may help gain a better understanding of the effects of verbal and visual questionnaire features, i.e., their effects on the respondents’ handling of potential ambiguities or difficulties in comprehending, and the way of processing relevant question components and reaching an answer (Christian, et al., 2009; Heerwegh, 2003; Stern, 2008). In general, a distinction is made between server-side and client-side paradata. Whereas server-side paradata is collected at screen level by simply recording visits to a Web page, client-side paradata is captured within a particular Web page enabling the logging of every user’s action (Heerwegh, 2003). Correspondingly, response times as one type of paradata can be gathered either server-sided or client-sided.

Server-side response times. Common software programs for the implementation of Web surveys enable the gathering of response times as standard by means of server-side time stamps. By capturing the time span between loading a target page and submitting the page to the server by selecting a 'Continue' button, server-side response times also include the time for downloading the content of a Web page and transmitting the respondents' answers to the server. Thus, server-side response time comprises more than just the respondents' time for performing the respective stages of the question-answer process (Olson & Parkhurst, 2013). Furthermore, the detailedness of response time information is reduced with each additional survey question on a single screen. Consequently, if there are two or more survey questions on a single screen, the information content of server-side paradata is strikingly decreased as compared to client-side paradata because no specifications for individual questions can be obtained (Heerwegh, 2003).

Client-side response times. Recordings of mouse events such as mousedown, mouseup, mouse click, and mousemove as a sort of client-side paradata using JavaScript, represent a more detailed measurement of response times because each single respondent's action within a Web page is recorded separately (Heerwegh, 2003; Olson & Parkhurst, 2013). This is particularly relevant for the analysis of rating scales that are commonly arranged in a multiple-item-per-screen format. As distinct from server-side response times, client-side mouse event recordings enable the survey researcher to analyze the time respondents spent on each single action. Moreover, the order in which respondents select their answers and the number of changes they make before submitting the final answer can also be assessed, allowing a better understanding of how respondents reach their answers to survey questions (Christian, et al., 2009; Heerwegh, 2003; Stern, 2008).

In this chapter, it has become apparent that rating scales in general and grid questions in particular often suffer from a respondent's inattentive and careless processing of rating scale items which is likely to result in impaired data accuracy. Impaired data accuracy can, among others, be reflected in systematic response tendencies, item missing data, or click-through behaviors. A thorough assessment of data accuracy in rating scales is important to ensure high total survey quality at the end. But even more important for achieving high total survey quality is to increase data accuracy from the outset by

preventing the respondents' inattentiveness and carelessness in processing a rating scale. In the next sections, factors related to questionnaire design and questionnaire administration are presented which help improve data accuracy.

6. IMPROVING DATA ACCURACY IN RATING SCALES

The likelihood of cognitive shortcuts in survey responding is thought to be affected by the interaction between a respondent's ability and motivation to optimally perform the response task, which in turn is dependent on the inherent difficulty of a response task (Krosnick, 1999). Whereas respondent ability cannot be influenced by the survey researcher, respondent motivation and task difficulty can be affected by a number of factors related to questionnaire design and questionnaire administration which are under direct control of the survey researcher. Aside from reducing the number of rating scale items or implementing balanced scales, advanced dynamic features in terms of dynamic input controls can be used in order to positively impact respondent motivation and task difficulty and thus, to improve data accuracy in rating scales.

6.1 Length of Rating Scales

Since increased respondent burden is likely to enhance the risk of cognitive shortcuts, the respondent effort required to answer a survey question should be limited. Thus, in order to restrict perceived and actual respondent burden, the number of rating scale items must not exceed a certain maximal amount of items (Andrews, 1984; Dillman, et al., 2009, p. 179; Toepoel, et al., 2008). Likewise, grid questions may induce an increased sensorimotor effort because they require respondents to combine a fairly large amount of information from rows and columns. This effort is likely to increase even further the more items are presented simultaneously in a single grid (Couper, et al., 2013; Dillman, et al., 2009, p. 179). This will be particularly true if scrolling is necessary (Couper, et al., 2013; Toepoel, et al., 2009b). Furthermore, an increasing number of rating scale items is expected to induce respondent fatigue and foster a respondent's inattentiveness and carelessness towards the response task (Drolet & Morrison, 2001; Hinkin, 1995). This in turn may tempt respondents to take various cognitive shortcuts, resulting in merely satisfying rather than optimal answers. For instance, "faced with more items, respondents distinguish less between them, and the influence of an earlier

item on a later item appears greater” (Drolet & Morrison, 2001, p. 200). Or, the risk of item nonresponse may increase with larger numbers of rating scale items presented in a single grid question (Toepoel, et al., 2009b).

In sum, the efficiency gains of grid questions related to the facilitation of processing and acceleration of completion may turn to the contrary, in terms of higher respondent burden and an increased risk of respondents relying on cognitive shortcuts, if grid questions consist of more than just a few items (Andrews, 1984; Couper, et al., 2013; Dillman, et al., 2009, p. 179). Although there is no binding recommendation on the optimal number of rating scale items in a grid, the number of items should not exceed a certain limit in order to prevent increased respondent burden and potential compromises in data accuracy (Andrews, 1984; Dillman, et al., 2009, p. 179; Toepoel, et al., 2008). This general statement is supported by previous findings on the respondents’ evaluation of a Web survey as a function of the number of items on a screen, even though the findings are admittedly mixed: Toepoel and colleagues (2009b), for example, found that the respondent evaluation of the survey layout significantly worsened with an increasing number of items per screen which is indicative of a respondent’s preference for a single-item-per-screen format. Similarly, Thorndike and colleagues (2009) revealed the respondents’ preference for a single-item-per-screen format even though it took longer to be completed. By contrast, Toepoel and colleagues (2008) found significantly better respondent evaluation of the survey layout with 5 items presented in a grid as compared to each item on a single screen or a grid with 10 items, indicating that respondents prefer grid formats that do not exceed a certain number of items. Callegaro and colleagues (2009) found no significant differences in perceived difficulty and enjoyment depending on the number of items per screen.

6.2 Balanced Rating Scales

The use of balanced rating scales consisting of an equal number of original and reverse worded items is useful in affecting data accuracy in two respects: First, the respondents’ carelessness towards individual item characteristics during survey completion can be attenuated, and second, the extent of respondents relying on systematic response tendencies can be assessed

(Barnette, 2000; Podsakoff, et al., 2003; Weijters & Baumgartner, 2012; Weijters, et al., 2009).

In the former case, reversed items follow the concept of “cognitive ‘speed bumps’ that require respondents to engage in more controlled, as opposed to automatic, cognitive processing” (Podsakoff, et al., 2003, p. 884). In this way, respondent boredom and uniform responding to a seemingly endless set of rating scale items worded in one direction shall be prevented (Harrison & McLaughlin, 1993; Weijters & Baumgartner, 2012). Hence, inattentive or careless respondents can be urged to bring more focus on individual item characteristics, whereby systematic responding may be averted (Podsakoff, et al., 2003; Weijters & Baumgartner, 2012).

In the latter case, systematic response tendencies such as acquiescence or nondifferentiation may compromise data accuracy because of the artificially inflated internal consistency of rating scales when no reversed items are included (Barnette, 2000; Weijters, et al., 2009). On the contrary, the risk of lowered correlations between the original and reversed items in balanced rating scales and hence, a decrease in the internal consistency of rating scale measures may also impair data accuracy (Barnette, 2000; Harrison & McLaughlin, 1993; Hinkin, 1995; Weijters, et al., 2009). Nevertheless, the examination of response inconsistencies gives some indication of the occurrence of systematic response tendencies, e.g., a respondent’s susceptibility to careless responding according to which he or she simply fails to notice the reverse wording of an item at the comprehension stage, or a respondent’s tendency to acquiescent responding emerging at the stage of reporting the answer (Weijters & Baumgartner, 2012).

When creating appropriate item reversals in balanced rating scales, a conceptual opposite is required rather than a mere logical reversal in terms of simply negating the assertion, e.g., by adding a ‘not’ or the negative half sentence ‘It is not the case that...’ (Paulhus, 1991; Peabody, 1961; Rorer & Goldberg, 1965). However, the problem still remains that for some items, there is no clear conceptual opposite, or reversed items have substantially changed in content. If the original item and its reverse equivalent are not completely contradictory, substantial correlation will be missing and double (dis)agreement with reversed item pairs will not necessarily indicate response inconsistencies that are due to a respondent’s susceptibility to cognitive shortcuts (Schuman & Presser, 1996, p. 206). Consequently, pairs of original and reverse worded items have to be selected carefully in order to ensure that

lacking correlations are actually due to careless or acquiescent responding rather than the result of poor item selection.

6.3 Interactive Rating Scales

In general, interactive Web surveys enrich conventional questionnaire design and administration with advanced dynamic features which “involve some kind of movement or change in what is displayed to the respondent” (Tourangeau, et al., 2013, p. 100). Dynamic features can either be responsive or non-responsive. In the former case, movements or changes are triggered by particular respondents’ actions, whereas in the latter case, each movement or change occurs irrespective of the respondents’ actions. Responsive dynamic features in particular, are expected to improve the respondents’ performance, and thus, improve data accuracy by offering immediate feedback to the respondents on their relevant actions (Conrad, Couper, Tourangeau, & Galesic, 2005; Tourangeau, et al., 2013, p. 101). More precisely, the implementation of responsive dynamic features can specifically be used in questionnaire design to draw the respondents’ attention to the relevant question components and render assistance to adequately complete the four stages of the question-answer process (Couper, et al., 2013; Kaczmirek, 2010; Tourangeau, et al., 2013, p. 99). Responsive dynamic features can further be used to enrich the respondent-survey interaction, whereby respondent involvement may be enhanced and respondent motivation is more likely to be maintained at a consistently high level over the course of survey completion (Couper, et al., 2013; Kaczmirek, 2010; Sikkel, et al., 2014; Stanley & Jenkins, 2007). Concerning the design of rating scales, increases in respondent involvement and respondent attention can be achieved by, among others, dynamic highlighting or dynamic input methods such as drag-and-drop.

Dynamic Highlighting

Responsive dynamic features in terms of visual feedback on item completion can be applied in order to reduce the risk of (unintentional) skipping of single items in a grid question. Basically, the concept of dynamic highlighting follows the principle of “providing feedback about what the user has done and what remains to be done” (Couper, et al., 2013, p. 14). In fact, visual

indication of item completion by dynamically graying out the entire row once an item was answered, significantly reduced item nonresponse rates as compared to a standard grid where no dynamic highlighting was applied (Couper, et al., 2013; Kaczmirek, 2010). Thus, shading in the post-selection phase may facilitate visual orientation and help identify those items a respondent has not yet addressed. By contrast, nondifferentiation was not affected by post-selection feedback (Kaczmirek, 2010). Hence, visually highlighting the response options that a respondent has chosen for previous items does not induce higher differentiation among a set of rating scale items.

At the same time, Kaczmirek (2010) also showed that visual feedback in grid questions may impede complete and accurate responses because visual guidance in terms of shading in the pre-selection phase—triggered by moving the mouse cursor over the response options—provoked more item nonresponse and more nondifferentiation than grid questions without any dynamic highlighting (Kaczmirek, 2010). Thus, pre-selection feedback compromised data accuracy rather than contributed to its improvement. Based on these findings, Kaczmirek (2010) concluded that shading in the pre-selection phase potentially distracts respondents from the actual response task. In this regard, respondents may invest less cognitive effort in the comprehension of an item and in the retrieval and judgment of an answer, which in turn results in less differentiated responses to rating scale measures.

Dynamic Dragging and Dropping

Beyond navigation by mouse clicking, Web surveys enable the implementation of more dynamic drag-and-drop techniques performed by mouse dragging. By the use of such responsive dynamic data input methods, the interactivity of a response task can be further enhanced. Slider scales as one special kind of the broader category of visual analogue scales can be considered a well-established question format using drag-and-drop: Respondents have to left-click on a slider, drag the slider to a desired position while holding the mouse button down, and then release the mouse button.

The basic idea behind using drag-and-drop techniques in Web surveys is to increase the respondent's involvement, or at least, counteract the respondent's fatigue by "making the [survey] a more active, lively process" (Sikkel, et al., 2014, p. 183). In fact, survey questions using drag-and-drop in many different forms are assessed more positively in terms of the respondent's self-rated interest, enjoyment, and overall satisfaction (Dolnicar,

Grün, & Yanamandram, 2013; Downes-Le Guin, Baker, Mechling, & Ruylea, 2012; Sikkel, et al., 2014; Sleep & Puleston, 2009; Stanley & Jenkins, 2007; Thomas, 2011; Tress, 2012). However, Sikkel and colleagues (2014) also reported that an initially positive evaluation of various drag-and-drop scales in the first wave of a panel survey turned into a more negative evaluation in the second wave as compared to a conventional radio button scale, which is why they recommended to use drag-and-drop scales only rarely. By contrast, Couper and colleagues (2006) found no significant differences in how interesting the survey was assessed depending on whether a slider bar, radio buttons or a text field was used for data input.

By counteracting respondent fatigue, the risk of the respondents' inattentiveness and carelessness towards the response task should be reduced, resulting in higher data accuracy. However, in most previous studies, no meaningful differences were found as related to response distributions, straight-lining, inconsistent responding to reversed item pairs, or extreme responding (Couper, et al., 2006; Downes-Le Guin, et al., 2012; Funke, et al., 2011; Thomas, 2011). Moreover, previous research indicated that the use of slider scales implemented through drag-and-drop resulted in longer response times and provoked more missing data in terms of item nonresponse and survey breakoff compared to conventional radio button scales. This increase in item missing data is assumed to be either an indication of a respondent's reluctance to use a slider scale as a result of the increased task difficulty and higher respondent burden, or an indication of technical problems arising with the drag-and-drop technique (Couper, et al., 2006; Funke, et al., 2011; Thomas, 2011). In line with the satisficing theory, Funke and colleagues (2011) also found that the increased risk of survey breakoff with slider scales particularly applies to respondents with less than average education. This finding supports the assumption that "certain formats are more challenging in terms of previous knowledge needed or cognitive load" (p. 227). Tourangeau and colleagues (2013) arrived at the rather sobering conclusion that "the current literature provides no compelling reason to use [slider scales implemented through drag-and-drop] other than it can be done" (p. 113).

Visual analogue scales can also be implemented with navigation by mouse clicking instead of mouse dragging. Respondents indicate their response by clicking on the desired position of an answer line with modifications being made by clicking again on any other position. By applying such visual analogue scales, Funke and Reips (2012) found neither a

difference in item nonresponse nor in response times between visual analog scales and radio button scales. However, the frequency of modified responses was significantly higher in the visual analog scales indicating that respondents actually made use of the finer gradations enabled by a visual analog scale. Furthermore, correlations between corresponding items were significantly higher with visual analog scales being regarded as a further indication of improved measurement. Based on these findings on visual analogue scales using mouse clicking versus mouse dragging, one possible recommendation concerns the straightforward implementation of survey questions to keep technological requirements for Web survey respondents as simple as possible. In order to prevent item missing data and impaired measurement properties of rating scales, “maneuvering should be kept at the simplest level” (Gräf, 2002, p. 62).

In sum, previous findings revealed that the use of advanced dynamic features for constructing survey questions may increase respondent satisfaction but otherwise, may also enhance task difficulty and respondent burden by imposing high cognitive and navigational requirements at least on some respondents. These findings clearly indicate that with the objective of implementing a respondent-friendly Web survey in order to promote an accurate completion of a questionnaire, it is of great importance to ensure an appropriate level of task difficulty which corresponds with respondent ability and keeps respondent motivation at a high level over the course of survey completion (Funke, et al., 2011; Jenkins & Dillman, 1997; Kaczmirek, 2010; Kieruj & Moors, 2013). Besides an appropriate degree of task difficulty concerning the basic understanding and handling of dynamic features, software and hardware compatibility has to be ensured to take full advantage of the potentials provided by advanced dynamic features (Dillman, et al., 2009, p. 213; Vicente & Reis, 2010). Otherwise, respondent frustration and confusion due to novel respondent tasks and possible software or hardware non-compatibility resulting, for example, in the incorrect displaying of screen contents, may distract respondents from their actual response task and interfere with their ability and motivation to provide complete and accurate responses (Funke, et al., 2011; Vicente & Reis, 2010). Based on a literature review on Web survey design and item missing data, Vicente and Reis (2010) similarly concluded that “the technical potential of the Internet offers survey researchers a wide range of possibilities for web surveys in terms of

questionnaire design”, while warning that “the abuse of technical facilities can detract respondents from cooperating rather than motivating them” (p. 251).

In the following section, new drag-and-drop rating procedures are introduced attempting to tap the full potential of visual enhancements and greater interactivity in the design and administration of survey questions provided in a Web survey environment, while taking the theoretical foundations of rating scale design and administration and lessons learned from previous implementations of interactive rating scales into account.

7. NEW DRAG-AND-DROP RATING SCALE DESIGNS

As outlined throughout the previous sections, optimal survey responding requires respondents to carefully process the four stages of the question-answer process. In Web surveys where information is commonly presented visually and no interviewer is present to guide and motivate the respondent, an adequate execution of the question-answer process additionally presupposes prior visual perception of and sufficient attention to the information presented. However, respondents frequently fall back on various kinds of cognitive shortcuts in order to minimize the cognitive and/or navigational effort necessary to answer survey questions which may affect data accuracy negatively (Jenkins & Dillman, 1997; Krosnick, 1991).

While in a single-item-per-screen design, a respondent's focus is on a single rating scale item, this question format may lack the necessary context facilitating a fluent question-answer processing and a comprehensive understanding of the item set as a whole (Couper, et al., 2001; Schwarz, Strack, & Mai, 1991). On the contrary, a fairly large amount of information presented simultaneously in a grid question and the necessity to combine information from rows and columns makes it more difficult for respondents to get a clear understanding of the respective item meanings and to perceive fine distinctions between them (Callegaro, Shand-Lubbers, et al., 2009; Toepoel, et al., 2008; Tourangeau, et al., 2004).

A rating scale format using drag-and-drop technique as a more interactive data input method than traditional radio button scales may potentially help prevent the shortcomings associated with conventional grid questions. With the aim of incorporating the advantages of a standard grid while overcoming its disadvantages, two drag-and-drop rating scales have been designed that allow for navigation by mouse dragging instead of mouse clicking: the 'drag-response' and 'drag-item' scale. While the context of a question necessary for a fluent question-answer processing of a set of rating scale items is still provided, both drag-and-drop rating scales aim at enhancing the focus on each single item and the respective response options in order to promote the attentive and careful processing of a rating scale.

The drag-response and drag-item scale are based on the same navigation technique. To answer the drag-and-drop rating scales, respondents have to left-click on a draggable element, drag this element to the desired position while holding the mouse button down before dropping the element in the field provided. However, the drag-and-drop rating scales differ regarding the question component which is draggable: In the drag-response scale, respondents are required to drag the response options, whereas in the drag-item scale, they have to drag the items. While dragging an element with the left mouse button held down, conventional scrolling bars cannot be used. Therefore, an auto-scrolling function during a drag-and-drop operation has been implemented, i.e., the display window is automatically slid upwards or downwards when the mouse cursor exceeds a minimum distance to the upper or lower edge of the screen while dragging the respective element. Hence, both drag-and-drop rating scales can be extended in their length with navigation being still ensured even on rather small browser window sizes.

In the following, the basic structure, the visual design, and the dynamic handling of the drag-and-drop rating scales are briefly described. Thereafter, implications of the structure, design, and handling of the drag-and-drop-scales introduced here are discussed with respect to the anticipated impact on cognitive and navigational processing of rating scales, especially with regard to the wider context of attentive and careful question-answer processing. Expectations are outlined concerning the effects of the drag-and-drop rating scales on the respondents' attentiveness and carefulness in cognitive and navigational processing. In this regard, it is explicated how the use of the drag-response and drag-item scale is expected to affect data accuracy in terms of altering the measures discussed above (see also Chapter 5): the prevention or evocation of systematic response tendencies, the occurrence of item missing data, and variations in response times.

7.1 Drag-Response Scale

The graphical layout of the drag-response scale (see Figure 2) basically resembles a conventional grid format: The general question request is followed by several rating scale items presented in the rows of a raster. Unlike traditional grid formats, however, the response options arranged in the top

line of the actual scale corpus are draggable. Thus, the difference to a standard grid is that instead of clicking a radio button, respondents are required to left-click on one of the draggable response options, and then drag a response option towards an item, before dropping the response option in the desired position. Respondents can use one selected response option for multiple different items, i.e., each response option can be selected as many times as there are items to be answered. In the drag-response scale, all items are presented simultaneously. As a result, question context is provided right from the beginning. Furthermore, the order in which the rating scale items are answered is left to the respondent and previous answers can be modified any number of times, even after a response option has been assigned to an item.

7.2 Drag-Item Scale

In the drag-item scale (see Figure 3), a general question request is followed by several rating scale items that are stacked on top of one another in the top line of the actual scale corpus. The items need to be dragged to the response options vertically arranged in rows by left-clicking on the draggable item on top of the stack, dragging one item at a time to the desired response option, and drop it in place. It follows that respondents have to move one selected item to its destination after another, whereby previous items need to be answered first before subsequent items can be dragged. Thus, on the one hand, a clear sequence of rating scale items is prescribed in the drag-item scale. On the other hand, question context is provided only gradually since the content of an item only becomes visible after the previous element has been taken from the stack. Once an item has been assigned to a response option, its position can be changed freely; the item can be moved to another response option, if the respondent wants to revise his or her answer. Also, an infinite number of items can be assigned to the very same response option. If respondents assign more than one item to the same response option, the items will not overlap, but will be displayed one underneath the other. Hence, this rating scale layout is in some way similar to a ranking task in terms of bringing the items into a kind of rank order but without enforcing distinctions between the items, when no difference actually exists from a respondent's point of view. Following the basic idea of a progress indicator used to prevent

premature termination of a survey, the number of the current item and the total number of items are indicated by small figures on the right-hand side of each item in order to keep the respondent informed about the current completion status and the number of items remaining.

Im Folgenden finden Sie eine Reihe von Aussagen, die sich auf Sie beziehen. Bitte geben Sie für jede Aussage an, inwieweit Sie der Aussage zustimmen oder diese ablehnen.

Um eine Antwort zu geben, klicken Sie mit der linken Maustaste auf das gewünschte Antwortkärtchen, ziehen Sie es mit gehaltener Maustaste auf das jeweilige blaue Antwortfeld und lassen Sie dann die Maustaste los.

	stimme voll und ganz zu	stimme eher zu	teils/teils	lehne eher ab	lehne voll und ganz ab
Auch in kritischen Situationen behalte ich einen kühlen Kopf.	<input type="text"/>				
Ich plane nicht gern, sondern lasse lieber alles auf mich zukommen.	<input type="text"/>		<input type="text"/>		
Ich kann besser denken, wenn ich ein leichtes Gefühl ängstlicher Spannung habe.	<input type="text"/>				
Ich halte es für wichtig, meine Zukunft vorzustrukturieren.	<input type="text"/>				
Wenn mich etwas in Spannung versetzt, kann ich weniger gut arbeiten als sonst.	<input type="text"/>				
Mir passiert es oft, dass ich in kritischen Situationen Fehler mache.	<input type="text"/>				

Figure 2: Graphical layout of the drag-response rating scale

Im Folgenden finden Sie eine Reihe von Aussagen, die sich auf Sie beziehen. Bitte geben Sie für jede Aussage an, inwieweit Sie der Aussage zustimmen oder diese ablehnen.

Um eine Antwort zu geben, klicken Sie mit der linken Maustaste auf das Fragekärtchen, ziehen Sie es mit gehaltener Maustaste auf das gewünschte blaue Antwortfeld und lassen Sie dann die Maustaste los. Mehrere Fragen können derselben Antwort zugewiesen werden.

Ich plane nicht gern, sondern lasse lieber alles auf mich zukommen.	2/6
stimme voll und ganz zu	<input type="text"/>
stimme eher zu	<input type="text"/>
teils/teils	<input type="text"/>
lehne eher ab	<input type="text"/>
Auch in kritischen Situationen behalte ich einen kühlen Kopf.	1/6
lehne voll und ganz ab	<input type="text"/>

Figure 3: Graphical layout of the drag-item rating scale

7.3 Expectations Regarding Data Accuracy

Both drag-and-drop rating scales introduced above are expected to have a positive impact on the respondents' attentiveness and carefulness when answering a set of rating scale items which may prevent systematic response tendencies typically occurring in rating scales in general, and grid questions in particular. As outlined above, systematic response tendencies may be prevented by: (a) counteracting respondent fatigue, (b) arousing attention and decelerating the speed of responding, and (c) by strengthening the link between items and response options.

Counteracting Respondent Fatigue

Respondents are likely to get tired and bored when being repeatedly confronted with rating scale items having similar content and the same response options (Alwin & Krosnick, 1985; Drolet & Morrison, 2001; Gräf, 2002; Knowles, et al., 1992). Besides the similarity of item content, the visual proximity of items in grid questions makes respondents even more susceptible to rely on the visual proximity of rating scale items rather than attentively and carefully processing each individual item (Alwin, 2010; Tourangeau, et al., 2004; Weijters, et al., 2013). By using dynamic drag-and-drop techniques in rating scales, respondent involvement may be maintained and fatigue effects can be counteracted (Downes-Le Guin, et al., 2012; Sikkel, et al., 2014). At questionnaire level, the use of drag-and-drop scales gives variety to otherwise rather uniform questionnaires and monotonous question formats. At question level, the repetitive nature of rating scale measures can be interrupted by means of dynamic drag-and-drop interactions which help break the monotony in the respondents' processing and responding.

While respondent involvement is a rather vague concept which has not been defined properly in previous studies, the way in which the drag-response scale and the drag-item scale are expected to counteract respondent fatigue, and at the same time encourage the respondents' attentiveness and carefulness towards the response task, is explained in greater detail further with regard to the cognitive and navigational requirements of the drag-response and drag-item scale.

Achieving Higher Attention & Decelerated Responding

Respondent attention to the various components of a survey question may vary in terms of the extent to which these components are spontaneously noticed and carefully attended to as a result of the limited capacity of a respondent's working memory and the fact that not every information is of equal interest to the respondent (Jenkins & Dillman, 1997; Kahneman, 1973, pp. 1ff, 52ff). In grid questions, this is aggravated by the fact that respondents tend to speed up and simply click through several rating scale items without sufficiently considering the content of the respective items, or reaffirming the meaning of the response options (Callegaro, et al., 2005; Stieger & Reips, 2010; Toepoel, et al., 2008).

By using the dynamic drag-and-drop scales, respondents may be encouraged to pay greater attention to the key components of a rating scale: Whereas the drag-item format clearly highlights the item content, thus bringing higher attention to it, the drag-response format primarily brings the response options into focus. Aside from the visual highlighting of the key components of a rating scale, the respondents' attention is continuously and repeatedly guided towards the respective draggable element, so that both the items, in case of the drag-item scale or the response options, in case of the drag-response scale are emphasized. Thus, because of the functional principle of the drag-and-drop technique per se, respondents may be encouraged to consciously and repeatedly attend to the key question components throughout the whole rating scale completion. In addition, similar to the approach of integrating reversed items within a set of rating scale items as "cognitive speed bumps" (Podsakoff, et al., 2003, p. 884), the drag-and-drop technique in rating scales may act as kind of 'navigational speed bumps' which cause respondents to slow down the process of responding, and thus, encourage them to process the four steps of the question-answer process more attentively and carefully.

Strengthening the Link between Items and Response Options

In grid questions, a considerable amount of information is presented simultaneously, while respondents are required to link the items presented at the beginning of each row with the response options arranged column by column at the top of the grid. This in turn implies a higher sensorimotor effort required for "greater hand-eye coordination, in that the respondent has to line

up both the row and column to identify the appropriate radio button” (Couper, et al., 2001, p. 236). Furthermore, the connection between the items and the response options is weakened because “when the respondent’s eyes focus on the answer categories of the question, the actual reference is out of view and vice versa” (Gräf, 2002, p. 56). Overall, while grid questions require a greater effort on the part of the respondent, they may also create a higher level of respondent confusion about the appropriate relation between rows and columns, compared to less complex question formats (Callegaro, Yang, et al., 2009; Couper, et al., 2013; Gräf, 2002; Jenkins & Dillman, 1997; Kaczmirek, 2010).

As part of their guidelines for good questionnaire design, Jenkins and Dillman (1997) stated that “one way to simplify the respondent’s task is to ask a comprehensive question, which both visually and logically consolidates the information for the respondent” (p. 186). By “presenting *conceptually connected but physically disconnected* information” (italics in original), grid questions explicitly conflict with this general design principle (Jenkins & Dillman, 1997, p. 186). By contrast, both drag-and-drop scales agree with this principle by strengthening the link between the items and the response options of rating scales. The drag-and-drop interaction is considered a ‘compound task’ in terms of integrating the performance of two separate tasks into one operation: the selection of a draggable element and the positioning at a desired position. In this regard, it is assumed that “the user models the compound task as a single entity” (Buxton, 1986, p. 4). Thus, in both drag-and-drop scales the link between single items and the respective response options is strengthened since both components are expected to be regarded as an entity by the respondents. In this regard, respondents are believed to be in a better position to thoughtfully bring together the items and the response option, thus being more carefully in matching an individual item with the respective response option.

Task Difficulty and Increasing Cognitive Load

Both versions of the drag-and-drop rating scales introduced here make use of a more sophisticated input method than conventional rating scales using radio buttons as data input method. Therefore, the task difficulty is likely to be increased. Compared to conventional grid questions with radio buttons simply necessitating a single click, rating scales using drag-and-drop technique “may

be more demanding and require more hand-eye coordination” (Funke, et al., 2011, p. 223).

In principle, two opposing effects of increased task difficulty on survey responding are conceivable. On the one hand, a high level of task difficulty and the accompanying cognitive stimulation may trigger the respondent’s involvement and attention. As Malhotra (2009) argued concerning the risk of cognitive shortcuts in simple and complex ranking tasks, the simplicity of the response tasks may encourage a respondent’s inattentiveness and carelessness due to boredom and fatigue. Therefore, making the response task easier for the respondents may compromise data accuracy because of a decline in respondent motivation to attentively and carefully process all relevant question components and to make sufficiently precise distinctions between several rating scale items (Alwin & Krosnick, 1985; Gräf, 2002; Hui & Triadis, 1985; Knowles, et al., 1992; Malhotra, 2009). Conversely, high task complexity may encourage the attentive and careful processing of rating scales as respondents invest more cognitive effort simply because they have to in order to complete a more demanding response task (Malhotra, 2009).

On the other hand, a high level of task difficulty and the associated cognitive load may result in frustration and distraction from the actual response task (Couper, et al., 2013; Gräf, 2002; Kieruj & Moors, 2013). Respondents may become frustrated if the use of a particular rating scale is too demanding. Increasing frustration may find its expression in reduced motivation to spend sufficient effort to answer survey questions accurately (Kieruj & Moors, 2013). Furthermore, as a result of the high level of task difficulty and the high cognitive load involved, the respondents’ attention may be reduced and removed from the actual response task since “anything that requires extra effort on the part of the respondent diverts concentration from answering the questions” (Gräf, 2002, p. 62). Thus, complex rating tasks may increase the risk of cognitive shortcuts in terms of the respondents’ susceptibility to systematic response tendencies due to distraction from or interference with the actual response task (Couper, et al., 2006; Couper, et al., 2013; Gräf, 2002). Accordingly, high levels of perceived and actual respondent burden may discourage respondents from investing the “cognitive or sensorimotor effort in answering the questions” (Couper, et al., 2013, p. 4). At its extreme, this may result in survey breakoff, or else in increased item

nonresponse or various kinds of systematic response tendencies (Couper, et al., 2013; Kieruj & Moors, 2013).

Hence, the use of the drag-response and drag-item scale introduced here may either result in the respondents' cognitive stimulation and a more attentive and careful processing of the rating scale items, or quite the opposite, it may interfere with the respondents' processing capacity discouraging them to attentively and carefully process the rating scale. In this regard, the length of a respective drag-and-drop rating scale is considered a decisive factor moderating the task difficulty and the accompanying respondent burden in both drag-and-drop rating scales.

Number of Items in a Grid and Increasing Cognitive Load

Since increased respondent burden is likely to enhance the risk of cognitive shortcuts, the respondent effort required to answer a survey question should be kept within reasonable limits. One factor which decisively influences the extent of respondent burden is the number of rating scale items presented in a rating scale. Thus, in order to control perceived and actual respondent burden, the number of rating scale items must not exceed a certain maximum amount of items (Andrews, 1984; Dillman, et al., 2009, p. 179; Toepoel, et al., 2008). Likewise, by necessitating the respondents to combine a fairly large amount of information from rows and columns, grid questions may require an increased sensorimotor effort. This effort is likely to rise even further with an increasing number of items simultaneously presented in a single grid (Couper, et al., 2013; Dillman, et al., 2009, p. 179). Furthermore, respondent fatigue is likely to increase with an increasing number of rating scale items (Andrews, 1984; Drolet & Morrison, 2001).

In summary, gains in efficiency and speed in grid questions are likely to be at the expense of increased respondent burden and a higher risk of respondents being engaged in inattentive and careless responding, thus relying on cognitive shortcuts if grid questions consist of more than just a few items (Andrews, 1984; Couper, et al., 2013; Dillman, et al., 2009, p. 179; Drolet & Morrison, 2001). Therefore, although binding information on the optimal number of items in a grid question is missing, survey researchers are generally advised not to exceed a certain limit of items in order to prevent increased respondent burden and potential compromises in data accuracy (Andrews, 1984; Dillman, et al., 2009, p. 180; Toepoel, et al., 2009b).

Also, with regard to the drag-response and drag-item scale introduced here, the number of rating scale items presented in a drag-and-drop scale is expected to be a decisive factor impacting both the task difficulty and the respondent burden. Despite the functionality of auto-scrolling implemented in the drag-and-drop rating scales, copying or moving the draggable elements to their destination would require substantially more navigational effort if the starting and finishing point of the mouse move cannot be seen on the screen at the same time. This may be the case if a respondent's individual browser window size is rather small, or if a rating scale exceeds a certain number of items. Therefore, the extent of sensorimotor effort in answering the drag-response and drag-item scale is likely to be considerably increased with the greater length of a rating scale. This may still hold true even though a certain learning effect is to be expected in both drag-and-drop rating scales in terms of more practice and improved handling with each additional item being answered. Thus, "if the use of a particular scale is too strenuous, respondents might become frustrated and as a result lose motivation to give accurate responses" (Kieruj & Moors, 2013, p. 196). A decreased respondent motivation in turn may result in a decreased willingness to provide complete and accurate responses to the drag-and-drop rating scales.

Current Expectations and Assumptions at a Glance

In this study, dynamic drag-and-drop rating scale procedures have been designed and implemented in several Web experiments. Grid questions as the well-established standard for the design and administration of rating scales have been used for the purpose of assessing the accuracy of rating scale measures obtained by the application of the drag-response and drag-item rating scale procedure. Because there are no objective criteria for the validation of rating scale measures in surveys, grid questions were used in similar terms as a gold standard approach according to which "the given approach is regarded as valid to the extent that its results match the gold standard method" (Alwin, 2010, p. 410). But instead of considering the grid question to be a 'gold' standard, it rather served as a 'minimum' standard of data accuracy which was expected to be exceeded by the drag-and-drop rating scale procedures introduced here. In this study, data accuracy was assessed on the basis of systematic response tendencies that are frequently accompanied by rating scale measures. In addition, item missing data in terms of item

nonresponse and survey breakoff, as well as the time spent on completing the respective rating scale measures were used as rather indirect indicators of data accuracy.

The assessment of the respondents' susceptibility to various cognitive shortcuts commonly applied in order to minimize the extent of cognitive and navigational effort in answering rating scales was made on the basis of five systematic response tendencies: careless, nondifferentiated, acquiescent, and extreme responding as well as the respondents' systematic tendency to select one of the response options presented first, so called primacy effects. All these systematic response tendencies have in common that they arise from the respondents' inattention towards the rating scale content and carelessness towards the response task (Krosnick, 1999; Paulhus, 1991; Weijters, et al., 2013).

The drag-response scale and drag-item scale examined in this study were expected to increase the respondents' focus on the relevant components of a rating scale and strengthen the link between individual items and their respective response options, whereas at the same time, decelerate the respondents' answering of a rating scale. Greater respondent attention and decelerate responding were assumed to encourage deeper cognitive processing and more careful responding to rating scales. Thus, it was hypothesized that both drag-and-drop rating scales would prevent respondents from relying on cognitive shortcuts which was expected to result in a reduction of the incidences of careless responding, nondifferentiated responding, and primacy effects in both the drag-response and the drag-item scale compared to a standard grid. The task difficulty and the extent of effort required for attentive and careful processing of rating scales was expected to increase with the growing number of items in a rating scale. This higher extent of effort required by longer rating scales was thought to increase respondent burden and thus, decrease respondent motivation to attentively and carefully process rating scales. Hence, respondents were expected to be more likely to rely on cognitive shortcuts in longer rating scales, which in turn would generally result in more careless and nondifferentiated responding in rating scales with their increasing length³.

³ The effect of rating scale length was not tested related to the occurrence of primacy effects in order to ensure an adequate level of complexity and a reasonably large number

Careless responding. Concerning the extent of careless responding, it was expected that the use of the drag-response and drag-item scale would result in fewer response inconsistencies between the original and reversed items compared to a grid question. Since both drag-and-drop rating scales were supposed to generally encourage higher respondent attentiveness and carefulness in the processing of a rating scale, it was assumed that respondents would be less likely to merely rely on information provided by the question context, and engage in a more attentive and careful processing of each individual rating scale item instead. Therefore, respondents were thought to be more likely to notice the reverse wording of rating scale items in both drag-and-drop rating scales compared to the grid question. In the drag-item scale, the content of the rating scale items was especially highlighted; hence, respondents were expected to pay special attention to the item content. As a consequence, the drag-item scale would prevent inconsistencies between the original and reversed items even more effectively than the drag-response scale. Response inconsistencies between the original and reversed items were expected to increase with a larger number of items presented in a rating scale. This effect of rating scale length was expected to similarly occur in all three rating scale designs.

Nondifferentiated responding. Concerning the extent of nondifferentiated responding, it was expected that the use of the drag-response and drag-item scale would lead to a higher degree of scale differentiation compared to the grid question. The reason for this was supposed to be twofold: While the drag-item scale would clearly highlight the item content and thus, promote higher attention to the item content, the drag-response scale would bring the response options into focus. Furthermore, both drag-and-drop rating scales were expected to strengthen the link between each single item and the respective response options. Both aspects—a higher attention to the items and response options and a strengthened connection between these two key components of a rating scale—were expected to encourage the respondents' attention and care in processing the rating scales. First, respondents would be more attentive to distinctions in item content and more attentive to the full range of the response options provided. Second,

of cases in each experimental condition. Respective experimental designs are explained in greater detail in Chapter 8.

respondents would be more likely to repeatedly reassess the appropriateness of response options with direct reference to the current item instead of providing a first answer that seemed reasonable and simply adjusting subsequent answers to this initial one. However, the degree of scale differentiation was expected to decrease with an increasing scale length which was expected to similarly occur in all three rating scale designs.

Primacy effects. Concerning the occurrence of primacy effects, the drag-response scale was expected to be less affected by primacy effects compared to the grid or the drag-item scale. More specifically, it was assumed that in the drag-response scale, there would be a more careful decision making process about which of the response options ought to be selected each time the respondent made a choice. This behavior was thought of to be the effect of highlighting the meaning of the respective response options, whereby respondents would be more encouraged in the drag-response scale—as compared to the grid or drag-item scale—to take all response options into account before selecting one of them instead of simply selecting the first response option that seemed reasonable. This more careful response option selection in the drag-response scale was expected to continue for the duration of scale completion, i.e., might apply for every new item being answered. By contrast, predictions concerning the drag-item scale were less clear. On the one hand, the drag-item scale might decrease the risk of primacy effects compared to a grid question, by generally encouraging higher respondent attentiveness and carefulness in the processing of a rating scale. On the other hand, it was conceivable that the drag-item scale would increase the risk of primacy effects compared to the grid question since a top-end bias might be more likely in the drag-item scale. This systematic preference for selecting one of the response options listed topmost in a drag-item scale might be the result of the higher navigational effort required to select one of the response options arranged at the bottom.

Acquiescent and extreme responding. Acquiescent and extreme responding as further kinds of systematic response tendencies in rating scales are deemed to be primarily affected by characteristics of the respondent: Whereas acquiescence can be considered the result of a respondent's effort to give a favorable and polite impression, extremity is rather the result of individual differences in the interpretation of response options and the translation of a judgment into one of those. In general, the occurrence of

acquiescent and extreme responding is relatively stable over time and mostly unaffected by situational factors (Paulhus, 1991). Thus, method-related characteristics are considered to be of minor importance in affecting the likelihood of acquiescent and extreme responding. Even though it was assumed that most characteristics of questionnaire design and administration had no major impact on the occurrence of acquiescent and extreme responding, both were taken into account in the present study since they are considered one of the most common systematic response tendencies in rating scales (Van Dijk, et al., 2009; Van Herk, et al., 2004). In this regard, it was expected that differing rating scale designs would have no effect on the extent of acquiescent and extreme responding in rating scales. Furthermore, the number of rating scale items was expected to have no effect on acquiescence and extremity which would equally apply to all three rating scale designs.

In the present study, item missing data in terms of item nonresponse and survey breakoff were considered a rather indirect indicator of data accuracy. In general, item missing data may be indicative of the extent of respondent burden and respondent motivation to attentively and carefully process a set of rating scale items. More specifically, item missing data may be indicative of a respondent's difficulty in processing a rating scale (Beatty & Herrmann, 2002; De Leeuw, et al., 2003). Furthermore, it may reflect a respondent's reluctance to even start processing a rating scale because of its mere visual complexity (Beatty & Herrmann, 2002; Couper, et al., 2013; Peytchev, 2006). In this context, it is important to note that in the present study, respondents were allowed to skip single items or even an entire rating scale without being prompted to enter up missing values and without being prevented from proceeding with the survey.

Item Nonresponse. Concerning the risk of item nonresponse, it was expected that the use of the drag-response and drag-item scale would lead to a higher item nonresponse rate compared to a grid question in two respects. First, there might be a higher risk that all items of a rating scale were skipped in both drag-and-drop scales due to difficulties in the basic understanding of the drag-and-drop technique and its specific handling and processing requirements. An alternative explanation would be that respondents might be deterred by the visual appearance and the complexity of the drag-response and drag-item scale before even starting the processing of the rating scale

items. Second, there might also be a higher risk that single items remained unanswered as a consequence of the higher navigational effort required to successfully complete the entire rating scale when the items were presented in a drag-response or drag-item scale compared to a grid question. The risk of item nonresponse was expected to increase with a larger number of rating scale items. This effect of the rating scale length was expected to similarly occur in all three rating scale designs.

Survey Breakoff. Survey breakoff is often considered a more aggravated form of item nonresponse. In interactive Web surveys, survey breakoff may potentially occur at every page. If respondents are, for whatever reason, not able or not willing to provide an answer to a survey question, they may skip the question, or in an extreme, prematurely terminate the survey (Beatty & Herrmann, 2002; De Leeuw, et al., 2003). Hence, the underlying reasons for survey breakoff are similar to those of item nonresponse. Accordingly, one might assume that the use of the drag-response and drag-item scale would lead to a higher survey breakoff rate compared to a grid question. This might be either due to a lack of respondent ability to understand the drag-and-drop technique and its specific handling and processing requirements, or due to a lack of respondent motivation in view of the visual complexity of the drag-response and drag-item scale. However, in the present study, no significant effect of the rating scale design on the risk of survey breakoff was expected since respondents could skip any survey question and thus, each of the rating scale items without being prompted to go back and complete the rating scale. Similarly, the risk of survey breakoff was expected to be unaffected by the length of a rating scale which was presumably true of all three rating scale designs.

Similar to item missing data, response times provide further evidence of the extent of respondent burden and respondent motivation to attentively and carefully process a set of rating scale items. Thus, differences in response times may reflect the respondents' difficulties in processing a rating scale on the one hand, or may be indicative of the amount of effort respondents spent on processing the rating scale on the other hand (Callegaro, et al., 2005; Heerwegh, 2003; Tourangeau, et al., 2004). Concerning the time spent on completing a rating scale, it was expected that the drag-response and drag-item scale would need more time to be completed relative to the grid question.

The reasons were thought to be the following: First, respondents were expected to spend more time on the attentive and careful processing of the content of a rating scale when the items were presented in a drag-response or drag-item scale as compared to a grid question. Second, the navigation in both drag-and-drop rating scales, i.e., dragging an element to the position intended, was deemed more challenging and thus, would require more time than simply clicking a radio button. And third, respondents were expected to need more time to get a basic understanding of the drag-response and drag-item scale and their respective handling and processing requirements, compared to a well-known grid question. This effect was expected to be even more pronounced in the drag-item scale since it was held to initially require a more thorough and thus, more time-consuming examination of its functionality than the drag-response scale. Even though both drag-and-drop scales were based on the same navigation technique, the drag-item scale was considered to be less intuitive in terms of its functionality than the drag-response scale. Respondents might become acquainted with the functionality of the drag-response scale in relative short time because the visual structure of the drag-response scale resembled more a well-known grid and all relevant information was provided at a glance. By contrast, the drag-item scale differed more from the established rating scale designs with respect to its visual structure and the arrangement of items, with just one item was visible at first. Hence, the response times would be even longer in the drag-item scale than in the drag-response scale. With respect to the length of a rating scale, time savings were expected in longer rating scales since the time needed to answer subsequent items was likely to be reduced once a couple of items has already been processed. These time savings are commonly explained in terms of efficiency gains: The necessary question context would be increasingly enriched with each additional item being answered contributing to a comprehensive understanding of the underlying construct of a rating scale and thus, enabling a faster processing of subsequent items. This kind of facilitation effect with an increasing number of rating scale items was expected to occur in all three rating scale designs. In a grid question, additional time savings would also be attributable to some kind of fatigue effects, tempting respondents to speed up their response times in order to quickly finish the rating scale. In the drag-response and drag-item scale, additional time savings in longer rating scales were expected to arise as a result of learning effects. The improved handling

of the drag-and-drop technique once a few items have been answered successfully would allow respondents to increase the speed of responding with later rating scale items. These additional time savings in both drag-and-drop scales due to the facilitation effects and learning effects were expected to lead to an approximation in response times in longer rating scales, between the drag-response and drag-item scale on the one hand and the grid question on the other hand. Nevertheless, a significant difference in response times would still remain between both drag-and-drop rating scales and the grid question because of the generally higher cognitive and navigational requirements outlined above.

Besides differing rating scale designs of various lengths, the categorial arrangement of the response options was altered primarily for the purpose of systematically examining the occurrence of primacy effects in rating scales. Given the limited number and lack of consistent findings concerning the effects of categorial arrangement on the extent of nondifferentiated or extreme responding, reliable predictions could hardly be made. Nevertheless, it was assumed that the negative-to-positive response option order would promote higher scale differentiation compared to a positive-to-negative response option order. When defying the respondents' expectation of starting with the affirmative response option—i.e., when using the less common negative-to-positive response option order—respondents were expected to be more attentive and careful in processing all response options and thus, would actually be more likely to make use of the full range of the response options. This effect of the scale arrangement on the degree of differentiation was expected to similarly occur in all three rating scale designs. Whereas the extent of careless and acquiescent responding was not examined as function of the categorial arrangement⁴, the extent of extreme responding was expected to be unaffected by the categorial arrangement of the response options for essentially the same reasons as described above. Similarly, no effect of scale arrangement on the risk of item nonresponse or survey breakoff was expected in any of the three rating scale designs. Related to response times, it was expected, in compliance with previous findings, that a negative-to-positive response option order would take longer to be processed and thus, would

⁴ See footnote 3.

increase the time spent on completing the rating scale items. This effect of categorial arrangement was expected to similarly occur in all three rating scale designs.

The effectiveness of the drag-response and drag-item scale and related expectations concerning the occurrence of systematic response tendencies, concerning the risk of item missing data, and concerning the impact on response times were experimentally tested. Detailed information on the study design, the experimental conditions, and the indicators of data accuracy are provided in the next sections.

8. METHODS

Within the scope of this study, three Web surveys were conducted in the period between July 2012 and August 2013: the Panel Survey 2012, the University Applicants Survey 2012, and the University Applicants Survey 2013 (section 8.1). Three experimental designs were developed to examine the effectiveness of the two drag-and-drop rating scales described above in reducing the risk of systematic response tendencies and the occurrence of item missing data in rating scales compared to a grid question as a conventional multiple-item-per-screen design and two single-item-per-screen designs. Based on the experimental manipulation of the factors scale format, scale length, scale arrangement, and scale sequence (section 8.2), three different experimental designs were distinguished (section 8.3), taking into account an adequate level of complexity and a reasonably large number of cases in each experimental condition. Using various instruments (section 8.4), the effectiveness of differing scale formats and scale lengths, as well as varying orders of items and response options were assessed in terms of different indicators of data accuracy (section 8.5).

For a better overview, the three experimental designs are displayed in Table 1. Each experimental design was implemented in the form of a between-subjects factorial design with respondents being randomly assigned to one of the experimental conditions. The experimental designs 1 and 3 were implemented several times across and within the three Web surveys, resulting in a total of six distinct experiments. Within the same survey, the experiments were independently randomized, i.e., the assignment of a respondent to an experimental condition in one experiment was independent of his or her assignment in another experiment to reduce possible carryover effects.

Table 1: Overview of the six experiments implemented in three Web surveys

Experiment #	Experimental Design 1			Experimental Design 2	Experimental Design 3	
	Length x Format			Arrangement x Format	Arrangement x Sequence x Format	
	1.1	1.2	1.3		2	3.1
Panel Survey 2012	X			X		
University Applicants Survey 2012		X				
University Applicants Survey 2013			X		X	X

8.1 Participants

8.1.1 Panel Survey 2012

The Panel Survey 2012 about family and private life was conducted from July 4 to July 21, 2012 among a sample of 2,264 opt-in panel members studying at the Darmstadt University of Technology (TU Darmstadt, Germany). In total, 960 respondents (42.4%) logged in to start the survey, 846 (37.4%) of which completed the survey⁵. In the final sample, 46.5% were women and the respondents' age ranged from 19 to 62 years ($M = 24.38$, $SD = 4.54$). Respondents predominantly had extensive experience in dealing with computer technologies and the Internet. Based on a self-assessment of computer and Internet literacy, the vast majority of respondents regarded themselves as computer and Internet users with advanced skills (64.2% and 63.8%, respectively), and expert skills (22.7% and 28.4%, respectively), whereas only a fairly small number of respondents indicated basic knowledge in using computers and the Internet (13.1% and 7.8%, respectively). Based on Bescherer and Spannagel's (2011) short scale for measuring computer-related self-efficacy (CUSE-D-r) on a 6-point rating scale, the respondents' self-evaluation testified a very high level of computer literacy ($M = 5.00$, $SD = 0.87$). On average, respondents spent more than three hours per day with the computer ($M = 3.53$, $SD = 1.14$) and the Internet ($M = 3.18$, $SD = 1.18$). A

⁵ The completion rate was calculated according to AAPOR RR6 and refers to the number of respondents who completed the survey among all eligible respondents who received an e-mail invitation. Respondents were included who answered the entire questionnaire ('complete respondents') as well as respondents who answered at least 50% of all applicable questions ('partial respondents') (AAPOR, 2011).

respondent's prior Web survey experience proved to be rather moderate, with an average of three Web surveys a respondent participated in within the last 12 months ($M = 3.32$, $SD = 6.54$). The average time to complete the survey was about 17 minutes ($M = 17.71$, $SD = 11.06$)⁶.

8.1.2 University Applicants Survey 2012

This Web survey was conducted among university applicants at the TU Darmstadt from August 4 to August 30, 2012 and was about their motivation for applying and their expectations of their studies at the TU Darmstadt. A total of 18,463 university applicants received an e-mail invitation, 9,464 respondents (51.3%) of which started the survey and 5,977 (32.4%) completed the survey⁷. In the final sample, 48.0% were women and the respondents' age ranged from 16 to 55 years ($M = 20.50$, $SD = 2.88$). Again, the respondents' knowledge in dealing with computers and the Internet was high with 65.4% and 69.3% accordingly assessing themselves as advanced computer and Internet users, 14.5% and 22.0% respectively accounting themselves as professionals, and merely 20.1% and 8.7% respectively indicating to be a beginner in dealing with computers and the Internet. On average, respondents spent more than 11 hours a week with computer and Internet usage ($M = 11.26$, $SD = 11.72$). By contrast, the respondents' prior Web survey experience was relatively low with less than two Web surveys a respondent participated in within the last 12 months ($M = 1.81$, $SD = 4.09$). Average survey completion time was about half an hour ($M = 31.02$, $SD = 18.76$)⁸.

⁶ For the sake of completeness, the proportion of missing values was indicated on the percentage basis of 846 cases of the final sample: 0.8% for age, 0.1% for Internet-related knowledge, 1.5% for the index on computer-related self-efficacy, 0.4% for time spent with the computer, 0.1% for time spent with the Internet, 1.1% for the number of prior Web surveys, and 4.4% for survey completion time.

⁷ See footnote 5.

⁸ Based on the final sample of 5,977 cases, the proportion of missing values amounted to 0.9% for gender, 1.3% for age, 0.3% for computer-related knowledge, 0.5% for Internet-related knowledge, 4.9% for time spent on computer and Internet usage, 2.3% for the number of prior Web surveys, and 16.7% for survey completion time.

8.1.3 University Applicants Survey 2013

This Web survey was conducted among university applicants at the TU Darmstadt from July 24 to August 26, 2013. Again, the questionnaire was about the applicants' motivation and their expectations with regard to studying at the TU Darmstadt. A total of 18,327 university applicants received an e-mail invitation, 9,992 (54.5%) of which logged in to start the survey and 7,395 (40.4%) completed the survey⁹. 45.4% of the final sample were women and the respondents' age ranged from 16 to 65 years ($M = 20.50$, $SD = 3.19$). The respondents' knowledge of computer and Internet usage could be regarded as fairly good since 63.1% and 67.6% respectively assessed themselves as advanced computer and Internet users, 12.4% and 20.2% respectively described themselves as professionals, and 24.5% and 12.3% respectively classified themselves as beginners in dealing with computers and the Internet. On average, respondents spent more than 18 hours a week with the computer and the Internet ($M = 18.65$, $SD = 14.94$). The respondents' prior Web survey experience proved to be rather moderate with an average of about two Web surveys a respondent participated in within the last 12 months ($M = 2.34$, $SD = 8.28$). The average time to complete the survey was about half an hour ($M = 28.50$, $SD = 14.78$)¹⁰.

8.2 Experimental Manipulation

8.2.1 Scale Format

The main manipulation in the present study was related to the format of rating scales with the primary focus being on the comparison of two different drag-and-drop scale formats against a standard grid. Both the drag-response and drag-item scale are described and illustrated in detail above (see also Chapter 7). As a kind of conventional multiple-item-per-screen design, a grid format is the most prevalent rating scale format where respondents are required to indicate their answers by simply clicking on a radio button. Two single-item-

⁹ See footnote 5.

¹⁰ Based on the final sample of 7,395 cases, the proportion of missing values was 2.9% for gender, 3.3% for age, 2.4% for computer-related knowledge, 2.6% for Internet-related knowledge, 7.8% for time spent on computer and Internet usage, 4.6% for the number of prior Web surveys, and 13.1% for survey completion time.

per-screen designs were implemented as additional standards of comparison since they were considered the only alternative to a standard grid well established in survey research up to now. Nonetheless, the present experiments focused on a comparison between the novel drag-and-drop rating scale formats on the one hand and conventional grid formats as a kind of multiple-item-per-screen design on the other hand. Therefore, both of the single-item-per-screen designs were not part of the hypothesis formulation and testing. Table 2 provides an overview of the various rating scale formats that were integrated in the experiments.

Table 2: Description of differing rating scales formats tested

Format	Description
a) Grid	A general question request introducing into the subject of a rating scale is followed by several rating scale items that are presented in rows with each row comprising a single item. The response options are arranged in the topmost cells.
b) Drag-Response	A general question request is followed by several rating scale items presented in rows with each row comprising a single item. Response options are arranged in the top line of the scale corpus. While items are fixed, response options need to be dragged towards the items (see also section 7.1 for details).
c) Drag-Item	A general question request is followed by several items which are stacked on top of one another in the top line of the scale corpus. Response options are presented in the rows with each row comprising a single response option. While response options are fixed, items need to be dragged towards the response options (see also section 7.2 for details).
d) One-Vertical	The introductory page containing the general question request is followed by the rating scale items each being presented separately on a single screen. Below each item response options are arranged vertically.
e) One-Horizontal	Introduced by a general question request, each item is presented separately on a single screen with response options being horizontally arranged below the respective item.

As distinct from conventional radio button scales in formats a, d, and e, both drag-and-drop scales were navigated by mouse dragging using JavaScript. Since this data input method has been used rather rarely in Web surveys up to now, specific input instructions were deemed necessary for the drag-and-drop scales. In trying to ensure a high degree of comparability by keeping all question components that did not belong to the experimental manipulation as constant as possible, respective input instructions were also included in the

radio button scales. The specific input instructions for the various rating scale formats varied by type of input method. Formats using conventional radio button scales instructed respondents as follows: ‘To give an answer, click on the desired radio button.’ Respondents presented with the drag-response format were given the instruction: ‘To give an answer, left-click on the desired answer-card, drag it to the blue response field with the mouse button pressed, and release the mouse button.’ To the drag-item format, the following instruction was added: ‘To give an answer, left-click on the item-card, drag it to the desired blue response field with the mouse button pressed, and release the mouse button. Several items can be assigned to the same response.’

8.2.2 Scale Length

Concerning the factor scale length, a 6-item, 10-item, and 16-item scale was chosen for range of a short, medium, and large rating scale. By experimentally varying the length of rating scales, it was examined whether the length of a rating scale is related to the likelihood of systematic response tendencies and item missing data. In addition, potential interaction effects arising between scale format and scale length could be observed.

The specification of the actual numbers of items tested in different experimental conditions was deduced from previous research findings. In general, the findings of a meta-analysis showed that, based on a total of 274 rating scales implemented in 75 different studies, 73% of the rating scales consisted of up to 6 items, 18% comprised 7 to 10 items, and merely 9% included more than 10 items (Hinkin, 1995). In general, the maximum length of rating scales commonly recommended by questionnaire design guidelines is reached with a 10-item scale, even though systematic testing and clear findings are rather rare (Toepoel, et al., 2009b). Thus, the 6-item scale used in the present experiments can be regarded quite common in practice and is mostly considered unproblematic because of its shortness, even when presented in a grid format. By contrast, the 10-item scale integrated in the experiments at hand corresponds to the maximum number of rating scale items generally recommended by survey researchers; although this scale length is certainly no exception in practice, even in grid formats. However, the 16-item scale as the third experimental condition can be considered a

comparably long scale which might be problematic in either rating scale format.

In Web surveys, the number of rating scale items that is visible simultaneously on a single screen depends on the respondent's individual screen resolution or browser window size, and the respective necessity of scrolling. Since scrolling is perceived to be more burdensome as indicated by less positive respondent evaluations in a Web survey, it is recommended to place 4 to 10 items on a single screen, so that no scrolling is required to answer the whole set of rating scale items (Toepoel, et al., 2009b). Toepoel and colleagues (2008) showed that the layout of a questionnaire was evaluated best in terms of how well-designed and answerable the questionnaire was assessed when a rating scale of 10 items was split up into two five-item-per-screen formats than when presenting all 10 items in a single grid or each item on a single screen. Based on these findings, it can be concluded that respondents prefer a grid format, however, solely up to a certain amount of items (Toepoel, et al., 2008).

8.2.3 Scale Arrangement

Concerning the factor scale arrangement, a positive scale arrangement presenting response options from positive-to-negative and a reverse scale arrangement from negative-to-positive were distinguished. By experimentally varying the response option order of a rating scale, the occurrence of primacy effects in terms of a respondent's systematic preference for response options listed first (further explanations in section 8.5.5) depending on varying scale formats could be systematically examined.

8.2.4 Scale Sequence

Concerning the factor scale sequence, an original and reverse item order was distinguished. By experimentally varying the sequence of rating scale items, the occurrence of semantic-order effects in terms of responses to later items being affected by preceding items (further explanations in section 8.5.6) depending on varying scale formats could be systematically examined.

8.3 Experimental Designs

8.3.1 *Experiment 1*

In Experiment 1, the format and length of a rating scale was varied. Factor scale format included each of the five different rating scale formats, i.e., the grid, drag-response, drag-item, one-vertical, and one-horizontal format. Factor scale length varied the number of items the rating scale was composed of, i.e., 6, 10, or 16 items. In Experiment 1.1 embedded in the Panel Survey 2012, respondents were randomly assigned to either a 10-item scale or a 16-item scale, resulting in a 2 (scale length: 10 vs. 16 items) x 5 (scale format: grid vs. drag-response vs. drag-item vs. one-vertical vs. one-horizontal) between-subjects factorial design. In Experiment 1.2 implemented in the University Applicants Survey 2012, the design of the first experiment was extended by another shorter scale length. Thus, respondents were randomly assigned to one of 15 experimental conditions in a 3 (scale length: 6- vs. 10- vs. 16-item scale) x 5 (scale format: grid vs. drag-response vs. drag-item vs. one-vertical vs. one-horizontal) between-subjects factorial design. Experiment 1.3 implemented in the University Applicants Survey 2013 aimed at replicating the findings of the former two experiments, which is why the same 3 x 5 between-subjects factorial design was used as in Experiment 1.2.

Table 3 to Table 5 display the number of completes per experimental condition in Experiments 1.1, 1.2, and 1.3. The differences between the number of completes per experimental condition and the total number of completes mentioned in section 8.1.1 were explained by the exclusion of respondents answering the questionnaire in a browser window smaller than 900 x 500 pixels. These respondents were automatically screened during survey completion and excluded from the experimental conditions in order to exclude respondents with a smart phone and tablet PC who could not drag and drop via their touch screen navigation system.

Random assignment of respondents to one of the experimental conditions is aimed at ensuring that the experimental conditions are comparable with respect to important respondent characteristics in order that “any observed differences between the two groups can be attributed to treatment effects rather than to differences in subsample composition” (Groves, et al., 2008, p. 834). In Experiments 1.1, 1.2, and 1.3, the randomization was successful in that there were no significant differences

between the experimental conditions¹¹ in regard to the respondent's gender, age, computer and Internet literacy, and prior Web survey experience¹². Despite random assignment of the remaining respondents to one of the experimental conditions using an intra-system random function of the survey software, the number of completes differed considerably in some of the experimental conditions. To anticipate one of the results reported later, this uneven distribution was not due to a differential risk of survey breakoff in the different experimental conditions.

Table 3: Number of completes per experimental condition (Experiment 1.1)

Length	Format					Total
	Grid	Drag-R	Drag-I	One-V	One-H	
10	76	72	78	82	73	381
16	77	83	79	77	74	390
Total	153	155	157	159	147	771

Table 4: Number of completes per experimental condition (Experiment 1.2)

Length	Format					Total
	Grid	Drag-R	Drag-I	One-V	One-H	
6	347	367	353	353	348	1,768
10	360	355	356	340	360	1,771
16	348	340	354	329	352	1,723
Total	1,055	1,062	1,063	1,022	1,060	5,262

Table 5: Number of completes per experimental condition (Experiment 1.3)

Length	Format					Total
	Grid	Drag-R	Drag-I	One-V	One-H	
6	395	388	383	392	374	1,932
10	395	398	408	393	401	1,995
16	399	398	362	412	398	1,969
Total	1,189	1,184	1,153	1,197	1,173	5,896

¹¹ In Experiment 1.3, the respondent's mean age ($F(2, 5,703) = 4.16, p < .05, \eta^2 = .001$) was slightly but significantly higher in the 6-item scale (20.7) compared to the 16-item scale (20.5).

¹² Calculations on gender and literacy in computer and Internet use were based on Pearson's chi-squared tests with an alpha level of .05. Variance analyses with an alpha level of .05 were used for calculations on age and Web survey experience.

8.3.2 Experiment 2

Aside from variations in scale format, Experiment 2 varied the categorical arrangement of the response options in a rating scale. According to the factor scale arrangement, respondents were randomly assigned to either a rating scale with response options arranged from the most positive to the most negative or, exactly the opposite, from the most negative to the most positive, resulting in a 2 (scale arrangement: positive-to-negative vs. negative-to-positive) x 5 (scale format: grid vs. drag-response vs. drag-item vs. one-vertical vs. one-horizontal) between-subjects factorial design that was embedded in the Panel Survey 2012. Table 6 displays the number of completes per experimental condition in Experiment 2.

Again, the differences between the number of completes per experimental condition and the total number of completes were explained by the exclusion of respondents answering the questionnaire in a browser window smaller than 900 x 500 pixels. The random assignment of respondents to one of the experimental conditions was successful since no significant differences between the experimental conditions in regard to the respondent's gender, age, computer and Internet literacy, and prior Web survey experience were found. As aforementioned, despite the random assignment to the experimental conditions, there were considerable differences in the number of completes in some experimental conditions which can be traced back to an unreliable intra-system random function of the survey software, and was not due to differential survey breakoff rates in the different experimental conditions.

Table 6: Number of completes per experimental condition (Experiment 2)

Arrangement	Format					Total
	Grid	Drag-R	Drag-I	One-V	One-H	
Pos-Neg	67	90	86	87	68	398
Neg-Pos	77	84	59	83	70	373
Total	144	174	145	170	138	771

8.3.3 Experiment 3

In Experiment 3, the format of the rating scale was varied but as distinct from Experiments 1 and 2, the variations in scale format were confined to the grid format, the drag-response format, and the drag-item format¹³. Furthermore, the categorial arrangement of the response options (replication of Experiment 2, see also section 8.3.2) and the sequence of the rating scale items were manipulated. According to the factor scale arrangement, response options were arranged either in a positive-to-negative or negative-to-positive order. With regard to factor scale sequence, items were presented either in the original order, with the first item of the rating scale being in first place or in the reverse order where the originally first item of the rating scale came last. Thus, respondents were randomly assigned to one of the 15 experimental conditions in a 2 (scale arrangement: positive-to-negative vs. negative-to-positive) x 2 (scale sequence: original vs. reverse) x 3 (scale format: grid vs. drag-response vs. drag-item) between-subjects factorial design.

Experiment 3 was twice-implemented in the University Applicants Survey 2013 based on two different experimental questions. The number of completes per experimental condition in Experiments 3.1 and 3.2 are displayed in Table 7 and Table 8, whereby the differences between the number of completes per experimental condition and the total number of completes were again explained by the exclusion of respondents answering the questionnaire in a browser window smaller than 900 x 500 pixels. Random assignment of the respondents to one of the experimental conditions was successful in Experiments 3.1 and 3.2 since no significant differences between the experimental conditions¹⁴ in regard to the respondent's gender,

¹³ The reasons for not including both single-item-per-screen designs in Experiment 3 were threefold: (a) the present study was focused on the examination of new drag-and-drop formats as compared to a standard grid format. Therefore, single-item-per-screen formats were not part of hypothesis formulation and testing, (b) the findings of Experiments 1 and 2 concerning the one-vertical and one-horizontal format as compared to one of the other three formats were largely consistent and revealed no striking insights which would have made replication necessary, and (c) given a three factorial design including format, arrangement, and sequence of a rating scale, a restriction to merely three scale formats was in favor of a manageable complexity of the experimental design and a reasonably large number of cases in each experimental condition.

¹⁴ In Experiment 3.1, there was a significant effect for the drag-response format ($\chi^2(2, 1,962) = 6.93, p < .05$) with a significantly higher proportion of experts (54.7%) versus

age, computer and Internet literacy, and prior Web survey experience were found. As noted in the former experiments, there were considerable differences in the number of completes in some experimental conditions, despite random assignment. It has to be stated again that this was due to the unreliable intra-system random function of the survey software and does not indicate differential survey breakoff rates in the different experimental conditions.

Table 7: Number of completes per experimental condition (Experiment 3.1)

Arrangement	Sequence	Format			Total
		Grid	Drag-R	Drag-I	
Pos-Neg	Original	466	524	501	1,491
	Reverse	480	455	492	1,427
	Total	946	979	993	2,918
Neg-Pos	Original	471	484	495	1,450
	Reverse	487	535	475	1,497
	Total	958	1,019	970	2,947
Total	Original	937	1008	996	2,941
	Reverse	967	990	967	2,924
	Total	1,904	1,998	1,963	5,865

novices (44.7%) in the positive-to-negative arrangement, and a significantly higher proportion of novices (55.3%) versus experts (45.3%) in the negative-to-positive arrangement. In Experiment 3.2, a significant interaction between scale format and scale arrangement ($F(2, 5,833) = 3.18, p < .05, \eta^2 = .001$) indicated a higher mean number of participations in Web surveys within the last 12 months in the negative-to-positive (2.8) compared to the positive-to-negative scale arrangement (2.2) for the grid format, whereas the opposite was true for the drag-item format with a higher number of participations in Web surveys within the last 12 months in the positive-to-negative (2.8) compared to the negative-to-positive scale arrangement (2.4).

Table 8: Number of completes per experimental condition (Experiment 3.2)

Arrangement	Sequence	Format			Total
		Grid	Drag-R	Drag-I	
Pos-Neg	Original	505	494	533	1,532
	Reverse	514	499	472	1,485
	Total	1,019	993	1,005	3,017
Neg-Pos	Original	465	505	487	1,457
	Reverse	511	496	517	1,524
	Total	976	1,001	1,004	2,981
Total	Original	970	999	1,020	2,989
	Reverse	1,025	995	989	3,009
	Total	1,995	1,994	2,009	5,998

8.4 Instruments

8.4.1 Experiment 1

Experiment 1.1 was embedded in the Panel Survey 2012 on the subject of family and private life with the experimental question addressing perceived emotional intelligence. The rating scale items were part of Otto and colleagues' (2001) three-dimensional scale on a person's perceived emotional intelligence in terms of (1) attention to emotions (13 items), (2) clarity of emotions (9 items), and (3) repair of emotions (6 items). The selection of a 10-item or a 16-item scale was based on three different criteria: First, an equal number of original and reversed items in total, as well as within each subscale were selected in order to obtain balanced (sub)scales. Second, within each subscale, paired items were selected measuring the same content in reverse wording (item pairs marked with the same subscript (a, b, c) in Appendix A, Table 48). Third, taking account of the first two criteria, items have been selected according to their factor loadings and discriminatory power. The item order, item assignment to subscale, item wording (German and English), and item selection for each scale of differing length are presented in Appendix A, Table 48. A respondent's level of agreement and disagreement to the items was assessed on a bipolar 5-point Likert-type scale ranging from 1 ('completely agree') to 5 ('completely disagree'). All response options were labeled verbally. After recoding half of the items (marked with a minus (-) in

Appendix A, Table 48), high values signified a high ability to attend to, distinguish among, and regulate emotions.

Experiments 1.2 and 1.3 were based on the same rating scale items that asked for a respondent's achievement motive and were implemented in the University Applicants Survey 2012 and 2013, respectively. Items were originally taken from Modick's (1977) three-dimensional scale on an individual's achievement motive in terms of (1) achievement motivation relating to the future (22 items), (2) performance-inhibiting anxiety (22 items), and (3) performance-enhancing tension (12 items). Again, the selection of a 6-item, a 10-item, or a 16-item scale was carried out according to three different criteria: First, in order to obtain balanced scales consisting of an equal number of original and reversed items in total, as well as within each subscale, some of the original items were analogously rephrased to obtain a reverse worded equivalent (marked with an asterisk (*) in Appendix A, Table 49). Particular emphasis was placed on retaining the meaning of the original item while considering two main requirements for appropriate item reversals (Peabody, 1961). Item reversals need to be (a) more than solely the logical reversal by simply adding a 'not', and (b) moderate in content. Second, within each subscale, paired items were selected measuring the same content but in reverse wording (item pairs marked with the same subscript (a, b, c) in Appendix A, Table 49). Third, taking account of the first two criteria, items have been selected according to their factor loadings and discriminatory power. Item order, item assignment to subscale, item wording (German and English), and item selection for each scale of differing length are presented in Appendix A, Table 49. A respondent's level of agreement and disagreement to the items was assessed on a bipolar 5-point Likert-type scale ranging from 1 ('completely agree') to 5 ('completely disagree'). All response options were labeled verbally. After recoding half of the items (marked with a minus (-) Appendix A, Table 49) high values signified (1) high forward-looking need for achievement, (2) high debilitating anxiety, and (3) high facilitating tension.

8.4.2 *Experiment 2*

Experiment 2 was embedded in the Panel Survey 2012 on the issue of family and private life with the experimental question being about the Big-Five based personality traits. A German version of the Ten-Item Personality

Inventory (TIPI) was used which was introduced by Gerlitz and Schupp (2005) and originally developed by Gosling and colleagues (2003). Personality traits were measured in terms of neuroticism, extraversion, openness, agreeableness, and conscientiousness, implemented in a five-dimensional scale. Item order, item assignment to subscale, and item wording (German and English) are presented in Appendix A, Table 50. A respondent's level of agreement and disagreement to the items was assessed on a bipolar 5-point Likert-type scale ranging from 1 ('completely agree') to 5 ('not agree at all') with the response order of negative-to-positive scale arrangement being recoded in advance of the analyses. All response options were labeled verbally. Experiment 2 also involved a balanced scale since each of the five personality traits was measured by means of two items, one of which had to be recoded with regard to its content (marked with a minus (-) in Appendix A, Table 50).

8.4.3 Experiment 3

Concerning Experiment 3, the same experimental design was twice-implemented in the University Applicants Survey 2013 in terms of two different experimental questions: first, reasons for social advancement (Experiment 3.1), and second, locus of control (Experiment 3.2). In Experiment 3.1, items were used from the two-dimensional scale on reasons for social advancement introduced by Weinhardt and Schupp (2011) which distinguishes between legitimate (5 items) and illegitimate (7 items) paths to success. In the 8-item rating scale used in Experiment 3.1, each of the two subscales was represented by four items, respectively. In Experiment 3.2, the two-dimensional scale on locus of control published in Weinhardt and Schupp (2011) was used. Based on this scale, an internal (3 items) and external (5 items) locus of control can be distinguished. Original item order, item assignment to subscale, and item wording (German and English) are presented in Appendix A, Table 51 for Experiment 3.1 and in Appendix A, Table 52 for Experiment 3.2. A respondent's level of agreement and disagreement to both rating scales was assessed on a bipolar 5-point Likert-type scale ranging from 1 ('completely agree') to 5 ('completely disagree') with each response option being verbally labeled.

8.5 Measures

In the present study, the benefits and challenges of implementing alternative rating scale procedures using drag-and-drop in Web surveys are examined with regard to their respective effects on data accuracy. In addition to the extent of item missing data in terms of item nonresponse and survey breakoff, systematic response tendencies such as careless, nondifferentiated, acquiescent, and extreme responding are commonly assumed to reflect a respondent's susceptibility to cognitive shortcuts and are held to be indicative of impaired data accuracy in rating scales. Moreover, the extent of primacy effects and semantic-order effects in rating scales are often examined for this purpose. Moreover, response times as one type of paradata can be used to gain a better understanding of the respondents' processing of survey questions and the extent of effort and attention respondents spent on completing them.

8.5.1 *Careless Responding*

In the present study, careless responding refers to a respondent's inattentiveness or carelessness towards the reverse wording of rating scale items. Because of the respondent's reduced attention to item specifics, the risk of missing the reverse wording of an item is increased (Podsakoff, et al., 2003; Weijters, et al., 2013).

Based on a balanced scale consisting of several pairs of original and reversed items, the extent of careless responding can be assessed by means of Pearson's correlations between the item pairs. A respondent's inattentiveness or carelessness towards the reverse wording of rating scale items may result in response inconsistencies which are reflected in lowered item-pair correlations between the original and reversed items. Item-pair correlations were calculated after inversion of reverse worded items.

8.5.2 *Nondifferentiated Responding*

In general, nondifferentiation as a kind of strong satisficing is considered the result of a respondent's inability or unwillingness to process all the steps of the question-answer process conscientiously, leading to identical or almost identical answers to different rating scale items (Herzog & Bachman, 1981; Krosnick, 1991; McCarty & Shrum, 2000).

Following the approach of McCarty and Shrum (2000), the degree of differentiation was measured by means of P_d (rho). This differentiation index reflects the variation in the number of different response options used by a respondent and can be calculated according to the following formula:

$$P_d = 1 - \sum_{i=1,n} P_i^2$$

Differentiation index could attain values from 0—if a respondent selected the same response option for every single item—to maximum 1 with the exact maximum value depending on the number of rating scale items and response options provided. A higher value indicated a higher degree of differentiation (McCarty & Shrum, 2000). Differentiation index was calculated without prior inversion of reverse worded items.

8.5.3 *Acquiescent Responding*

Acquiescent responding is characterized by the tendency to agree rather than disagree with rating scale items irrespective of the item content. Acquiescence is commonly measured at scale level by presenting respondents with a mix of original and reverse worded rating scale items, arranged within the same scale.

Following Van Herk and colleagues (2004), a respondent's extent of acquiescent and non-acquiescent responding was examined on the basis of a balanced rating scale consisting of an equal number of original and reverse worded rating scale items. An acquiescence index was calculated by counting the number of clearly positive responses (response options 1 or 2), subtracting the number of clearly negative responses (response options 4 or 5), and dividing the resultant value by the total number of rating scale items as described in the following formula:

$$ARB = (\sum_i N_{i_{positive}} - \sum_i N_{i_{negative}}) \div N$$

Acquiescence index could attain values from -1 to +1, with a value around 0 indicating balanced responses without any systematic tendency to agree or disagree with the rating scale items. A positive value indicated that a respondent was more likely to agree rather than disagree with the rating scale

items (Van Herk, et al., 2004). Acquiescence index was calculated without prior inversion of reverse worded items.

8.5.4 *Extreme Responding*

Extreme responding is described as a respondent's tendency to select the most extreme response options while excluding intermediate ones irrespective of item content. According to the procedure of Van Herk and colleagues (2004), extremity index was computed by counting the number of extreme responses on both ends of the rating scale (response options 1 or 5) and dividing it by the total number of rating scale items as specified in the following formula:

$$ERB = \sum_i N i_{extreme} \div N$$

Extremity index could attain values from 0 to +1 with higher values denoting a higher incidence of extreme responding (Van Herk, et al., 2004). Extremity index was computed without prior inversion of reverse worded items.

8.5.5 *Primacy Effects*

Within the scope of the satisficing theory, primacy effects as a kind of response-order effect are explained as follows: A respondent considers oneself satisfied with an acceptable rather than optimal answer in order to minimize his or her cognitive and/or navigational effort which finds expression in the selection of the first acceptable response option while considering remaining response options to a lesser degree or ignoring them completely (Krosnick, 1999; Krosnick & Alwin, 1987; Malhotra, 2009; Tourangeau, 1984).

One possible approach to systematically examine the occurrence of primacy effects in rating scales is counterbalancing the direction of response option presentation. In rating scales with ordered response options, complete randomization of the response option order is not useful. Thus, counterbalancing allows for two different scale directions: either a positive scale direction, presenting response options from positive-to-negative, or in a reverse scale direction from negative-to-positive (Krosnick & Presser, 2010).

Given this experimental variation in scale arrangement, primacy effects can be quantified in terms of differences in response distributions and item

means. A systematic shift towards the left side of a horizontally arranged rating scale (left-side bias) or to the top end of a vertically arranged rating scale (top-end bias) indicates a primacy effect in terms of systematic shifts in response distributions and item means towards the leftmost or topmost side of a rating scale.

8.5.6 Semantic-Order Effects

Semantic-order effects as a kind of item-order effects are typically examined in terms of mean shifts and reliability shifts (Knowles, et al., 1992). For examining the occurrence of semantic-order effects, the sequence of rating scale items needs to be counterbalanced. Items can either be presented in a completely randomized order, whereby item content is counterbalanced over every conceivable position within a rating scale, or in a forward and backward order (Bishop, 2008; Bradlow & Fitzsimons, 2001). If the cognitive processing of a single item is not affected by its context, the item means and item reliabilities should remain unchanged, irrespective of the order of the rating scale items. However, if a systematic shift in item means and item reliabilities occurs depending on the respective scale sequence, this is regarded as evidence of responses to later items being affected by preceding items (Knowles, et al., 1992).

While in terms of mean shifts, it can merely be said that a mutual influence of rating scale items exists, the assessment of item-total correlations additionally enables statements about potential increases in the consistency and reliability of measures (Knowles, 1988; Knowles & Byers, 1996; Knowles, et al., 1992; Krosnick, et al., 2005). Since “similar items can clarify the meaning of the current question, eliminating superfluous meanings and sharpening the meaning held in common by the items” (Knowles & Byers, 1996, p. 1081), item-total correlations measuring the relationship between a single item and the total rating scale measure minus the item assessed¹⁵ may increase with the number of similar items being already answered earlier and thus, may improve the measurement of the underlying construct.

In the present study, scale sequence was not completely randomized but varied in a forward and backward item order. Since the rating scales applied

¹⁵ More correctly, an item-total correlation excluding the item assessed is generally indicated as the corrected item-total correlation. For the sake of simplicity, the designation item-total correlation was nevertheless used further.

in this study all measured multidimensional constructs, within-subscale item-total correlations were considered more appropriate than within-scale item-total correlations. Whereas in the former case only items of the same subscale are used for calculating item-total correlations, in the latter case, all items of a scale are considered.

8.5.7 *Item Missing Data*

Item nonresponse occurs when a respondent completes the survey but skips one or more questions within a questionnaire. In the present experiments, respondents could proceed to the next Web page without providing a substantive response as they were not explicitly prompted to enter a response. Thus, skipping one or more survey questions was of no consequence for the respondents. Furthermore, none of the rating scales tested here included an explicit nonsubstantive response option (e.g., ‘don’t know’, no opinion).

Item nonresponse rates can be calculated on the basis of the entire questionnaire with the sum of questions with missing values being divided by the total number of survey questions that should have been answered by a respondent. In addition, item nonresponse rates can be calculated on the basis of individual survey questions that comprise several components such as grid questions, for instance (Couper, et al., 2001; Heerwegh & Loosveldt, 2006; Vicente & Reis, 2010). In the present study, item nonresponse rates were calculated on the basis of rating scales and referred to the share of items with missing values divided by the total number of items in a rating scale.

Accordingly, the breakoff rate refers either to the share of respondents who dropped out of the survey among all respondents who have started the survey, or question-specifically, to the share of respondents who prematurely abandoned the questionnaire at a specific question among all respondents who reached and completed the question at which the breakoff occurred (AAPOR, 2011; Heerwegh & Loosveldt, 2006; Vicente & Reis, 2010). In the former case, survey breakoff can occur immediately after starting the survey, after answering a certain number of questions, or even towards the end of the questionnaire, whereas in the latter case, survey breakoff is predetermined by the position of the respective question within the questionnaire at which the breakoff occurred. In the present study, breakoff rates were reported for the respective rating scales assessed and thus, referred to the share of respondents who dropped out of the survey at the rating scale under investigation among

all respondents who reached and completed the rating scale at which the breakoff occurred.

8.5.8 Response Times

In Web surveys, response times can be collected automatically during the data collection process with a distinction being made between server-side and client-side paradata. Server-side response times are collected at screen level by simply recording visits to a Web page, whereas client-side response times are captured within a particular Web page enabling the logging of every user action (see also section 5.4). From this point on, all the server-side and client-side response times are indicated in seconds.

In the present experiments, outliers were excluded in accordance with the established criterion of excluding cases with values ± 2 standard deviations around the group mean. Prior to this standard method, extreme cases with unreasonably high response times were removed by excluding cases that exceeded the session timeout of 7,200 seconds due to interruptions on the target page comprising the rating scale under investigation. Differences in download speeds for radio button scales and drag-and-drop scales using client-side JavaScript programming proved to be rather small and thus, were not considered any further in the analysis of server-side and client-side response times.

In the Panel Survey 2012, the University Applicants Survey 2012, and the University Applicants Survey 2013, response times were gathered by means of standard server-side time stamps implemented in the Web survey environment by default. Server-side response times reflected the time span between loading and submitting of a single Web page. Since the experimental conditions tested here predominantly contained multiple-item-per-screen formats, server-side response times could be analyzed in terms of the total time for scale completion as well as in terms of the averaged item-response time. Whereas the total time for scale completion reflected the aggregated time spent on processing the total number of rating scale items presented either on a single screen or across several screens, averaged item-response times were calculated as the total time for scale completion divided by the number of rating scale items included in the calculation of the total time for scale completion.

In the University Applicants Survey 2013, response times were additionally gathered on the basis of client-side mouse event recordings which allowed for more detailed response time calculations in multiple-item-per-screen designs: (1) item-response times, (2) initial-reaction times, and (3) adjusted item-response times were examined on the basis of client-side response times. Exclusively for the drag-and-drop scales, (4) dragging times were reported.

Similar to server-side item-response times, client-side item-response times were calculated by means of dividing the total time for scale completion by the total number of respective rating scale items. Unlike server-side item-response times, however, client-side item-response times assessed the total time for scale completion by adding up single time spans for each action within a particular Web page. Thereby, the time span between the last answer and the click on the 'Continue' button was not included in client-side response time calculations.

Initial-reaction times reflected the time span between the loading of a Web page and the first respondent action, i.e., clicking a radio button in a radio button scale, or starting to drag a question-element or answer-element in a drag-and-drop scale. This distinction between the first and following actions was drawn because the initial time span between the loading of a Web page and a respondent's first action gave some indication of the time required for getting a first overview of the screen, for understanding the type and structure of the question, for reading and comprehending the general question request and response option labels, and particularly in a drag-and-drop scale, for reading and comprehending the specific data input instructions and navigational requirements. By contrast, for each of the following actions it could be assumed that response time was mainly allotted to the cognitive processing of the respective item rather than on surrounding information.

The calculation of adjusted item-response times took into account the increased requirements in regard to the time for initial orientation on the screen in the drag-and-drop scales compared to the radio button scales. The adjusted item-response time was assessed as the difference between the item-response time and the initial-reaction time. This was considered an attempt to subtract the share of time which was not primarily targeted at the cognitive processing of the item content but was rather spent on the processing of surrounding information. By implication, adjusted item-response times were

assumed to reflect the actual time for cognitive processing more precisely than overall item-response times including the time for initial orientation.

Finally, dragging times were separately captured for the drag-and-drop scales, indicating the time span between the start of dragging a question-element or answer-element to the desired position and dropping it to indicate an answer. Dragging times could be used to gain a better understanding of response behaviors in drag-and-drop scales, i.e., a respondent's navigational and cognitive processing of rating scale items presented in a drag-and-drop scale. More generally, this measure would give some indication of how the navigational and cognitive processing in drag-and-drop scales affected data accuracy of rating scale measures.

8.5.9 Respondent Evaluation

Strictly speaking, the respondent evaluation of a survey is no indicator of data accuracy. Nevertheless, respondent evaluation may serve as an indicator of the extent of the actual and perceived respondent burden. A semantic differential was used to assess the respondents' evaluation of different aspects of the survey. The scale was composed of 13 pairs of two bipolar adjectives that were assigned to three different components: (1) navigation with the adjective pairs (a) complicated to navigate – easy to navigate, (b) user-unfriendly – user-friendly, (c) unpractical to use – practical to use; (2) design with the adjective pairs (a) monotonous – diversified, (b) conventional – innovative, (c) ordinary – inventive; and (3) overall survey perception including, among others, the adjective pairs (a) boring – entertaining, (b) uninteresting – interesting, (c) unimportant – important. Respondents made their judgment on a 6-point semantic differential scale with negative adjectives being arranged on the left and positive adjectives on the right.

The findings regarding each of the direct and indirect data accuracy indicators outlined here are presented in Chapter 9, with special emphasis on a comparison of the two drag-and-drop rating scales with a conventional grid question. Possible implications of these findings will be discussed later in Chapter 10.

9. RESULTS

9.1 Item Nonresponse

9.1.1 Experiment 1

In Experiment 1, the incidence of item nonresponse was evaluated as a function of two (three) different scale lengths and five different scale formats in a 2 x 5 (3 x 5) between-subjects factorial design. Both drag-and-drop formats were expected to be accompanied by a higher risk of item nonresponse due to an increased extent of perceived and actual effort regarding their basic understanding and handling. With respect to scale length, item nonresponse was expected to increase with an increasing number of rating scale items. Accordingly, the 6-item scale would show the lowest risk of item nonresponse (in Experiments 1.2 and 1.3), followed by the 10-item scale, and lastly, the 16-item scale with the highest incidence of item nonresponse in Experiments 1.1, 1.2, and 1.3.

In order to create a proper basis of comparison for different scale lengths, the number of items with missing values was counted separately for the 6-item, 10-item, and 16-item scale. Total numbers of items with missing values were classified as follows: (1) no missing values, (2) partially missing values, and (3) completely missing values. Item nonresponse comprised the categories of partially and completely missing values, with the former referring to cases with missing values for at least one item to 5, 9, or 15 items depending on the respective scale length and the latter applying to cases with all items being missing values. The following analyses were based on Pearson's chi-squared tests with an alpha level of .05.

In Experiment 1.1, item nonresponse in terms of partially and completely missing values amounted to 8.8% across all the experimental conditions. By implication, 91.2% of the respondents provided substantive answers to all rating scale items (see Table 9). Overall, item nonresponse differed significantly as a function of scale format ($\chi^2(8, 771) = 68.45, p < .001$). As expected, the drag-response and drag-item format clearly carried the greatest risk of item nonresponse in terms of partially and completely missing values (22.0 and 13.4, respectively) as compared to the grid format (5.9), the one-vertical format (1.3), and the one-horizontal format (1.4). Despite a

strikingly sharp difference in the proportion of partially missing values between the drag-response format (18.1) and the drag-item format (8.3), this difference was not statistically significant. Separate analyses of the relationship between scale format and item nonresponse revealed largely the same results for the 10-item scale ($\chi^2(8, 381) = 17.87, p < .05$) and the 16-item scale ($\chi^2(8, 390) = 50.26, p < .001$).

Item nonresponse also varied significantly as a function of scale length ($\chi^2(2, 771) = 17.88, p < .001$) with the 16-item-scale expectably having a higher risk of item nonresponse (13.1) compared to the 10-item scale (4.5). Separate analyses of the relationship between scale length and item nonresponse revealed a significantly higher item nonresponse rate in the 16-item scale (19.0) compared to the 10-item scale (7.7) for the drag-response format ($\chi^2(2, 155) = 9.34, p < .01$). This effect was primarily due to a considerable increase in the number of partially missing values in the 16-item scale (25.3) compared to the 10-item scale (9.7). By contrast, differences for the remaining scale formats were non-significant (grid: $\chi^2(1, 153) = 2.88, ns$; drag-item: $\chi^2(2, 157) = 4.39, ns$; one-vertical: $\chi^2(1, 159) = 2.16, ns$; one-horizontal: $\chi^2(1, 147) = 0.00, ns$). Thus, prior expectations concerning an increase in item nonresponse in longer rating scales could be confirmed solely for the drag-response scale.

Table 9: Proportion of no missing, partially missing, and completely missing values (in %) depending on scale format and scale length (Experiment 1.1, $n = 771$)

Length		Format					Total
		(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 10	NM	97.4	88.9 ^{d,B}	92.3 ^B	100.0 ^b	98.6	95.5 ^B
	PM	2.6	9.7 ^{d,B}	5.1	0.0 ^b	1.4	3.7 ^B
	CM	0.0	1.4	2.6	0.0	0.0	0.8 ^B
(B) 16	NM	90.9 ^b	68.7 ^{a,d,e,A}	81.0 ^{d,e,A}	97.4 ^{b,c}	98.6 ^{b,c}	86.9 ^A
	PM	9.1	25.3 ^{d,e,A}	11.4	2.6 ^b	1.4 ^b	10.3 ^A
	CM	0.0	6.0	7.6	0.0	0.0	2.8 ^A
Total	NM	94.1 ^b	78.1 ^{a,d,e}	86.6 ^{d,e}	98.7 ^{b,c}	98.6 ^{b,c}	91.2
	PM	5.9 ^b	18.1 ^{a,d,e}	8.3 ^{d,e}	1.3 ^{b,c}	1.4 ^b	7.0
	CM	0.0 ^c	3.9	5.1 ^{a,d}	0.0 ^c	0.0	1.8

Note. NM = no missing values, PM = partially missing values, CM = completely missing values. Calculations were based on multiple Pearson's chi-squared tests (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the 10-item (A) and the 16-item scale (B). Deviations from 100% when adding the percentages of no missing, partially missing, and completely missing values are due to rounding errors.

In Experiment 1.2, the overall item nonresponse rate amounted to 9.8% (see Table 10). Pearson's chi-squared tests revealed a significant effect of scale format ($\chi^2(8, 5,262) = 382.02, p < .001$) and scale length ($\chi^2(4, 5,262) = 12.57, p < .05$). The expectations concerning the scale format could be confirmed since the drag-response and drag-item format had the highest item nonresponse rates in terms of partially and completely missing values (21.4 and 16.3, respectively) compared to the grid format (6.3), one-vertical format (2.0), and one-horizontal format (2.9). A significant difference between the drag-response and drag-item format was found in terms of a significantly higher proportion of partially missing values in the drag-response format (11.7) compared to the drag-item format (7.1). The effects were largely found in separate analyses of the 6-item scale ($\chi^2(8, 1,768) = 135.22, p < .001$), the 10-item scale ($\chi^2(8, 1,771) = 100.76, p < .001$), and the 16-item scale ($\chi^2(8, 1,723) = 169.38, p < .001$).

Concerning scale length, the proportion of partially missing values differed significantly depending on the number of rating scale items, with the 16-item scale comprising a higher risk of partially missing values (6.7)

compared to the 6-item scale (4.4). Hence, expectations concerning an increase in item nonresponse with longer rating scales could partially be confirmed. Again, separate analyses of the correlation between scale length and item nonresponse for each of the five scale formats revealed that the 16-item scale had a significantly higher proportion of partially missing values (17.6) compared to the 6-item scale (7.4) and the 10-item scale (10.4) when rating scale items were presented in the drag-response format ($\chi^2(4, 1,062) = 20.12, p < .001$). All the other scale formats remained unaffected by scale length (grid: $\chi^2(4, 1,055) = 4.61, ns$; drag-item: $\chi^2(4, 1,063) = 4.13, ns$; one-vertical: $\chi^2(2, 1,022) = 0.12, ns$; one-horizontal: $\chi^2(4, 1,060) = 5.25, ns$). Again, prior expectations concerning an increase in item nonresponse in longer rating scales could be confirmed solely for the drag-response scale.

Table 10: Proportion of no missing, partially missing, and completely missing values (in %) depending on scale format and scale length (Experiment 1.2, $n = 5,262$)

Length		Format					Total
		(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 6	NM	96.0 ^{b,c}	81.7 ^{a,d,e,C}	80.7 ^{a,d,e}	98.0 ^{b,c}	97.4 ^{b,c}	90.7
	PM	2.3 ^{b,c}	7.4 ^{a,d,e,C}	7.9 ^{a,d,e}	2.0 ^{b,c}	2.0 ^{b,c}	4.4 ^C
	CM	1.7 ^{b,c}	10.9 ^{a,d,e}	11.3 ^{a,d,e}	0.0 ^{b,c}	0.6 ^{b,c}	3.7
(B) 10	NM	92.8 ^{b,c,d}	81.1 ^{a,d,e,C}	85.4 ^{a,d,e}	98.2 ^{a,b,c}	96.7 ^{b,c}	90.8
	PM	4.7 ^b	10.4 ^{a,d,e,C}	7.0 ^d	1.8 ^{b,c}	3.3 ^b	5.5
	CM	2.5 ^{b,c,d,e}	8.5 ^{a,d,e}	7.6 ^{a,d,e}	0.0 ^{a,b,c}	0.0 ^{a,b,c}	3.7
(C) 16	NM	92.5 ^{b,c,d}	72.6 ^{a,c,d,e,A,B}	85.0 ^{a,b,d,e}	97.9 ^{a,b,c}	97.2 ^{b,c}	89.0
	PM	4.9 ^b	17.6 ^{a,c,d,e,A,B}	6.2 ^b	2.1 ^b	2.8 ^b	6.7 ^A
	CM	2.6 ^{b,c,d,e}	9.7 ^{a,d,e}	8.8 ^{a,d,e}	0.0 ^{a,b,c}	0.0 ^{a,b,c}	4.2
Total	NM	93.7 ^{b,c,d,e}	78.6 ^{a,c,d,e}	83.7 ^{a,b,d,e}	98.0 ^{a,b,c}	97.1 ^{a,b,c}	90.2
	PM	4.0 ^{b,c}	11.7 ^{a,c,d,e}	7.1 ^{a,b,d,e}	2.0 ^{b,c}	2.7 ^{b,c}	5.5
	CM	2.3 ^{b,c,d,e}	9.7 ^{a,d,e}	9.2 ^{a,d,e}	0.0 ^{a,b,c}	0.2 ^{a,b,c}	4.3

Note. NM = no missing values, PM = partially missing values, CM = completely missing values. Calculations were based on multiple Pearson's chi-squared tests (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale lengths, i.e., compared to the 6-item (A), 10-item (B), and 16-item scale (C). Deviations from 100% when adding the percentages of no missing, partially missing, and completely missing values are due to rounding errors.

In Experiment 1.3, item nonresponse amounted to 8.4% across all experimental conditions. By implication, 91.6% of the respondents completed the total set of rating scale items (see Table 11). Item nonresponse rates differed significantly as a function of scale format ($\chi^2(8, 5,896) = 381.75, p < .001$). As expected, the drag-response and drag-item formats clearly carried the greatest risk of item nonresponse in terms of partially and completely missing values (19.7 and 12.7, respectively) as compared to the grid format (4.5), the one-vertical format (2.8), and the one-horizontal format (2.6). Furthermore, the drag-response format showed significantly more item nonresponse compared to the drag-item format which was again primarily due to the high proportion of partially missing values in the drag-response format (14.9) compared to the drag-item format (7.7). When considering the relationship between scale format and item nonresponse separately for each of the three scale lengths, largely the same patterns were found within the 6-item scale ($\chi^2(8, 1,932) = 109.90, p < .001$), the 10-item scale ($\chi^2(8, 1,995) = 111.89, p < .001$), and the 16-item scale ($\chi^2(8, 1,969) = 201.07, p < .001$).

Furthermore, item nonresponse differed significantly depending on scale length ($\chi^2(4, 5,896) = 51.03, p < .001$). Consistent with prior expectations, the 16-item scale featured the highest proportion of item nonresponse (11.3) followed by the 10-item scale (7.6) and the 6-item scale (6.3) with further significant differences being indicated in Table 11. This significant increase in item nonresponse in longer rating scales was primarily due to the increasing number of cases with partially missing values. When considering the relationship between scale length and item nonresponse separately for each of the five scale formats, astonishingly, the drag-item scale was the only scale format which remained unaffected by scale length ($\chi^2(4, 1,153) = 2.35, ns$) since the proportion of item nonresponse amounted to nearly 13% in all three scale lengths. By contrast, the drag-response format showed a clear effect of scale length ($\chi^2(4, 1,184) = 51.49, p < .001$) with an approximately doubled proportion of item nonresponse in the 16-item scale (29.9) compared to the 6-item scale (13.4) and the 10-item scale (15.6). Once again, this was primarily due to a conceivable increase in partially missing values in the 16-item scale (23.6) compared to the 6-item scale (7.5) as well as the 10-item scale (13.3). Further significant differences for the grid format ($\chi^2(4, 1,189) = 13.10, p < .05$), the one-vertical format ($\chi^2(2, 1,197) = 10.36, p < .01$), and the one-horizontal format ($\chi^2(2, 1,173) = 8.10, p < .05$) are indicated in Table 11. Thus, the expectably higher item nonresponse rates in

longer rating scales were also found for all scale formats, except for the drag-item scale.

Table 11: Proportion of no missing, partially missing, and completely missing values (in %) depending on scale format and scale length (Experiment 1.3, $n = 5,896$)

Length		Format					Total
		(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 6	NM	97.5 ^{b,c}	86.6 ^{a,d,e,C}	87.2 ^{a,d,e}	98.2 ^{b,c,C}	99.2 ^{b,c,C}	93.7 ^C
	PM	2.3 ^{b,c}	7.5 ^{a,d,e,B,C}	7.0 ^{a,d,e}	1.8 ^{b,c,C}	0.8 ^{b,c,C}	3.9 ^{B,C}
	CM	0.3 ^{b,c}	5.9 ^{a,d,e,B}	5.7 ^{a,d,e}	0.0 ^{b,c}	0.0 ^{b,c}	2.4
(B) 10	NM	94.7 ^{b,c,d}	84.4 ^{a,d,e,C}	87.3 ^{a,d,e}	98.5 ^{a,b,c,C}	97.3 ^{b,c}	92.4 ^C
	PM	3.5 ^b	13.3 ^{a,d,e,A,C}	7.4 ^{d,e}	1.5 ^{b,c,C}	2.7 ^{b,c}	5.7 ^{A,C}
	CM	1.8	2.3 ^{d,e,A,C}	5.4 ^{d,e}	0.0 ^{b,c}	0.0 ^{b,c}	1.9
(C) 16	NM	94.5 ^{b,c}	70.1 ^{a,c,d,e,A,B}	87.3 ^{a,b,d,e}	95.1 ^{b,c,A,B}	96.0 ^{b,c,A}	88.7 ^{A,B}
	PM	5.3 ^b	23.6 ^{a,c,d,e,A,B}	8.8 ^b	4.9 ^{b,A,B}	4.0 ^{b,A}	9.3 ^{A,B}
	CM	0.3 ^{b,c}	6.3 ^{a,d,e,B}	3.9 ^{a,d,e}	0.0 ^{b,c}	0.0 ^{b,c}	2.0
Total	NM	95.5 ^{b,c}	80.3 ^{a,c,d,e}	87.3 ^{a,b,d,e}	97.2 ^{b,c}	97.4 ^{b,c}	91.6
	PM	3.7 ^{b,c}	14.9 ^{a,c,d,e}	7.7 ^{a,b,d,e}	2.8 ^{b,c}	2.6 ^{b,c}	6.3
	CM	0.8 ^{b,c}	4.8 ^{a,d,e}	5.0 ^{a,d,e}	0.0 ^{b,c}	0.0 ^{b,c}	2.1

Note. NM = no missing values, PM = partially missing values, CM = completely missing values. Calculations were based on multiple Pearson's chi-squared tests (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale lengths, i.e., compared to the 6-item (A), 10-item (B), and 16-item scale (C). Deviations from 100% when adding the percentages of no missing, partially missing, and completely missing values are due to rounding errors.

9.1.2 Experiment 2

In Experiment 2, item nonresponse was examined as a function of scale arrangement and scale format in a 2 x 5 between-subjects factorial design. As already hypothesized in previous experiments, both drag-and-drop scales were expected to provoke more item nonresponse compared to the grid format and the single-item-per-screen formats. With respect to scale arrangement, no significant differences in item nonresponse were expected depending on a positive-to-negative or negative-to-positive response option order. To ensure comparability with prior findings, the missing values were classified according to the categories of no missing, partially missing, and completely

missing values. Analyses were based on Pearson's chi-squared tests with significance being reported at an alpha level of .05.

In Experiment 2, the aggregated proportion of partially and completely missing values amounted to 5.4% across all experimental conditions, thus 94.7% of the respondents having no missing values (see Table 12). Item nonresponse varied significantly depending on scale format ($\chi^2(8, 768) = 39.32, p < .001$). Whereas the drag-response format differed significantly from the one-vertical format, the drag-item format was significantly different from the one-vertical and one-horizontal format. Both drag-and-drop formats revealed no significantly increased item nonresponse rate compared to the grid format. Thus, prior expectations of an increased susceptibility to item nonresponse in both drag-and-drop scales could only be partly confirmed. Separate analyses of the two scale arrangements revealed significant scale format effects on item nonresponse in both the positive-to-negative scale arrangement ($\chi^2(8, 398) = 27.91, p < .001$) and the negative-to-positive scale arrangement ($\chi^2(4, 370) = 12.28, p < .05$).

Furthermore, item nonresponse also differed significantly depending on the scale arrangement ($\chi^2(2, 768) = 8.49, p < .05$). Contrary to expectations, the positive-to-negative response option order had a significantly higher proportion of completely missing values (2.3) compared to the negative-to-positive response option order (0.0). This effect was largely due to the drag-item format ($\chi^2(2, 144) = 6.58, p < .05$) with a significantly lower proportion of no missing values in the positive-to-negative scale arrangement (83.7) compared to the negative-to-positive scale arrangement (96.6). Aside from this, no significant differences appeared when analyzing the relationship between scale arrangement and item nonresponse separately for the various scale formats (grid: $\chi^2(1, 144) = 0.77, ns$; drag-response: $\chi^2(2, 172) = 1.87, ns$; one-vertical: $\chi^2(1, 170) = 0.00, ns$; one-horizontal: $\chi^2(1, 138) = 1.04, ns$).

Although most pairwise comparisons between different scale formats and scale arrangements were non-significant in Experiment 2, it should nevertheless be noted that item nonresponse rates in both drag-and-drop formats were considerably higher than in the grid or in one of the single-item-per-screen formats. In the drag-response format, this was again due to a considerably higher proportion of partially missing values. In the drag-item format, increased item nonresponse rates were attributable to higher proportions of both partially and completely missing values which were particularly high in a positive-to-negative scale arrangement. A reason for the

lack of significant differences in Experiment 2 was certainly the smaller sample size, and thus, a smaller number of cases per experimental condition.

Table 12: Proportion of no missing, partially missing, and completely missing values (in %) depending on scale format and scale arrangement (Experiment 2, $n = 768$)

Arrangement	Format					Total	
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H		
(A) Pos-Neg	NM	98.5 ^c	92.2	83.7 ^{a,b}	98.9	97.1	93.7
	PM	1.5	6.7	8.1	1.1	1.5	4.0
	CM	0.0	1.1	8.1	0.0	1.5	2.3 ^B
(B) Neg-Pos	NM	96.1	89.0	96.6 ^A	98.8	98.6	95.7
	PM	3.9	11.0	3.4	1.2	1.4	4.3
	CM	0.0	0.0	0.0	0.0	0.0	0.0 ^A
Total	NM	97.2	90.7 ^d	88.9 ^{d,e}	98.8 ^{b,c}	97.8 ^c	94.7
	PM	2.8	8.7 ^d	6.3	1.2 ^b	1.4	4.2
	CM	0.0	0.6	4.9 ^d	0.0 ^c	0.7	1.2

Note. NM = no missing values, PM = partially missing values, CM = completely missing values. Calculations were based on multiple Pearson's chi-squared tests (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the positive-to-negative (A) and the negative-to-positive (B) scale arrangement. Deviations from 100% when adding the percentages of no missing, partially missing, and completely missing values are due to rounding errors.

9.1.3 Experiment 3

In Experiment 3, item nonresponse was again examined depending on scale arrangement and three different scale formats in a 2×3 between-subjects factorial design¹⁶. As in the previous experiments, both drag-and-drop formats were expected to yield higher item nonresponse rates compared to a standard grid format, whereas no differences were expected depending on a positive-to-negative or negative-to-positive scale arrangement. Again, missing values were classified according to the categories of no missing, partially missing,

¹⁶ Even though scale sequence was an inherent factor of Experiment 3 that was originally implemented as a $2 \times 2 \times 3$ between-subjects factorial design, the factor scale sequence was left out of subsequent analyses of item nonresponse since variations in scale sequence were not expected to have a systematic effect on the incidence of item nonresponse in rating scales.

and completely missing values and were analyzed with Pearson's chi-squared tests with an alpha level of .05.

In Experiment 3.1, item nonresponse reached 11.1% across all experimental conditions (see Table 13). In line with previous findings and prior expectations, the proportion of no missing, partially missing, and completely missing values varied significantly depending on scale format ($\chi^2(4, 5,865) = 221.88, p < .001$): The drag-response format (17.3) had a significantly higher risk of item nonresponse than the drag-item format (12.7), whereas both had a significantly higher item nonresponse rate compared to the grid format (3.1). Moreover, the drag-response format had a significantly higher item nonresponse rate compared to the drag-item format which was due to the significantly higher proportion of partially missing values in the drag-response format (11.4) compared to the drag-item format (6.5). Comparable differences between the scale formats were found when separately considering the positive-to-negative ($\chi^2(4, 2,918) = 91.44, p < .001$) and negative-to-positive scale arrangement ($\chi^2(4, 2,947) = 140.77, p < .001$).

As expected, no significant differences in item nonresponse rates were found depending on a positive-to-negative or negative-to-positive scale arrangement ($\chi^2(2, 5,865) = 0.43, ns$). Furthermore, no significant effects were found when considering the relationship between scale arrangement and item nonresponse separately for each of the three scale formats (grid: $\chi^2(2, 1,904) = 0.34, ns$; drag-response: $\chi^2(2, 1,998) = 3.84, ns$; drag-item: $\chi^2(2, 1,963) = 5.60, ns$).

Table 13: Proportion of no missing, partially missing, and completely missing values (in %) depending on scale format and scale arrangement (Experiment 3.1, $n = 5,865$)

Arrangement	Format			Total	
	(a) Grid	(b) Drag-R	(c) Drag-I		
(A) Pos-Neg	NM	96.8 ^{b,c}	84.4 ^a	86.3 ^a	89.1
	PM	2.2 ^{b,c}	10.1 ^a	7.8 ^{a,B}	6.8
	CM	1.0 ^{b,c}	5.5 ^a	5.9 ^a	4.2
(B) Neg-Pos	NM	96.9 ^{b,c}	81.2 ^{a,c}	88.4 ^{a,b}	88.6
	PM	2.4 ^{b,c}	12.7 ^{a,c}	5.2 ^{a,b,A}	6.9
	CM	0.7 ^{b,c}	6.2 ^a	6.5 ^a	4.5
Total	NM	96.8 ^{b,c}	82.7 ^{a,c}	87.3 ^{a,b}	88.8
	PM	2.3 ^{b,c}	11.4 ^{a,c}	6.5 ^{a,b}	6.8
	CM	0.8 ^{b,c}	5.9 ^a	6.2 ^a	4.3

Note. NM = no missing values, PM = partially missing values, CM = completely missing values. Calculations were based on multiple Pearson's chi-squared tests (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale formats, i.e., compared to grid (a), drag-response (b), and drag-item format (c). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the positive-to-negative (A) and the negative-to-positive (B) scale arrangement. Deviations from 100% when adding the percentages of no missing, partially missing, and completely missing values are due to rounding errors.

In Experiment 3.2, item nonresponse added up to 12.3% across all experimental conditions (see Table 14). The examination of the effects of the scale format and scale arrangement on the incidence of item nonresponse largely yielded the same results as in Experiment 3.1. Overall, the effect of the scale format on item nonresponse was significant ($\chi^2(4, 5,998) = 193.02$, $p < .001$). As expected, the drag-response format (17.7) had a significantly higher risk of item nonresponse compared to the drag-item format (14.8), whereas both scale formats significantly exceeded the item nonresponse rate of the grid format (4.5). This was due to higher proportions of both partially and completely missing values in the drag-and-drop formats. Again, the proportion of partially missing values in the drag-response format (10.7) was even higher than in the drag-item format (7.3). These differences between various scale formats were largely the same in separate analyses of the positive-to-negative ($\chi^2(4, 3,017) = 101.09$, $p < .001$) and the negative-to-positive scale arrangement ($\chi^2(4, 2,981) = 97.81$, $p < .001$). Also, as expected,

the effect of scale arrangement was non-significant ($\chi^2(2, 5,998) = 0.32, ns$). Separate analyses of the relationship of scale arrangement and item nonresponse for each of the three scale formats remained non-significant as well (grid: $\chi^2(2, 1,995) = 1.55, ns$; drag-response: $\chi^2(2, 1,994) = 2.59, ns$; drag-item: $\chi^2(2, 2,009) = 1.63, ns$).

Table 14: Proportion of no missing, partially missing, and completely missing values (in %) depending on scale format and scale arrangement (Experiment 3.2, $n = 5,998$)

Arrangement		Format			Total
		(a) Grid	(b) Drag-R	(c) Drag-I	
(A) Pos-Neg	NM	96.1 ^{b,c}	82.3 ^a	85.2 ^a	87.9
	PM	2.1 ^{b,c}	10.0 ^a	7.9 ^a	6.6
	CM	1.9 ^{b,c}	7.8 ^a	7.0 ^a	5.5
(B) Neg-Pos	NM	95.1 ^{b,c}	82.2 ^a	85.2 ^a	87.4
	PM	2.3 ^{b,c}	11.5 ^{a,c}	6.8 ^{a,b}	6.9
	CM	2.7 ^{b,c}	6.3 ^a	8.1 ^a	5.7
Total	NM	95.6 ^{b,c}	82.2 ^{a,c}	85.2 ^{a,b}	87.7
	PM	2.2 ^{b,c}	10.7 ^{a,c}	7.3 ^{a,b}	6.7
	CM	2.3 ^{b,c}	7.0 ^a	7.5 ^a	5.6

Note. NM = no missing values, PM = partially missing values, CM = completely missing values. Calculations were based on multiple Pearson's chi-squared tests (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale formats, i.e., compared to grid (a), drag-response (b), and drag-item format (c). Deviations from 100% when adding the percentages of no missing, partially missing, and completely missing values are due to rounding errors.

9.1.4 Summary

The present experiments consistently showed that both drag-and-drop scales displayed significantly increased item nonresponse rates compared to conventional rating scale formats, such as a grid or single-item-per-screen format. Furthermore, the drag-response scale showed a higher risk of item nonresponse compared to the drag-item scale. The drag-response format was particularly susceptible to high proportions of partially missing values. This higher risk of partially missing values in the drag-response scale was even more pronounced in longer rating scales. Other scale formats, however, were

mostly unaffected by scale length. The findings on scale arrangement were mostly non-significant.

On the evidence of this considerably increased risk of item nonresponse in both drag-and-drop scales, the differential item nonresponse potentially arising from differing rating scale formats was examined with respect to the respondents' gender and age, knowledge in dealing with computers and the Internet, and prior Web survey experience. Findings revealed that nonrespondents in both drag-and-drop scales did not differ significantly from nonrespondents in the grid, one-vertical, or one-horizontal scale in terms of their gender as well as computer and Internet knowledge as indicated by Pearson's chi-squared tests with an alpha level of .05. Non-significant differences were found in all six experiments. Similarly, variance analyses revealed no significant differences between nonrespondents in the various conditions of scale format in terms of age and prior Web survey experience in all six experiments¹⁷. Hence, although both drag-and-drop scales were generally more likely to provoke item nonresponse, analyses of several of the respondents' variables that might have been systematically related to the risk of skipping items in different scale formats suggested that nonrespondents in the drag-response and drag-item scale did not differ from nonrespondents in other scale formats.

9.2 Survey Breakoff

9.2.1 Experiment 1

In Experiment 1, the incidence of survey breakoff was evaluated as a function of two (three) different scale lengths and five different scale formats in a 2 x 5 (3 x 5) between-subjects factorial design. No significant effects of the scale

¹⁷ In Experiment 3.2, a significant main effect of age ($F(2, 696) = 6.08, p < .01, \eta^2 = .017$) was found with a Bonferroni post-hoc test indicating that among nonrespondents the mean age was slightly but significantly higher in the drag-item format (20.1) compared to the drag-response format (20.0, $p < .01$); concerning prior Web survey experience, there was a significant main effect in Experiment 3.1 ($F(2, 600) = 4.26, p < .05, \eta^2 = .014$) and Experiment 3.2 ($F(2, 670) = 4.30, p < .05, \eta^2 = .013$) with Bonferroni post-hoc tests indicating that among nonrespondents a respondent's prior Web survey experience was slightly but significantly higher in the drag-response format (2.15 and 2.12, respectively) compared to the drag-item format (1.94 and 1.89, $p < .05$, respectively).

format or scale length on the risk of survey breakoff were expected since respondents were allowed to skip any question and thus, could also skip an entire rating scale without being prompted to go back and complete the rating scale items.

In Experiment 1.1, the examination of the question-specific breakoff rates showed that only 0.9% of overall breakoff occurred in the rating scale assessed (see Table 15). Concerning breakoff rates depending on scale format and scale length, no significant differences were found between the experimental conditions for scale format ($\chi^2(4, 812) = 6.18, ns$) and scale length ($\chi^2(1, 812) = 3.31, ns$).

Nevertheless, it should be emphasized that the drag-item scale yielded a comparably high overall survey breakoff rate of 2.4%. It was also striking to see that the risk of survey breakoff was particularly high with 4.5% when the drag-item scale was used for a 16-item scale. The smaller number of cases per experimental condition might again be a probable reason for the lack of significant differences in Experiment 1.1.

Table 15: Breakoff rate (in %) depending on scale format and scale length (Experiment 1.1, $n = 812$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 10	0.0	1.4	0.0	0.0	0.0	0.3
(B) 16	1.3	0.0	4.5	1.3	0.0	1.4
Total	0.6	0.6	2.4	0.6	0.0	0.9

Note. Calculations were based on multiple Pearson's chi-squared tests (Bonferroni correction).

Breakoff rates were again at a low level with 1.1% in Experiment 1.2 (see Table 16). No significant effects were found depending on scale format ($\chi^2(4, 5,486) = 5.61, ns$) or scale length ($\chi^2(2, 5,486) = 4.07, ns$). Separate analyses of the various scale formats showed that there was a significantly higher breakoff rate in the 16-item scale (2.5) compared to the 6-item scale (0.3) when items were presented in a grid format ($\chi^2(2, 1,099) = 8.08, p < .05$). Similarly, the risk of survey breakoff tended to increase with increasing length of a drag-response scale. However, apart from the grid format, no significant effects were found for the remaining scale formats.

Table 16: Breakoff rate (in %) depending on scale format and scale length (Experiment 1.2, $n = 5,486$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 6	0.3 ^C	1.0	0.5	0.8	0.8	0.7
(B) 10	0.8	1.8	0.8	1.7	0.5	1.1
(C) 16	2.5 ^A	1.9	0.8	0.9	0.8	1.4
Total	1.2	1.6	0.7	1.1	0.7	1.1

Note. Calculations were based on multiple Pearson's chi-squared tests (Bonferroni correction): Uppercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale lengths, i.e., compared to the 6-item (A), 10-item (B), and 16-item scale (C).

In Experiment 1.3, again, the breakoff rates were low, at 0.6% (see Table 17). No significant effects were found depending on scale format ($\chi^2(4, 6,250) = 8.67, ns$). Concerning scale length, a significant effect of scale length was found ($\chi^2(2, 6,250) = 15.89, p < .05$) with the 16-item scale (1.9) featuring a higher breakoff rate than the 6-item scale (0.6) and 10-item scale (1.0). Separate analyses showed that this difference was primarily attributable to the one-horizontal format ($\chi^2(2, 1,236) = 6.10, p < .05$).

It was noteworthy that the drag-response and drag-item scales also tended to increase the risk of survey breakoff, particularly in the case of longer rating scales. However, the differences in survey breakoff rates between the drag-response and drag-item scales on the one hand, and a grid or single-item-per-screen design on the other failed to reach statistical significance.

Table 17: Breakoff rate (in %) depending on scale format and scale length (Experiment 1.3, $n = 6,250$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 6	0.2	0.7	1.0	0.2	0.8	0.6 ^C
(B) 10	1.7	1.7	0.9	0.5	0.2	1.0 ^C
(C) 16	0.9	3.0	2.3	1.4	1.9	1.9 ^{A,B}
Total	0.2	0.7	1.0	0.2	0.8	0.6

Note. Calculations were based on multiple Pearson's chi-squared tests (Bonferroni correction): Uppercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale lengths, i.e., compared to the 6-item (A), 10-item (B), and 16-item scale (C).

9.2.2 Experiment 2

In Experiment 2, survey breakoff was examined as a function of scale arrangement and scale format in a 2 x 5 between-subjects factorial design. Again, neither a significant main effect of scale format nor a significant main effect of scale arrangement was expected. While the breakoff rate of 1.3% could be considered very low, actually no significant effects of varying rating scale design emerged (see Table 18). As expected, no significant effects of scale format ($\chi^2(4, 833) = 1.55, ns$) or scale length ($\chi^2(1, 833) = 1.01, ns$) were found.

Table 18: Breakoff rate (in %) depending on scale format and scale arrangement (Experiment 2, $n = 833$)

Arrangement	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) Pos-Neg	0.0	1.0	2.1	0.0	1.4	0.9
(B) Neg-Pos	3.6	1.1	1.6	2.2	0.0	1.7
Total	1.9	1.1	1.9	1.1	0.7	1.3

Note. Calculations were based on multiple Pearson's chi-squared tests (Bonferroni correction).

9.2.3 Experiment 3

In Experiment 3, survey breakoff was examined depending on scale arrangement and scale formats in a 2 x 3 between-subjects factorial design¹⁸. No effects related to scale format and scale arrangement were expected. Breakoff rates were again at a low level of 0.7% in Experiment 3.1 (see Table 19) and 0.5% in Experiment 3.2 (see Table 20).

In Experiment 3.1, a significant effect of scale format ($\chi^2(2, 6,034) = 6.33, p < .05$) was found through the separate analyses of the two scale arrangements, indicating a significantly higher breakoff rate in the drag-response (1.2) and drag-item format (0.7) compared to the grid format (0.0) in case of a positive-to-negative scale arrangement ($\chi^2(2, 3,008) = 11.17, p < .01$). No significant differences were found in case of a negative-to-positive scale arrangement ($\chi^2(2, 3,026) = 1.80, ns$).

¹⁸ Even though scale sequence was an inherent factor of Experiment 3 that was originally implemented as a 2 x 2 x 3 between-subjects factorial design, the factor scale sequence was left out of subsequent analyses of survey breakoff since variations in scale sequence were not expected to have a systematic effect on the risk of survey breakoff in rating scales.

Furthermore, no significant overall effect on breakoff rates was found as a function of scale arrangement ($\chi^2(1, 6,034) = 0.09, ns$). However, separate analyses of the three scale formats revealed a significantly higher breakoff rate in the negative-to-positive (0.8) compared to the positive-to-negative scale arrangement (0.0) for the grid format ($\chi^2(1, 1,959) = 7.86, p < .01$). In the drag-response and drag-item formats, the opposite trend for different scale arrangements was observed, however, no significant differences were found neither in the drag-response ($\chi^2(1, 2,059) = 0.55, ns$) nor drag-item format ($\chi^2(1, 2,016) = 0.71, ns$).

Table 19: Breakoff rate (in %) depending on scale format and scale arrangement (Experiment 3.1, $n = 6,034$)

Arrangement	Format			Total
	(a) Grid	(b) Drag-R	(c) Drag-I	
(A) Pos-Neg	0.0 ^{b,c,B}	1.2 ^a	0.7 ^a	0.6
(B) Neg-Pos	0.8 ^A	0.9	0.4	0.7
Total	0.4	1.0	0.5	0.7

Note. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale formats, i.e., compared to the grid (a), drag-response (b), and drag-item format (c). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the positive-to-negative (A) and negative-to-positive (B) scale arrangement.

In Experiment 3.2, neither significant differences depending on scale format ($\chi^2(2, 6,147) = 0.12, ns$) and scale arrangement ($\chi^2(1, 6,147) = 1.39, ns$), nor any noticeable percentages were found.

Table 20: Breakoff rate (in %) depending on scale format and scale arrangement (Experiment 3.2, $n = 6,147$)

Arrangement	Format			Total
	(a) Grid	(b) Drag-R	(c) Drag-I	
(A) Pos-Neg	0.8	0.6	0.6	0.6
(B) Neg-Pos	0.4	0.5	0.4	0.4
Total	0.6	0.5	0.5	0.5

Note. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction).

9.2.4 Summary

Overall, prior expectations concerning survey breakoff could be confirmed since neither scale format nor scale length or scale arrangement had a significant effect on the risk of survey breakoff in the vast majority of cases. Only very few exceptions existed. Solely in Experiment 3.1, the breakoff rates of the drag-response and drag-item format exceeded the breakoff rate of the grid format. Hence, the findings provided consistent evidence supporting that both drag-and-drop scales did not present an increased risk of survey breakoff in the experiments at hand.

9.3 Scale Properties

9.3.1 Dimensionality

In all six experiments, multi-item measures were applied with a latent construct being captured by a set of several rating scale items. In a multidimensional measure, items can commonly be assigned to different scale components, which in turn jointly represent a latent construct. The replication of the underlying structure was considered a prerequisite for a meaningful interpretation of substantive survey data, and thus, for the basic suitability of a rating scale format. For this reason, scale multidimensionality was examined first, before the various systematic response tendencies were assessed and discussed with reference to their implications for data accuracy.

In order to examine scale multidimensionality in terms of whether different scale formats were equally suited to replicate the assumed latent structures¹⁹, a principal components analysis using Varimax rotation with Kaiser Normalization was performed. The percentage of variance explained by the principal components was reported as a first indication of the goodness of fit of the proposed component solution. Recoding of the reverse item scoring was required in Experiments 1 and 2, so that original and reversed items were all positively correlated. Cases with one or more missing values for the respective rating scale items were excluded from the analysis.

¹⁹ Despite certain modifications of the original multi-item scales for the purpose of scale construction in the present experiments (i.e., number of items per factor, changes in item wording at least for some of the original items), replication of the original factor structures was assumed in the present experiments.

It was presumed that differing scale formats would measure the same theoretical construct and could equally replicate the multidimensional structure of a rating scale. In fact, the three-dimensional structure of Otto and colleagues' (2001) scale on perceived emotional intelligence applied in Experiment 1.1 could equally be replicated by means of all five scale formats (see Appendix B, Table 53). Besides the analysis of the latent structure of the data, the percentage of variance explained by the three extracted components enabled an overall assessment of the quality of measurement provided by each of the scale formats. The percentage of explained variance (70.7% and above) suggested a good fit of the component solution irrespective of scale format. In Experiments 1.2 and 1.3 (see Appendix B, Table 54 and Table 55), the three-dimensional scale structure of Modick's (1977) scale on achievement motive could adequately be replicated irrespective of scale format. Again, percentages of explained variance (ranging between 68.8% and 70.1% in Experiment 1.2 and between 67.5% and 69.6% in Experiment 1.3) indicated still a good fit of the proposed model and were comparable for the five scale formats. Moreover, the two-dimensional structures of the reasons for social advancement scale used in Experiment 3.1, and the locus of control scale used in Experiment 3.2 could be equally obtained by each of the three scale formats applied (see Appendix B, Table 57 and Table 58). However, the percentages of explained variance did not exceed 49.7% in Experiment 3.1 and 49.5% in Experiment 3.2, suggesting a rather poor fit of the component solution irrespective of scale format. In Experiment 2, none of the five scale formats could account for the five-dimensional structure of the Ten-Item Personality Inventory (TIPI) scale (see Appendix B, Table 56). Thus, irrespective of scale format, the scale structure could not be replicated. Therefore, an insufficient applicability of the rating scale measure by itself could have been assumed.

In sum, it was observed that both drag-and-drop scales measured the same theoretical construct and could equally replicate a multidimensional scale structure as did the conventional grid or single-item-per-screen formats. Nevertheless, it should be pointed out that in some instances, the values of salient factor loadings considerably varied between different rating scale formats. However, each of the rating scale designs were equally affected by these variations and no consistent patterns could be found.

9.3.2 *Internal Consistency Reliability and Item Means*

An examination of the internal consistency reliability of a rating scale followed the principal component analysis in order to assess how well a set of rating scale items comprising a (sub)scale measures the latent construct (Viswanathan, 2005, p. 18). In general, scale reliability in terms of the relative absence of random errors in measurement is an important component of construct validity and is therefore considered essential in reducing measurement error (Krosnick, et al., 2005). Cronbach's alpha as a coefficient of the internal consistency of an entire rating scale was reported to indicate the reliability of a set of several rating scale items that were all presumed to measure the same underlying construct (Alwin, 2010).

Although a high Cronbach's alpha is basically intended in scale construction in terms of maximizing the proportion of variance attributable to common sources, increases in Cronbach's alpha may also carry the risk of "trivial redundancies and a narrow operationalisation of one subdomain of the overall construct" (Viswanathan, 2005, p. 26). Even more important in this study is the caveat that an increased internal consistency may also be an indication of systematic measurement error due to the respondents' failure to perceive distinctions in item meaning and to sufficiently differentiate between the rating scale items (Alwin, 2010; Callegaro, Shand-Lubbers, et al., 2009; Peytchev, 2006; Tourangeau, et al., 2004; Viswanathan, 2005, p. 110). In unbalanced rating scales, this may result in an arbitrary inflation of internal consistency. In counterbalanced rating scales, however, the internal consistency of a rating scale is expected to decrease because of the inclusion of reversed items and their higher susceptibility to response inconsistencies (Barnette, 2000; Callegaro, Shand-Lubbers, et al., 2009; Harrison & McLaughlin, 1993; Weijters, et al., 2009). Following Callegaro and colleagues' (2009) reasoning, response inconsistencies may be even more likely in grid questions because "there are more chances to 'miss' the meaning of the items" (p. 5895). Previous studies that systematically investigated the effect of presenting rating scale items in a grid or single-item-per-screen format on the internal consistency of a rating scale measure were inconclusive since most findings found no clear evidence of any differences between varying rating scale designs (see also section 5.1) (Callegaro, Shand-Lubbers, et al., 2009; Couper, et al., 2001; Toepoel, et al., 2008, 2009b; Toepoel & Dillman, 2008; Tourangeau, et al., 2004).

In the present study, the interpretation of the internal consistency reliability of a rating scale was further aggravated by the absence of the objective scale reliabilities with which the observed Cronbach's alpha values of the various scale formats could have been compared²⁰. Thus, an evaluation of differing scale formats based on Cronbach's alpha values was considered improper because the determination of minimum and maximum values of Cronbach's alpha would have been completely arbitrary. Instead, Cronbach's alpha values were rather indicated to assess the basic suitability of the drag-and-drop rating scales, in terms of ensuring satisfactory internal consistency and their basic comparability with conventional rating scales. Given the multidimensionality of a rating scale, the subscale internal consistency was considered more informative than overall scale internal consistency. In addition, item means were inspected for differences for varying scale formats. Even though different rating scale formats were expected to affect the accuracy of responses, they were still assumed to measure the same underlying construct. Hence, the substantive responses in terms of mean responses to single rating scale items should have been unaffected by scale format.

The recoding of reverse item scorings prior to analysis was required in Experiments 1 and 2. Cases with one or more missing values for the respective rating scale items were excluded from analysis. Concerning comparisons of item means in Experiments 1 and 2, hardly any significant differences were found depending on different scale formats²¹. In Experiment

²⁰ Although information on the internal consistency of (sub)scales used in the present experiments was mostly well documented, this information was unhelpful due to scale modifications made regarding the number and content of rating scale items.

²¹ Two-way ANOVAs revealed no significant main effects of scale length or significant interactions between scale format and scale length. Thus, for better clarity, illustration of item means was restricted to differences depending on scale format. The following exceptions could be observed: In Experiment 1.1, a significant main effect of scale length for item #6 ($F(1, 714) = 5.84, p < .05, \eta^2 = .001$) due to a significantly lower mean in the 10-item scale than in the 16-item scale; in Experiment 1.2, a significant main effect of scale length for item #5 ($F(2, 4,813) = 3.39, p < .05, \eta^2 = .001$) due to a significantly lower mean in the 6-item scale than in the 10-item scale ($p < .05$), a significant main effect of scale length for item #6 ($F(2, 4,813) = 9.51, p < .001, \eta^2 = .004$) with a lower mean in the 10-item scale than in the 16-item scale ($p < .001$), and a significant interaction between scale format and scale length for item #5 ($F(8, 4,813) = 3.88, p < .001, \eta^2 = .006$) concerning the one-vertical ($F(2, 1,009) = 10.43, p < .001, \eta^2 = .020$)

1.1 (see Appendix B, Table 59), findings of a one-way ANOVA comparing item means as a function of scale format revealed a significant main effect for item #7 ($F(4, 714) = 2.84, p < .05, \eta^2 = .016$) with a Bonferroni post-hoc test indicating a significantly higher mean in the drag-item format (3.74) compared to the one-vertical format (3.36, $p < .05$). A significant main effect for item #8 ($F(4, 714) = 6.14, p < .001, \eta^2 = .033$) was due to a significantly higher mean in the drag-item format (3.69) compared to the grid (3.25, $p < .01$), drag-response (3.30, $p < .05$), and one-vertical format (3.07, $p < .001$). In Experiment 1.2 (see Appendix B, Table 60), findings of a one-way ANOVA revealed a significant main effect of scale format for item #1 ($F(4, 4,813) = 5.58, p < .001, \eta^2 = .005$) with the drag-item format (1.84) yielding a lower mean compared to the grid (1.97, $p < .01$) and one-horizontal format (1.94, $p < .05$). Additionally, the drag-response format (1.86) had a significantly lower mean for item #1 than the grid format ($p < .05$). In Experiments 1.3 (see Appendix B, Table 61) and 2 (see Appendix B, Table 62), no significant differences in item means were found depending on different scale formats. In Experiment 3, however, mean responses depending on scale format revealed a strikingly different picture²²: In Experiment 3.1 (see Appendix B, Table 63), findings of a one-way ANOVA revealed a significant main effect of scale format for 6 out of 8 items²³. Significant differences were predominantly due to either significantly lower means in the drag-item format compared to the grid format for items of the first factor being about legitimate reasons for social advancement, or quite the contrary, significantly higher means for

and one-horizontal format ($F(2, 1,037) = 4.88, p < .01, \eta^2 = .009$); in Experiment 1.3, a significant main effect of scale length for item #5 ($F(2, 5,529) = 7.73, p < .001, \eta^2 = .003$) due to a significantly lower mean in the 6-item scale compared to the 10-item and 16-item scale ($p < .01$, respectively) and for item #6 ($F(2, 5,529) = 15.43, p < .001, \eta^2 = .006$) due to a significantly lower mean in the 10-item scale compared to the 6-item ($p < .01$) and 16-item scale ($p < .001$).

²² Three-way ANOVAs revealed significant main effects relating to scale arrangement and scale sequence as well as significant interactions with scale format. These differences in item means are discussed in further detail in section 9.8.2 and section 9.9. Therefore, illustration of item means was further restricted to differences depending on scale format.

²³ One-way ANOVAs revealed significant main effects for item #1 ($F(2, 5,211) = 8.05, p < .001, \eta^2 = .003$), item #3 ($F(2, 5,211) = 22.14, p < .001, \eta^2 = .008$), item #4 ($F(2, 5,211) = 11.57, p < .001, \eta^2 = .004$), item #5 ($F(2, 5,211) = 3.60, p < .05, \eta^2 = .001$), item #6 ($F(2, 5,211) = 30.90, p < .001, \eta^2 = .012$), and item #8 ($F(2, 5,211) = 9.03, p < .001, \eta^2 = .003$).

items of the second factor dealing with illegitimate reasons for social advancement. In Experiment 3.2 (see Appendix B, Table 64), all seven items showed a significant main effect of scale format²⁴. Again, similar patterns were found since the drag-item format yielded significantly lower means with items of the first factor measuring internal locus of control, and significantly higher means on items of the second factor dealing with external locus of control.

The inspection of subscale internal consistencies revealed no uniform patterns²⁵. With the exception of Experiment 1.3, both drag-and-drop scales tended to result in a lower Cronbach's alpha compared to the grid format. However, since an ideal Cronbach's alpha value, or at least its upper and lower limits, could not be determined properly in the present experiments, differences in Cronbach's alpha as a function of differing scale formats could not be considered an adequate indicator of a respondent's susceptibility to cognitive shortcuts in rating scales.

In sum, it could be stated that relating to substantive responses, each of the two drag-and-drop scales, the grid, and the single-item-per-screen designs yielded comparable results in Experiments 1 and 2. With only minor exceptions that did not indicate any clear pattern, no significant differences in mean responses were found as a function of scale format. In Experiment 3, however, differences in substantive responses were found in (almost) all rating scale items primarily between the drag-item and grid format. In Experiments 3.1 and 3.2, the drag-item format showed a stronger agreement with items that were generally answered affirmatively, whereas on the contrary, there was stronger disagreement with items primarily answered negatively. Differences in Cronbach's alpha emerging as a function of differing scale formats did not indicate any clear pattern. As aforementioned,

²⁴ One-way ANOVAs revealed significant main effects for item #1 ($F(2, 5,227) = 15.00, p < .001, \eta^2 = .006$), item #2 ($F(2, 5,227) = 4.01, p < .05, \eta^2 = .002$), item #3 ($F(2, 5,227) = 5.32, p < .01, \eta^2 = .002$), item #4 ($F(2, 5,227) = 16.79, p < .001, \eta^2 = .006$), item #5 ($F(2, 5,227) = 5.61, p < .01, \eta^2 = .002$), item #6 ($F(2, 5,227) = 6.43, p < .01, \eta^2 = .002$), and item #8 ($F(2, 5,227) = 18.75, p < .001, \eta^2 = .007$).

²⁵ It should be noted that in Experiment 2, subscale 5 showed a negative Cronbach's alpha in the grid format. However, this was no recoding error since the items #5 ('sympathetic, warm') and #7 ('critical, quarrelsome') were actually negatively correlated (after reversing item #5) even though the original conception of TIPI provided for a positive correlation.

Cronbach's alpha values were only suitable to a limited extent for assessing the respondents' susceptibility to cognitive shortcuts and response (in)consistency in (counter)balanced rating scales. Specific analysis of original and reversed item pairs and their susceptibility to response inconsistencies as a function of various scale formats would presumably be more appropriate presented in the next section.

9.4 Careless Responding

In Experiment 1, a respondent's tendency to careless responding was evaluated as a function of two (three) different scale lengths and five different scale formats in a 2 x 5 (3 x 5) between-subjects factorial design. Since both drag-and-drop rating scales were assumed to encourage the respondent's attention to the specific item content rather than relying on information provided by the item context, respondents using the drag-response or drag-item format might be more likely to notice the reverse wording of an item, and thus, would respond more consistently to original and reversed items as if rating scale items were presented in a grid format. Furthermore, in the drag-item format, enhanced attention should be given to the content of each single rating scale item, which is why the drag-item format was expected to be even more effective in preventing response inconsistencies between the original and reversed items than the drag-response format. Regarding the number of rating scale items, it was presumed that careless responding would increase with growing scale length, which is why response inconsistencies between the original and reversed items would also increase with a larger number of rating scale items.

The effect of scale format and scale length on the extent of careless responding was examined at item level in terms of item-pair correlations between items that both measured the same content but in reverse wording. Since careless responding would affect the strength of item-pair correlations negatively, the drag-item format was expected to yield stronger item-pair correlations than the drag-response format, whereas both drag-and-drop scales were expected to show stronger item-pair correlations than the grid format. Furthermore, item-pair correlations were supposed to diminish the more items were presented in a rating scale. This effect of scale length was expected to occur irrespective of the rating scale format. In order to evaluate whether the

risk of careless responding varied as a function of scale format and scale length, Fisher's z transformed correlations between an original item and its reverse equivalent were examined. The calculations were based on cases with substantive answers for all ten items that were part of both the 10-item and 16-item scale in Experiment 1.1, and for all six items that were part of the 6-item, 10-item, and 16-item scale in Experiments 1.2 and 1.3.

In Experiment 1.1, hardly any significant differences in item-pair correlations were found depending on scale format (see Table 21). Also, separate analyses of the 10-item and 16-item scale largely found no evidence of stronger item-pair correlations in the drag-and-drop scales compared to the grid format. Solely in the 16-item scale, a higher item-pair correlation was found for the drag-response format compared to the grid format for item-pair #1a (.624 versus .300 with Fisher's $z = 2.42$, $p < .05$). Also in the 16-item scale, the drag-item format showed significantly higher item-pair correlations compared to the grid format for item-pair #1a (.663 versus .300 with Fisher's $z = 2.80$, $p < .01$) and #2a (.781 versus .582 with Fisher's $z = 2.19$, $p < .05$). However, exactly the opposite was found in the 10-item scale, where the drag-item format showed significantly lower item-pair correlations than the grid format for item-pair #2a (.387 versus .651 with Fisher's $z = 2.18$, $p < .05$) and #2b (.308 versus .622 with Fisher's $z = 2.43$, $p < .05$). Thus, no persuasive evidence was found supporting prior expectations of higher item-pair correlations between the original and reversed items in the drag-and-drop scales compared to the grid format since there were no systematic differences in the extent of careless responding depending on different scale formats.

Furthermore, hardly any significant differences in item-pair correlations were found depending on scale length. Solely for item-pair #1b (.513 versus .644 with Fisher's $z = 5.61$, $p < .001$) and item-pair #2a (.556 versus .624 with Fisher's $z = 2.96$, $p < .01$), overall higher item-pair correlations were found in the 16-item scale compared to the 10-item scale. This finding was contrary to prior expectations. Based on separate analyses of the five scale formats, the drag-item format yielded higher item-pair correlations in the 16-item scale compared to the 10-item scale. However, a significant difference was found for item-pair #2a only (.387 versus .781 with Fisher's $z = 3.64$, $p < .001$). In the drag-response format, no significant differences in item-pair correlations were found depending on scale length.

Table 21: Item-pair correlations depending on scale format and scale length (Experiment 1.1, $n = 714$)

Item-Pair #	Length	Format					Total
		(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
1a	(A) 10	.554	.634	.533	.670	.583	.595
	(B) 16	.300 ^{b,c,d}	.624 ^a	.663 ^{a,e}	.730 ^{a,e}	.419 ^{c,d}	.556
	Total	.438 ^{b,d}	.633 ^a	.601	.697 ^{a,e}	.481 ^d	.574
1b	(A) 10	.570	.459 ^B	.518	.556	.476	.513 ^B
	(B) 16	.549	.717 ^A	.640	.700	.588	.644 ^A
	Total	.560	.588	.584	.631	.509	.580
2a	(A) 10	.651 ^c	.605	.387 ^{a,B}	.613	.549	.556 ^B
	(B) 16	.582 ^C	.620	.781 ^{a,e,A}	.700 ^e	.403 ^{c,d}	.624 ^A
	Total	.621	.611	.576	.653 ^e	.487 ^d	.589
2b	(A) 10	.622 ^c	.628 ^c	.308 ^{a,b,d,e}	.598 ^c	.713 ^c	.578
	(B) 16	.566 ^d	.504 ^d	.526 ^d	.759 ^{a,b,c,e}	.551 ^d	.574
	Total	.596 ^c	.562	.421 ^{a,d,e}	.655 ^c	.645 ^c	.577
3a	(A) 10	.472 ^e	.501	.460 ^e	.649	.711 ^{a,c}	.553
	(B) 16	.632	.632	.550	.561	.596	.585
	Total	.562	.565	.510	.609	.660	.570

Note. Item pair numeration corresponds with item labeling in Appendix A, Table 48. Calculations were based on multiple z-tests each comparing two independent Fisher's z transformed correlations (no alpha correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the 10-item (A) and the 16-item scale (B).

In Experiment 1.2, contrary to expectations, no significant differences in item-pair correlations were found between the drag-response and drag-item format on the one hand and the grid format on the other hand. The few significant differences occurring as a function of scale format provided no uniform conclusions (see Table 22). Consistent with the expectations, a significantly higher item-pair correlation was found in the 10-item scale compared to the 16-item scale for item-pair #1a (.398 versus .327 with Fisher's $z = 2.31$, $p < .05$). Also, for item-pair #3a, a significantly higher correlation was found in the 6-item scale (.330) compared to the 10-item scale (.264 with Fisher's $z = 2.06$, $p < .05$) and 16-item scale (.247 with Fisher's $z = 2.55$, $p < .05$). As shown by separate analyses of the five scale formats, the

significant differences depending on scale length were primarily due to the drag-response and one-vertical format.

Table 22: Item-pair correlations depending on scale format and scale length (Experiment 1.2, $n = 4,813$)

Item-Pair #	Length	Format					Total
		(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
1a	(A) 6	.418	.469 ^{e,C}	.365	.421 ^B	.324 ^b	.398 ^C
	(B) 10	.397	.455 ^{c,d}	.290 ^b	.288 ^{b,A}	.354	.352
	(C) 16	.318	.320 ^A	.401	.324	.277	.327 ^A
	Total	.379	.417 ^e	.350	.349	.319 ^b	.360
2a	(A) 6	.341	.400	.341	.393	.296 ^{B,C}	.357
	(B) 10	.306	.312	.391	.377	.428 ^A	.360
	(C) 16	.352	.309 ^{d,e}	.352	.451 ^b	.459 ^{b,A}	.384
	Total	.322	.345 ^d	.362	.411 ^b	.388	.366
3a	(A) 6	.401 ^{e,C}	.328 ^C	.325	.397 ^{B,C}	.202 ^a	.330 ^{B,C}
	(B) 10	.294	.235	.313	.233 ^A	.234	.264 ^A
	(C) 16	.259 ^A	.148 ^{c,A}	.319 ^b	.218 ^A	.256	.247 ^A
	Total	.316	.240	.318	.284	.277	.279

Note. Item pair numeration corresponds with item labeling in Appendix A, Table 49. Calculations were based on multiple z-tests each comparing two independent Fisher's z transformed correlations (no alpha correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale lengths, i.e., compared to the 6-item (A), 10-item (B), and 16-item scale (C).

In Experiment 1.3, hardly any significant differences in item-pair correlations depending on scale format were found, and no differences were found between the grid format on the one side and the drag-response and drag-item format on the other side (see Table 23). With only one exception, with item-pair #2a in the 10-item scale showing a significantly higher item-pair correlation when the scale was presented in the drag-item format compared to the grid format (.401 versus .254 with Fisher's $z = 2.23$, $p < .05$). Also, scale length had no effect on item-pair correlations, except in the drag-response format for item-pair #1a with a significantly higher item-pair correlation in the 6-item (.498) compared to the 10-item scale (.302 with Fisher's $z = 3.09$, $p < .01$) and 16-item scale (.342 with Fisher's $z = 2.46$, $p < .05$). Thus, contrary to expectations, neither scale format nor scale length had a major

impact on the extent of item-pair correlations between the original and reversed rating scale items.

Table 23: Item-pair correlations depending on scale format and scale length (Experiment 1.3, $n = 5,529$)

Item-Pair #	Length	Format					Total
		(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
1a	(A) 6	.428	.498 ^{d,e,B,C}	.398	.347 ^b	.334 ^b	.400
	(B) 10	.322	.302 ^{e,A}	.372	.420	.436 ^b	.372
	(C) 16	.318	.342 ^A	.403	.349	.394	.362
	Total	.357	.381	.394	.372	.388	.378
2a	(A) 6	.322	.280	.353	.304	.345	.322
	(B) 10	.254 ^c	.336	.401 ^{a,d}	.237 ^c	.328	.311
	(C) 16	.257 ^e	.255 ^e	.268 ^e	.271 ^e	.410 ^{a,b,c,d}	.288
	Total	.282 ^e	.289	.338	.272 ^e	.360 ^{a,d}	.308
3a	(A) 6	.254	.280	.344	.326	.295	.301
	(B) 10	.272	.312	.279	.294	.348	.303
	(C) 16	.245	.208 ^c	.359 ^b	.316	.270	.279
	Total	.256	.268	.330	.311	.305	.294

Note. Item pair numeration corresponds with item labeling in Appendix A, Table 49. Calculations were based on multiple z-tests each comparing two independent Fisher's z transformed correlations (no alpha correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale lengths, i.e., compared to the 6-item (A), 10-item (B), and 16-item scale (C).

In sum, in all three experiments, hardly any significant differences in item-pair correlations were found depending on scale format and scale length. Obviously, the risk of response inconsistencies due to a respondent's inattention to the reverse wording of rating scale items was largely unaffected by varying scale formats or differing scale lengths. Despite the fact that differences in item-pair correlations as a function of scale format and scale length were mostly non-significant, some of these differences were nevertheless striking. However, such noticeable differences were found for each of the experimental conditions and showed no consistent pattern. In addition, item-pair correlations in Experiments 1.2 and 1.3 could be considered rather low with correlation values less than .40 across all experimental conditions.

9.5 Nondifferentiated Responding

9.5.1 Experiment 1

In Experiment 1, a respondent's tendency to nondifferentiated responding was evaluated as a function of two (three) different scale lengths and five different scale formats in a 2×5 (3×5) between-subjects factorial design. Both drag-and-drop formats were expected to yield more differentiated responses in relation to the grid format since respondents would be more likely to perceive distinctions between the items and attend to the full range of possible response options. Moreover, respondents would be encouraged to reassess each response option in view of the current item, instead of simply adjusting the current answer to any previously given answers. With respect to the number of rating scale items, scale differentiation was expected to decrease with increasing scale length: The 10-item scale was assumed to show a higher degree of differentiation than the 16-item scale in all three experiments, whereas the 6-item scale in Experiments 1.2 and 1.3 was expected to show even higher scale differentiation. No significant interaction between scale format and scale length was expected.

In Experiment 1.1, only cases with substantive answers for all ten items of both the 10-item and 16-item scale were included. Since the differentiation index was calculated on the basis of ten items being rated on a 5-point Likert-type rating scale, the values could theoretically range from .00 to .80. In fact, the differentiation index varied between .00 and .80 with an average value of .676. Table 24 depicts varying degrees of differentiation depending on scale format and scale length. The findings of the ANOVA revealed a significant main effect of scale format ($F(4, 714) = 5.24, p < .001, \eta^2 = .029$). As expected, the drag-item format yielded a higher degree of differentiation compared to the grid format as well as compared to all the other formats (except the one-vertical): A Bonferroni post-hoc test revealed a significantly higher degree of differentiation in the drag-item format (.704) compared to the standard grid (.669, $p < .01$), the one-horizontal (.662, $p < .01$) as well as compared to the drag-response format (.668, $p < .01$). These differences in the degree of differentiation also applied when considering the relationship between scale format and scale differentiation separately for the 16-item scale ($F(4, 350) = 5.49, p < .001, \eta^2 = .060$) but were non-significant for the 10-item scale ($F(4, 364) = 1.66, ns$). Contrary to prior expectations, the drag-

response format did not differ neither from the grid format nor the single-item-per-screen formats ($p = 1.000$, respectively).

Furthermore, the main effect of scale length ($F(1, 714) = 0.07$, ns) and the two-way interaction between scale format and scale length were non-significant ($F(4, 714) = 2.12$, ns). Thus, contrary to expectations, the 10-item and 16-item scale did not differ in their degree of differentiation. Hence, scale length seemed to have no effect on scale differentiation. However, separate analyses of the relationship between scale length and scale differentiation for varying scale formats revealed a significantly and notably higher degree of differentiation in the 16-item scale (.718) compared to the 10-item scale (.691, $F(1, 136) = 6.35$, $p < .05$, $\eta^2 = .045$) when items were presented in the drag-item format.

Table 24: Differentiation index depending on scale format and scale length (Experiment 1.1, $n = 714$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 10	.679 (.08)	.663 (.08)	.691 ^B (.07)	.686 (.08)	.663 (.12)	.677 (.09)
(B) 16	.658 ^C (.11)	.674 ^C (.08)	.718 ^{a,b,d,e,A} (.05)	.665 ^C (.09)	.661 ^C (.08)	.674 (.09)
Total	.669 ^C (.10)	.668 ^C (.08)	.704 ^{a,b,e} (.06)	.676 (.08)	.662 ^C (.10)	.676 (.09)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the 10-item (A) and the 16-item scale (B).

In Experiment 1.2, the analysis was based on cases with substantive answers for all six items the 6-item, 10-item, and 16-item scale had in common. Given these six items that were answered on a 5-point Likert-type rating scale, the differentiation index could theoretically range from .00 to .78. Actual values ranged between this minimum and maximum value with an average differentiation index of .647. Findings on differentiation indices as a function of scale format and scale length are shown in Table 25. Results of an ANOVA revealed a significant main effect of scale format ($F(4, 4,813) = 18.25$, $p < .001$, $\eta^2 = .015$) and scale length ($F(2, 4,813) = 4.63$, $p < .05$, $\eta^2 =$

.002), whereas the two-way interaction between scale format and scale length was non-significant ($F(8, 4,813) = 0.87, ns$). A Bonferroni post-hoc test was used to assess the statistical significance of pair-wise comparisons. Concerning scale format, the differentiation index in the drag-item format (.668) differed significantly from all the other formats (grid: .640, $p < .001$; drag-response: .651, $p < .01$; one-vertical: .646, $p < .001$; one-horizontal: .636, $p < .001$). Hence, consistent with expectations, the highest degree of differentiation was obtained by using the drag-item format. By contrast, the drag-response format only obtained a significantly higher degree of differentiation compared to the one-horizontal format ($p < .01$) but—contrary to expectations—not compared to the grid ($p = .081$) or one-vertical format ($p = 1.000$). These differences in a respondent's degree of differentiation largely held true when considering the relationship between scale format and scale differentiation separately for the 6-item ($F(4, 1,603) = 8.69, p < .001, \eta^2 = .021$), the 10-item ($F(4, 1,632) = 5.03, p < .01, \eta^2 = .012$), and the 16-item scale ($F(4, 1,578) = 6.36, p < .001, \eta^2 = .016$).

Concerning scale length, the expectations regarding a decrease in scale differentiation with an increasing number of rating scale items could partially be confirmed since the 6-item scale (.652) yielded a significantly higher degree of differentiation compared to the 16-item scale (.643, $p < .05$). However, the effect size was close to zero and significant differences between the 6-item and 16-item scale disappeared when separately analyzing the relationship between scale length and scale differentiation for each of the five formats.

Table 25: Differentiation index depending on scale format and scale length (Experiment 1.2, $n = 4,813$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 6	.644 ^c (.10)	.660 (.08)	.676 ^{a,d,e} (.07)	.649 ^c (.09)	.638 ^c (.09)	.652 ^c (.09)
(B) 10	.639 ^c (.11)	.649 (.08)	.663 ^{a,e} (.08)	.652 (.08)	.634 ^c (.10)	.647 (.09)
(C) 16	.636 ^c (.09)	.643 ^c (.09)	.666 ^{a,b,d,e} (.08)	.636 ^c (.09)	.636 ^c (.10)	.643 ^A (.09)
Total	.640 ^c (.10)	.651 ^{c,e} (.08)	.668 ^{a,b,d,e} (.08)	.646 ^c (.09)	.636 ^{c,b} (.10)	.647 (.09)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale lengths, i.e., compared to the 6-item (A), 10-item (B), and 16-item scale (C).

In Experiment 1.3, the same rating scale was used as in Experiment 1.2 and an ANOVA was conducted including cases with substantive answers for all six items that were part of the 6-item, the 10-item, and the 16-item scale. The differentiation index theoretically ranged between .00 and .78. The differentiation index actually varied between this minimum and maximum value with an average of .650. The differentiation indices for each experimental condition are presented in Table 26. Findings indicated a significant main effect of scale format ($F(4, 5,529) = 17.32, p < .001, \eta^2 = .012$). A post-hoc test using Bonferroni correction revealed that, as expected, the drag-item format (.670) yielded a significantly higher degree of differentiation relative to the grid format (.639, $p < .001$) as well as compared to all the other formats (drag-response: .647, $p < .001$; one-vertical: .650, $p < .001$; one-horizontal: .649, $p < .001$). These differences in a respondent's degree of differentiation largely persisted when considering the relationship between scale format and scale differentiation separately for the 6-item scale ($F(4, 1,811) = 8.38, p < .001, \eta^2 = .018$), the 10-item scale ($F(4, 1,879) = 4.46, p < .01, \eta^2 = .009$), and the 16-item scale ($F(4, 1,839) = 6.47, p < .001, \eta^2 = .014$). Yet again, contrary to expectations, there was no significant difference between the drag-response format on the one hand and the grid (p

= .456), one-vertical ($p = 1.000$), and one-horizontal format ($p = 1.000$) on the other hand.

Concerning the extent of scale differentiation depending on scale length, neither the main effect of scale length ($F(2, 5,529) = 0.09, ns$) nor the two-way interaction between scale format and scale length reached statistical significance ($F(8, 5,529) = 1.10, ns$). Also, no significant differences were found when separately analyzing the relationship between scale length and scale differentiation for differing scale formats. Hence, expectations concerning a decrease in scale differentiation with increasing length of a rating scale could not be confirmed.

Table 26: Differentiation index depending on scale format and scale length (Experiment 1.3, $n = 5,529$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 6	.634 ^{c,d} (.10)	.650 ^c (.08)	.672 ^{a,b,e} (.08)	.655 ^a (.10)	.646 ^c (.09)	.651 (.09)
(B) 10	.642 ^c (.10)	.649 ^c (.09)	.668 ^{a,b,d,e} (.08)	.648 ^c (.08)	.646 ^c (.09)	.650 (.09)
(C) 16	.640 ^c (.10)	.641 ^c (.09)	.671 ^{a,b,d} (.09)	.646 ^c (.09)	.654 (.09)	.650 (.09)
Total	.639 ^{c,d} (.10)	.647 ^c (.09)	.670 ^{a,b,d,e} (.08)	.650 ^{a,c} (.09)	.649 ^c (.09)	.650 (.09)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e).

9.5.2 Experiment 2

In Experiment 2, the impact of scale arrangement and scale format on the degree of scale differentiation was examined in a 2 x 5 between-subjects factorial design. As already hypothesized in Experiment 1, both drag-and-drop formats were expected to yield a higher degree of differentiation than the grid format. Furthermore, the effect of categorial scale arrangement on a respondent's degree of differentiation was examined, with the negative-to-positive response option order expected to promote higher scale differentiation compared to the positive-to-negative response option order. A

significant two-way interaction between scale format and scale arrangement on scale differentiation was not expected.

Analysis included cases with substantive answers for all ten items the rating scale consisted of in Experiment 2. Based on these ten items using a 5-point Likert-type rating scale, the differentiation index could theoretically range from .00 to .80. In fact, the differentiation index varied between .32 and .80. Apparent from Table 27 indicating the degree of differentiation as a function of scale format and scale arrangement, respondents displayed an average differentiation index of .692. Findings of the ANOVA revealed a significant main effect of scale format ($F(4, 727) = 3.55, p < .01, \eta^2 = .019$). Consistent with prior expectations, a Bonferroni post-hoc test revealed a significantly higher degree of differentiation for the drag-item format (.711) compared to the grid (.683, $p < .01$) and the one-vertical format (.688, $p < .05$). Separate analyses of varying response option orders revealed that scale differentiation differed significantly depending on scale format in the negative-to-positive response option order ($F(4, 354) = 3.33, p < .05, \eta^2 = .037$), whereas the effect was non-significant in the positive-to-negative response option order ($F(4, 373) = 2.36, ns$). Again, the drag-response format (.690) failed to encourage a higher degree of differentiation compared to the grid format ($p = 1.000$) or the single-item-per-screen formats ($p = 1.000$, respectively).

Furthermore, neither the main effect of scale arrangement ($F(1, 727) = 0.67, ns$) nor the two-way interaction between scale format and scale arrangement was significant ($F(4, 727) = 1.98, ns$). Apparently, scale arrangement in terms of a positive-to-negative response option order or vice versa had no effect on the extent of nondifferentiated responding. Solely when considering the relationship between scale format and scale differentiation separately for the five scale formats, a significantly higher degree of differentiation was found in the negative-to-positive order (.702) compared to the positive-to-negative order for the one-horizontal format (.675, $F(1, 135) = 4.65, p < .05, \eta^2 = .034$). Thus, as opposed to initial expectations, the negative-to-positive response option order did not promote higher scale differentiation compared to the more conventional positive-to-negative response option order.

Table 27: Differentiation index depending on scale format and scale arrangement (Experiment 2, $n = 727$)

Arrangement	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) Pos-Neg	.689 (.06)	.695 (.60)	.708 ^e (.06)	.684 (.07)	.675 ^{c,B} (.09)	.690 (.07)
(B) Neg-Pos	.678 ^c (.07)	.685 (.07)	.714 ^a (.05)	.693 (.06)	.702 ^A (.05)	.693 (.06)
Total	.683 ^c (.07)	.690 (.06)	.711 ^{a,d} (.06)	.688 ^c (.06)	.689 (.07)	.692 (.07)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the positive-to-negative (A) and negative-to-positive (B) scale arrangement.

9.5.3 Experiment 3

In Experiment 3, a respondent's degree of differentiation was evaluated depending on scale arrangement and scale format in terms of a 2 x 3 between-subjects factorial design²⁶. With respect to the grid, drag-item, and drag-response format, it was again hypothesized that both drag-and-drop formats would yield a higher degree of differentiation compared to the grid format. As already suggested in Experiment 2, the negative-to-positive response option order was expected to yield a higher scale differentiation as the positive-to-negative response option order. A significant two-way interaction between scale format and scale arrangement on the extent of scale differentiation was not expected.

The assumptions were examined in two independent experiments. In Experiments 3.1 and 3.2, a two-way ANOVA was conducted solely including cases which had substantive answers for all eight rating scale items,

²⁶ Even though scale sequence was an inherent factor of Experiment 3 that was originally implemented as a 2 x 2 x 3 between-subjects factorial design, the factor scale sequence was left out of subsequent analyses of scale differentiation since variations in scale sequence were not expected to have a systematic effect on a respondent's degree of differentiation in rating scales.

respectively. Considering these eight items using a 5-point Likert-type rating scale²⁷, differentiation index could theoretically range between .00 and .78.

In Experiment 3.1, the differentiation index actually ranged between .00 and .78 with an average value of .698 (see Table 28). The main effect of scale format reached statistical significance ($F(2, 5,211) = 54.25, p < .001, \eta^2 = .020$) with a Bonferroni post-hoc test revealing a significantly higher degree of differentiation in the drag-item format (.712) compared to the drag-response (.696, $p < .001$) and the grid format (.687, $p < .001$). Contrary to the findings in Experiments 1 and 2 but in line with initial expectations, the drag-response format yielded a slightly but significantly higher scale differentiation than the grid format ($p < .001$). These differences were also found in separate analyses of the relationship between scale format and scale differentiation for a positive-to-negative response option order ($F(2, 2,599) = 31.53, p < .001, \eta^2 = .024$) as well as for a negative-to-positive response option order ($F(2, 2,612) = 22.86, p < .001, \eta^2 = .017$).

As indicated by a significant main effect of scale arrangement ($F(1, 5,211) = 11.15, p < .01, \eta^2 = .002$), predictions could be confirmed because a negative-to-positive response option order (.701) actually yielded a higher scale differentiation compared to the positive-to-negative response option order (.695). Separate analyses of the relationship between scale arrangement and scale differentiation for various scale formats revealed a significant difference for the grid format ($F(1, 1,844) = 7.30, p < .01, \eta^2 = .004$) but a non-significant difference for the drag-response ($F(1, 1,653) = 2.87, ns$) and drag-item format ($F(1, 1,714) = 1.76, ns$). Thus, scale differentiation in both drag-and-drop scales seemed to be less affected by scale arrangement. Furthermore, as previously hypothesized, the two-way interaction between scale format and scale arrangement was non-significant ($F(2, 5,211) = 0.74, ns$).

²⁷ Although in Study 3.2 item #7 was removed from analysis due to a negative correlation with the actual factor 'internal locus of control', the theoretical range of the differentiation index remained unchanged.

Table 28: Differentiation index depending on scale format and scale arrangement (Experiment 3.1, $n = 5,211$)

Arrangement	Format			Total
	(a) Grid	(b) Drag-R	(c) Drag-I	
(A) Pos-Neg	.682 ^{b,c,B} (.08)	.693 ^{a,c} (.08)	.710 ^{a,b} (.06)	.695 ^B (.07)
(B) Neg-Pos	.692 ^{c,A} (.07)	.700 ^c (.07)	.714 ^{a,b} (.06)	.701 ^A (.07)
Total	.687 ^{b,c} (.08)	.696 ^{a,c} (.07)	.712 ^{a,b} (.06)	.698 (.07)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale formats, i.e., compared to grid (a), drag-response (b), and drag-item format (c). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the positive-to-negative (A) and negative-to-positive (B) scale arrangement.

In Experiment 3.2, differentiation index actually ranged between .00 and .78 with an overall mean of .669 (see Table 29). The main effect of scale format reached statistical significance ($F(2, 5,227) = 31.85, p < .001, \eta^2 = .012$) with a Bonferroni post-hoc test again supporting the theoretical prediction of a significantly higher degree of differentiation in the drag-item format (.682) compared to the grid format (.662, $p < .001$) as well as compared to the drag-response format (.664, $p < .001$). In line with the findings of Experiments 1 and 2 but contrary to prior expectations, the drag-response format did not differ significantly from the grid format ($p = 1.000$). Separate analyses of the two scale arrangements revealed the same differences in the positive-to-negative ($F(2, 2,639) = 10.35, p < .001, \eta^2 = .008$) and negative-to-positive response option order ($F(2, 2,588) = 23.30, p < .001, \eta^2 = .018$).

Contrary to expectations, there was no significant main effect of scale arrangement ($F(1, 5,227) = 1.05, ns$) which means that the degree of differentiation remained unaffected by variations in the response option order from positive-to-negative to negative-to-positive. This was also confirmed by separate analyses of the relationship between scale arrangement and scale differentiation for the grid ($F(1, 1,899) = 0.33, ns$), drag-response ($F(1, 1,616) = 0.42, ns$), and drag-item format ($F(1, 1,711) = 3.22, ns$). As expected, the two-way interaction between scale format and scale arrangement was non-significant ($F(2, 5,227) = 1.33, ns$).

Table 29: Differentiation index depending on scale format and scale arrangement (Experiment 3.2, $n = 5,227$)

Arrangement	Format			Total
	(a) Grid	(b) Drag-R	(c) Drag-I	
(A) Pos-Neg	.663 ^c (.09)	.663 ^c (.09)	.679 ^{a,b} (.08)	.668 (.08)
(B) Neg-Pos	.660 ^c (.09)	.665 ^c (.08)	.685 ^{a,b} (.07)	.670 (.08)
Total	.662 ^c (.09)	.664 ^c (.08)	.682 ^{a,b} (.08)	.669 (.08)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale formats, i.e., compared to grid (a), drag-response (b), and drag-item format (c).

9.5.4 Summary

In sum, findings of Experiments 1, 2, and 3 consistently showed that different rating scale formats induced differing scale differentiation. As initially proposed, a significantly higher degree of differentiation was unexceptionally found in the drag-item format compared to the grid format. With only few exceptions, the drag-item format also showed a significantly higher degree of differentiation compared to the drag-response, one-vertical, and one-horizontal format. This increase in scale differentiation applied irrespective of the length of a rating scale. However, against prior expectations, the drag-response format did not promote a higher degree of differentiation compared to a standard grid (except in Experiment 3.1) or the single-item-per-screen formats (except compared to the one-horizontal format in Experiment 1.2). With respect to the number of rating scale items and their impact on scale differentiation, the results were inconclusive: Whereas in Experiments 1.1 and 1.3, no evidence of an increasing scale differentiation with larger scale lengths was found, Experiment 1.2 revealed a significantly higher degree of differentiation in the shortest scale with 6 items compared to the longest scale comprising 16 items. Also, results regarding the effect of scale arrangement on scale differentiation were mixed: Experiments 2 and 3.2 revealed no significant effects depending on variations in scale arrangement, whereas Experiment 3.1 provided evidence of higher scale differentiation when

response options were arranged in a negative-to-positive response option order.

9.6 Acquiescent Responding

In Experiment 1, a respondent's tendency to acquiescent responding was evaluated as a function of two (three) different scale lengths and five different scale formats in a 2 x 5 (3 x 5) between-subjects factorial design. Acquiescent responding refers to a respondent's tendency to agree rather than disagree with a rating scale item irrespective of its content, which in turn is strongly determined by a respondent's disposition to strive for compliance with the "social convention to be polite" (Krosnick, 1999, p. 554). Therefore, previous research considered characteristics of the method to be of secondary importance, and hence, no variations in the degree of acquiescent responding were expected depending on scale format in Experiment 1. Equally, neither a significant effect of scale length nor a significant interaction between scale format and scale length was expected.

In Experiment 1.1, an ANOVA was conducted including cases with substantive answers for all ten items that were part of both the 10-item and 16-item scale. The acquiescence index actually ranged from -.60 to .50 with an overall mean of -.041 indicating a slight tendency to disagree rather than agree with the rating scale items (see Table 30). Findings of the ANOVA revealed neither a significant main effect of scale format ($F(4, 714) = 1.36, ns$) nor a significant main effect of scale length ($F(1, 714) = 1.81, ns$), or a significant two-way interaction between scale format and scale length ($F(4, 714) = 1.51, ns$). Thus, as expected, the drag-response and drag-item format did not differ from either the grid or the single-item-per-screen formats in terms of acquiescent responding. Also, consistent with expectations, the extent of acquiescent responding was unaffected by scale length. Solely one exception was noticeable. Despite the lack of a significant main effect of scale length, separate analyses of the relationship between scale format and acquiescence revealed a significant difference between the 10-item (-.063) and 16-item scale (.022) for the drag-item format ($F(1, 136) = 6.91, p < .05, \eta^2 = .049$).

Table 30: Acquiescence index depending on scale format and scale length (Experiment 1.1, $n = 714$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 10	-.064 (.19)	-.056 (.19)	-.063 ^B (.20)	-.037 (.15)	-.031 (.18)	-.049 (.18)
(B) 16	-.068 ^C (.19)	-.045 (.18)	.022 ^{B,A} (.17)	-.032 (.16)	-.036 (.19)	-.033 (.18)
Total	-.066 (.19)	-.051 (.18)	-.023 (.19)	-.034 (.16)	-.033 (.18)	-.041 (.18)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the 10-item (A) and the 16-item scale (B).

In Experiment 1.2, the analysis was based on cases with substantive answers for all six items that were part of the 6-item, 10-item and 16-item scale. The acquiescence index actually ranged from $-.83$ to 1.00 with an overall mean of $-.046$ indicating a slight tendency to disagree (see Table 31). Findings of the ANOVA indicated a significant main effect of scale format ($F(4, 4,813) = 3.61, p < .01, \eta^2 = .003$) with a Bonferroni post-hoc test revealing a significantly different acquiescence index in the one-vertical format ($-.028$) compared to the one-horizontal format ($-.062, p < .001$). Consistent with prior expectations, both drag-and-drop scales did not differ significantly from all the other scale formats. When considering the relationship between scale format and acquiescence separately for the three scale lengths, a significant effect was found for the 10-item scale ($F(4, 1,632) = 4.29, p < .01, \eta^2 = .010$) with the drag-item format ($-.019$) being significantly different from the one-horizontal format ($-.062, p < .01$), whereas the effects for the 6-item ($F(4, 1,603) = 2.11, ns$) and 16-item scale ($F(4, 1,578) = 1.32, ns$) were non-significant.

Furthermore, a significant main effect of scale length ($F(2, 4,813) = 3.24, p < .05, \eta^2 = .001$) was found with a Bonferroni post-hoc test revealing a significantly stronger tendency to disagree in the 10-item scale ($-.059$) compared to the 16-item scale ($-.038, p < .05$). Again, considering the relationship between scale length and acquiescence separately for the five scale formats, a significant effect was found for the one-vertical format ($F(2,$

1,037) = 6.53, $p < .01$, $\eta^2 = .012$) with the 6-item scale (-.029) being significantly different from the 10-item scale (-.097, $p < .01$). Contrary to expectations, the two-way interaction between scale format and scale length was significant ($F(8, 4,813) = 1.96$, $p < .05$, $\eta^2 = .003$). This significant interaction in turn showed that a significant overall effect of scale length was principally due to the one-vertical format with its significant difference between the 6-item and 10-item scale, whereas no significant differences were found in the remaining scale formats.

Table 31: Acquiescence index depending on scale format and scale length (Experiment 1.2, $n = 4,813$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 6	-.055 (.27)	-.056 (.21)	-.050 (.23)	-.011 (.25)	-.029 ^B (.26)	-.039 (.25)
(B) 10	-.053 (.26)	-.068 (.24)	-.019 ^e (.24)	-.054 (.25)	-.097 ^{C,A} (.24)	-.059 ^C (.25)
(C) 16	-.044 (.25)	-.048 (.26)	-.023 (.25)	-.019 (.26)	-.057 (.24)	-.038 ^B (.25)
Total	-.051 (.26)	-.058 (.24)	-.031 (.24)	-.028 ^e (.25)	-.062 ^d (.25)	-.046 (.25)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale lengths, i.e., compared to the 6-item (A), 10-item (B), and 16-item scale (C).

In Experiment 1.3, the acquiescence index actually ranged from -.83 to 1.00. An overall mean of -.055 indicated quite balanced responses with a slight tendency to disagree rather than agree with the rating scale items (see Table 32). The two-way ANOVA indicated a non-significant main effect of scale format ($F(4, 5,529) = 1.86$, ns). Thus, a respondent's extent of acquiescent responding was unaffected by differing scale formats. In this respect, prior expectations could be confirmed. A significant main effect of scale length ($F(2, 5,529) = 4.59$, $p < .05$, $\eta^2 = .002$) was due to a slightly but significantly higher tendency to disagree in the 10-item scale (-.068) compared to the 6-item scale (-.042, $p < .01$). However, when considering the relationship between scale length and acquiescence separately for the five different scale

formats, no significant differences were found for the grid ($F(2, 1,155) = .55, ns$), drag-response ($F(2, 1,036) = 1.38, ns$), drag-item ($F(2, 1,006) = 2.44, ns$), one-vertical ($F(2, 1,173) = 2.90, ns$), and one-horizontal format ($F(2, 1,159) = 1.75, ns$). Thus, prior expectations were deemed confirmed despite a significant overall effect of scale length since the effect size of the overall effect was close to zero and significance disappeared with separate analyses of each scale format. As expected, the two-way interaction between scale format and scale length was non-significant ($F(8, 5,529) = 1.00, ns$).

Table 32: Acquiescence index depending on scale format and scale length (Experiment 1.3, $n = 5,529$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 6	-.044 (.27)	-.055 (.25)	-.052 (.24)	-.029 (.25)	-.033 (.26)	-.042 ^B (.25)
(B) 10	-.063 (.26)	-.085 (.24)	-.058 (.24)	-.070 (.23)	-.063 (.24)	-.068 ^A (.24)
(C) 16	-.060 (.28)	-.076 ^C (.26)	-.018 ^b (.26)	-.061 (.25)	-.061 (.23)	-.056 (.26)
Total	-.055 (.27)	-.072 (.25)	-.043 (.25)	-.054 (.25)	-.052 (.24)	-.055 (.25)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale lengths, i.e., compared to the 6-item (A), 10-item (B), and 16-item scale (C).

Based on the findings of Experiment 1, prior expectations could be confirmed as a respondent's tendency to agree rather than disagree with a rating scale item occurred irrespective of the format and length of a rating scale. Findings on acquiescent responding in Experiment 2 were not reported here because a meaningful interpretation of the acquiescence index requires the use of a counterbalanced rating scale. This, however, was not the case in the rating scale used in Experiment 2 as analyses of scale multidimensionality and internal consistency reliability have shown. For this reason, the meaningfulness of the results concerning acquiescent responding in Experiment 2 was considered limited.

9.7 Extreme Responding

9.7.1 Experiment 1

In Experiment 1, a respondent's tendency to extreme responding was evaluated as a function of two (three) different scale lengths and five different scale formats in a 2 x 5 (3 x 5) between-subjects factorial design. To recall the definition, extreme responding is regarded as a respondent's tendency to select the most extreme response options irrespective of the item content. The extent of extreme responding is strongly determined by characteristics of the respondents and individual differences in interpreting and mapping a judgment onto the response options. Similar to acquiescent responding, the visual presentation of response options is considered of secondary importance in view of the fact that the wording, number, etc., of response options remained unchanged (Paulhus, 1991). Therefore, the degree of extreme responding was deemed to be unaffected by differences in scale format and length. Furthermore, no significant interaction between both factors was expected.

In Experiment 1.1, an ANOVA was conducted including cases with substantive answers for all ten items that were part of both the 10-item and 16-item scale. The extremity index actually ranged from .00 to 1.00 with an overall mean of .209 (see Table 33). Findings of the ANOVA revealed neither a significant main effect of scale format ($F(4, 714) = 1.83, ns$), nor a significant main effect of scale length ($F(1, 714) = 1.60, ns$) or a significant two-way interaction between scale format and scale length ($F(4, 714) = 2.15, ns$). Thus, consistent with prior expectations, the degree of extreme responding did not vary depending on the use of different scale formats or varying scale lengths.

Table 33: Extremity index depending on scale format and scale length (Experiment 1.1, $n = 714$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 10	.186 (.18)	.184 (.22)	.206 (.17)	.211 (.19)	.214 (.20)	.201 (.19)
(B) 16	.216 (.22)	.213 (.20)	.295 (.19)	.195 (.20)	.177 (.20)	.217 (.21)
Total	.201 (.20)	.198 (.21)	.248 (.18)	.203 (.20)	.195 (.20)	.209 (.20)

Note. Standard deviations are indicated in parentheses.

In Experiment 1.2, the analysis included cases with substantive answers for all six items that were part of the 6-item, 10-item, and 16-item scale. Extremity index actually ranged from .00 to 1.00 with an overall mean of .194 (see Table 34). Results of an ANOVA revealed a significant main effect of scale format ($F(4, 4,813) = 11.22, p < .001, \eta^2 = .009$). As indicated by a Bonferroni post-hoc test, a respondent's tendency to extreme responding was significantly higher in the drag-item format (.233) compared to all the other formats (grid: .187, $p < .001$; drag-response: .190, $p < .001$; one-vertical: .185, $p < .001$; one-horizontal: .179, $p < .001$). This higher tendency to extreme responding in the drag-item format became fully apparent when considering the relationship between scale format and extremity in the 6-item scale ($F(4, 1,603) = 4.72, p < .01, \eta^2 = .011$), whereas significant differences partly disappeared in the 10-item ($F(4, 1,632) = 3.93, p < .01, \eta^2 = .010$) and 16-item scale ($F(4, 1,578) = 4.29, p < .01, \eta^2 = .011$). Nonetheless, prior expectations could not be confirmed since the extent of extreme responding actually differed depending on different scale formats. By contrast, as expected, the extremity index was unaffected by the number of rating scale items as indicated by a non-significant main effect of scale length ($F(2, 4,813) = 0.05, ns$). The two-way interaction between scale format and scale length was non-significant either ($F(8, 4,813) = 0.65, ns$).

Table 34: Extremity index depending on scale format and scale length (Experiment 1.2, $n = 4,813$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 6	.182 ^c (.19)	.190 ^c (.19)	.236 ^{a,b,d,e} (.20)	.185 ^c (.20)	.179 ^c (.19)	.193 (.20)
(B) 10	.186 ^c (.19)	.190 (.21)	.231 ^{a,e} (.20)	.200 (.19)	.173 ^c (.19)	.195 (.20)
(C) 16	.194 (.21)	.189 (.18)	.233 ^{d,e} (.20)	.170 ^c (.18)	.185 ^c (.20)	.193 (.20)
Total	.187 ^c (.20)	.190 ^c (.19)	.233 ^{a,b,d,e} (.20)	.185 ^c (.19)	.179 ^c (.20)	.194 (.20)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e).

In Experiment 1.3, the calculation of the extremity index included cases with substantive answers for all six items the 6-item, 10-item and 16-item scale had in common. The extremity index actually ranged from .00 to 1.00 with an overall mean of .185 (see Table 35). A two-way ANOVA revealed a significant main effect of scale format ($F(4, 5,529) = 10.97, p < .001, \eta^2 = .008$) with a Bonferroni post-hoc test providing evidence of a significantly higher level of extreme responding in the drag-item format (.219) compared to the grid (.180, $p < .001$), drag-response (.179, $p < .001$), one-vertical (.179, $p < .001$), and one-horizontal format (.174, $p < .001$). Overall, the drag-item format induced respondents to select extreme response options more frequently than all the other scale formats. Again, the drag-item format also differed significantly from all the other scale formats when considering the relationship between scale format and extremity in the 6-item scale ($F(4, 1,811) = 6.56, p < .001, \eta^2 = .014$). By contrast, significant differences again largely disappeared in the 10-item ($F(4, 1,879) = 3.12, p < .05, \eta^2 = .007$) and 16-item scale ($F(4, 1,839) = 2.73, p < .05, \eta^2 = .006$). Nevertheless, prior expectations assuming that a respondent's tendency to select extreme responses would occur irrespective of the scale format could not be confirmed. By contrast, the main effect of scale length ($F(2, 5,529) = 0.21, ns$) as well as the two-way interaction between scale format and scale length ($F(8, 5,529) = .67, ns$) were non-significant as expected.

Table 35: Extremity index depending on scale format and scale length (Experiment 1.3, $n = 5,529$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 6	.173 ^c (.19)	.173 ^c (.18)	.226 ^{a,b,d,e} (.19)	.188 ^c (.17)	.164 ^c (.17)	.184 (.18)
(B) 10	.182 (.20)	.178 (.18)	.216 ^d (.19)	.173 ^c (.18)	.179 (.18)	.185 (.19)
(C) 16	.185 (.19)	.186 (.19)	.216 ^{d,e} (.18)	.177 ^c (.17)	.178 ^c (.17)	.187 (.18)
Total	.180 ^c (.19)	.179 ^c (.18)	.219 ^{a,b,d,e} (.19)	.179 ^c (.17)	.174 ^c (.17)	.185 (.18)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e).

9.7.2 Experiment 2

In Experiment 2, a respondent's tendency to extreme responding was examined depending on scale arrangement and scale format in a 2 x 5 between-subjects factorial design. Again, the extent of extreme responding was expected to be unaffected by scale format. Furthermore, whether response options were arranged from positive-to-negative or from negative-to-positive should not affect the extremity of responses. Equally, a two-way interaction between scale format and scale arrangement on extreme responding was not expected.

Analysis included cases with substantive answers for all ten items the rating scale was composed of in Experiment 2. The extremity index actually ranged from .00 to .80 with an overall mean of .227 (see Table 36). Findings of the ANOVA revealed a significant main effect of scale format ($F(4, 727) = 4.29, p < .01, \eta^2 = .023$). Contrary to expectations, a Bonferroni post-hoc test revealed a significantly higher incidence of extreme responding for the drag-item format (.276) compared to the drag-response (.210, $p < .01$), one-vertical (.218, $p < .05$), and one-horizontal format (.215, $p < .05$). The effect of scale format on the extent of extreme responding were also found in the negative-to-positive response option order ($F(4, 354) = 2.50, p < .05, \eta^2 = .028$), which was attributable to a significant difference between the drag-item format (.277) and the drag-response format (.197, $p < .05$), whereas no

significant differences existed in the positive-to-negative response option order ($F(4, 373) = 2.29, ns$). Furthermore, neither the main effect of scale arrangement ($F(1, 727) = 0.00, ns$) nor the two-way interaction between scale format and scale arrangement was significant ($F(4, 727) = 0.50, ns$). Thus, consistent with the expectations, scale arrangement did not affect the extent of extreme responding.

Table 36: Extremity index depending on scale format and scale arrangement (Experiment 2, $n = 727$)

Arrangement	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) Pos-Neg	.229 (.17)	.220 (.15)	.275 (.16)	.217 (.16)	.202 (.14)	.228 (.16)
(B) Neg-Pos	.222 (.15)	.197 ^c (.13)	.277 ^b (.15)	.218 (.15)	.228 (.13)	.226 (.15)
Total	.225 (.16)	.210 ^c (.14)	.276 ^{b,d,e} (.16)	.218 ^c (.15)	.215 ^c (.14)	.227 (.15)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e).

9.7.3 Experiment 3

In Experiment 3, extreme responding was evaluated depending on scale arrangement and scale format in terms of a 2×3 between-subjects factorial design²⁸. As already suggested in Experiment 2, scale format and scale arrangement were expected to have no effect on the extent of extreme responding. A two-way interaction between scale format and scale arrangement on extreme responding was not expected either.

In Experiment 3.1, the calculations were based on cases with substantive answers for a total of eight rating scale items. The extremity index actually ranged between .00 and 1.00 with an average extremity index of .337 across all experimental conditions (see Table 37). The main effect of scale format reached statistical significance ($F(2, 5,211) = 24.22, p < .001, \eta^2 = .009$) with a Bonferroni post-hoc test revealing a significantly higher extent of

²⁸ Again, the factor scale sequence was left out of subsequent analyses of extreme responding since variations in scale sequence were not expected to have a systematic effect on the extent of extreme responding in rating scales.

extreme responding in the drag-item format (.359) compared to the grid (.334, $p < .001$) and drag-response format (.316, $p < .001$) as well as a significantly higher extent of extreme responding in the grid compared to the drag-response format ($p < .01$). This held largely true when considering the relationship between scale format and extremity separately for the positive-to-negative ($F(2, 2,599) = 16.93, p < .001, \eta^2 = .013$) and the negative-to-positive response option order ($F(2, 2,612) = 9.15, p < .001, \eta^2 = .007$). Thus, contrary to expectations the extent of extreme responding actually varied depending on scale format.

As expected, the main effect of scale arrangement ($F(1, 5,211) = 3.82, ns$) and the two-way interaction between scale format and scale arrangement ($F(2, 5,211) = 1.79, ns$) were non-significant. However, separate analyses of the relationship between scale arrangement and extremity for each of the three scale formats revealed significantly more extreme responding in the positive-to-negative (.344) compared to negative-to-positive response option order (.324) when the items were presented in a grid format ($F(1, 1,844) = 5.16, p < .05, \eta^2 = .003$), whereas the extent of extreme responding in the drag-response ($F(1, 1,653) = 0.11, ns$) and drag-item format ($F(1, 1,714) = 2.18, ns$) remained unaffected by scale arrangement.

Table 37: Extremity index depending on scale format and scale arrangement (Experiment 3.1, $n = 5,211$)

Arrangement	Format			Total
	(a) Grid	(b) Drag-R	(c) Drag-I	
(A) Pos-Neg	.344 ^{b,c,B} (.19)	.315 ^{a,c} (.17)	.365 ^{a,b} (.17)	.342 (.18)
(B) Neg-Pos	.324 ^{c,A} (.18)	.318 ^c (.18)	.353 ^{a,b} (.18)	.332 (.18)
Total	.334 ^{b,c} (.19)	.316 ^{a,c} (.18)	.359 ^{a,b} (.17)	.337 (.18)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale formats, i.e., compared to grid (a), drag-response (b), and drag-item format (c). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the positive-to-negative (A) and negative-to-positive (B) scale arrangement.

In Experiment 3.2, calculations were based on cases with substantive answers for a total of eight rating scale items²⁹. The extremity index actually ranged between .00 and 1.00 with an overall mean of .301 (see Table 38). Contrary to expectations, the main effect of scale format reached statistical significance ($F(2, 5,227) = 28.02, p < .001, \eta^2 = .011$) with the drag-item format featuring once again a significantly higher extremity index (.336) compared to the grid (.285, $p < .001$) and drag-response format (.282, $p < .001$). This equally applied to the relationship between scale format and extremity separately for the positive-to-negative ($F(2, 2,639) = 12.75, p < .001, \eta^2 = .010$) and negative-to-positive response option order ($F(2, 2,588) = 15.61, p < .001, \eta^2 = .012$).

Contrary to prior expectations, a significant main effect of scale arrangement was found ($F(1, 5,227) = 4.37, p < .05, \eta^2 = .001$), indicating that the negative-to-positive response option order (.294) yielded a slightly but significantly lower incidence of extreme responding compared to the positive-to-negative response option order (.308). As shown in separate analyses of the three scale formats, a significantly lower incidence of extreme responding was found in the negative-to-positive response option order when presenting items in a grid format ($F(1, 1,899) = 3.99, p < .05, \eta^2 = .002$), whereas the drag-response ($F(1, 1,617) = 0.64, ns$) and drag-item format ($F(1, 1,711) = 0.77, ns$) remained unaffected by scale arrangement. As expected, the two-way interaction between scale format and scale arrangement was non-significant ($F(2, 5,227) = 0.40, ns$).

²⁹ Extremity index was based on seven rating scale items because item #7 was again excluded due to a negative correlation with the actual factor 'internal locus of control'.

Table 38: Extremity index depending on scale format and scale arrangement (Experiment 3.2, $n = 5,227$)

Arrangement	Format			Total
	(a) Grid	(b) Drag-R	(c) Drag-I	
(A) Pos-Neg	.296 ^{c,B} (.24)	.287 ^c (.24)	.341 ^{a,b} (.24)	.308 ^B (.24)
(B) Neg-Pos	.274 ^{c,A} (.24)	.277 ^c (.24)	.331 ^{a,b} (.24)	.294 ^A (.24)
Total	.285 ^c (.24)	.282 ^c (.24)	.336 ^{a,b} (.24)	.301 (.24)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale formats, i.e., compared to grid (a), drag-response (b), and drag-item format (c). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the positive-to-negative (A) and negative-to-positive (B) scale arrangement.

9.7.4 Summary

The present findings concerning the extent of extreme responding consistently showed significantly more extreme responding in the drag-item format compared to all the other scale formats (except for Experiment 1.1). On the contrary, the drag-response format did not differ from other scale formats regarding its extent of extreme responding (except for Experiment 3.1). Overall, prior expectations assuming that the extent of extreme responding would be unaffected by scale format could not be confirmed. By contrast, neither the number of items in a rating scale nor the categorial arrangement of response options seemed to be of great importance in affecting the extent of extreme responding since there were no significant effects of scale length or scale arrangement (except for Experiment 3.2).

In general, a higher degree of extreme responding is considered a kind of response tendency affecting data accuracy negatively. However, it may be assumed that a respondent's tendency to choose extreme response options has no negative effect on data accuracy, provided that a certain limiting value of extremity is not exceeded. For example, in Experiments 1.2 and 1.3 where respondents answered six rating scale items on a 5-point Likert-type scale, the maximum variation could be achieved by selecting each response option once, whereas one response option needed to be selected twice. In such a case, at least two out of six answers would count as extreme values. By

implication, a conceivable limiting value of the extremity index is .333 as an approximation to this value would make for a higher degree of differentiation, whereas an extremity index beyond this limit would affect scale differentiation negatively. In Experiments 1.2 and 1.3, the average extremity value in the drag-item format was clearly below this critical limit. Similarly, in Experiment 2, the mean extremity value of the drag-item format clearly undercut the critical limit of .400 (with at least four out of ten answers accounting for extreme values). Solely in Experiment 3.1, the critical limit of .250 (with at least two out of eight answer choices accounting for extreme values) was exceeded in every scale format. Similarly, in Experiment 3.2, the extremity index in the drag-item format clearly exceeded this limit, while the grid and drag-response format were somewhat above the critical limit of .286 (with at least two out of seven answer choices accounting for extreme values).

All in all, it can be noted that each of the present experiments indicated a rather moderate tendency to use extreme response options. And although the drag-item format showed a comparatively higher susceptibility to extreme responding than the other rating scale formats, this higher tendency to use extreme response options is considered still moderate.

9.8 Primacy Effects

9.8.1 *Experiment 2*

In Experiment 2, the liability of varying scale formats to primacy effects in terms of a left-side bias in horizontally arranged rating scales or in terms of a top-end bias in vertically arranged rating scales was assessed in a 2 x 5 between-subjects factorial design. In addition to the format of a rating scale, scale arrangement was varied from positive-to-negative or from negative-to-positive to reveal systematic shifts towards the leftmost or topmost side of a rating scale. Primacy effects were measured in terms of significant differences in response distributions and item means depending on a positive-to-negative or a negative-to-positive scale arrangement.

It was assumed that primacy effects were prevented most effectively in the drag-response format as compared to the grid and drag-item format since the drag-response format was expected to draw a respondent's attention directly and repeatedly to the response options. With every new item being answered in the drag-response format, respondents might review the meaning

of the response options again and would select the respective response option more carefully. Concerning the drag-item format, two outcomes were conceivable. Respondents would be discouraged from selecting the first response option that seemed reasonable to them without taking sufficient account of the latter ones because the drag-item format might generally encourage the respondents' attentiveness and carefulness in processing a rating scale. Or, quite the opposite, respondents in the drag-item format would be tempted to simply select one of the first response options, particularly because of the higher navigational effort required to select a response option at the bottom of a vertically arranged response scale. In the first case, the drag-item format was expected to decrease the risk of primacy effects compared to the grid format, whereas in the latter case it was expected to be increased.

The analyses included cases with substantive answers for all ten items the rating scale consisted of in Experiment 2. Items with a negative-to-positive response option order were reversed prior to analyses to enable an adequate comparison of the items being all coded in the same direction (1 = 'completely agree' to 5 = 'not agree at all'). Figure 4 shows the response distributions and the item means depending on scale arrangement separately for the grid, drag-response, and drag-item format³⁰. Comparing differences in response distributions as a function of the positive-to-negative and negative-to-positive response option order separately for each scale format, Pearson's chi-squared tests revealed no significant primacy effects³¹. Apart from item #2 where a significantly higher mean was found for the drag-response format ($F(1, 156) = 4.09, p < .05, \eta^2 = .026$) when starting with the negative end of a rating scale compared to beginning with the positive scale end, one-way ANOVAs revealed no further significant mean differences. Thus, there was only limited evidence for primacy effects in Experiment 2, and contrary to expectations, there were basically no differences between various scale formats.

³⁰ Response distributions and item means for the one-vertical and one-horizontal format were not depicted for better clarity. Furthermore, there were no significant differences in the risk of primacy effects for these two formats except for item #8 in the one-horizontal format ($F(1, 135) = 7.19, p < .01, \eta^2 = .051$).

³¹ Although in the grid format response distributions significantly differed for item #6 ($\chi^2(4, 140) = 10.64, p < .05$), there was no significant mean difference as indicated by a one-way ANOVA ($F(1, 140) = 0.00, ns$).

Besides, it was noticeable that for some items, one of the three scale formats showed a comparatively even distribution of the responses over the range of response options. This primarily applied to the grid format in a negative-to-positive scale arrangement (item #2, #3, #6), and the drag-response format in both scale arrangements (item #10) or in a negative-to-positive scale arrangement (item #4). However, except for item #6 in the grid format, no significant differences in response distributions were found depending on scale arrangement (see footnote 31).

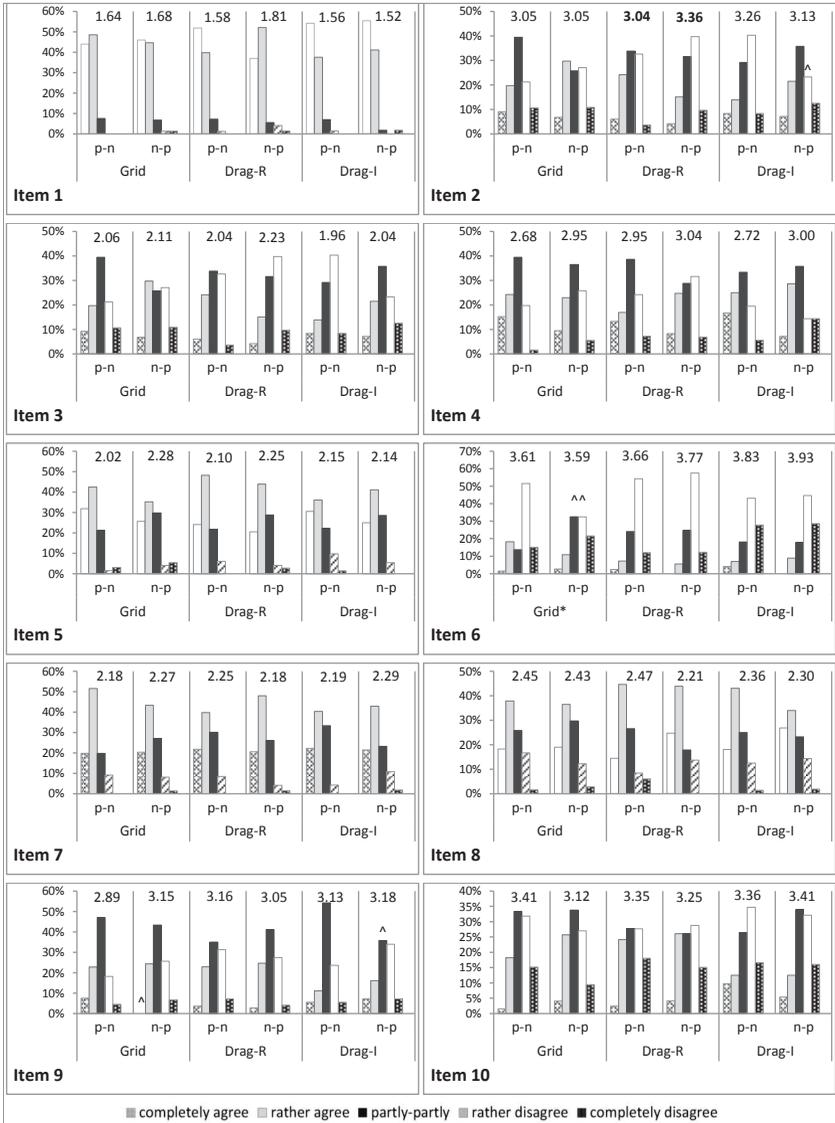


Figure 4: Response distributions and item means depending on a positive-to-negative (p-n) or negative-to-positive (n-p) scale arrangement separately for scale format (Experiment 2, $n = 727$).

Pearson's chi-squared tests with pair-wise comparisons (Bonferroni correction: $\wedge p < .05$ or less): *** $p < .001$, ** $p < .01$, * $p < .05$. One-way ANOVAs for independent samples: item means in bold $p < .05$ or less.

9.8.2 Experiment 3

In Experiment 3, primacy effects were evaluated depending on scale arrangement and scale format in terms of a 2 x 3 between-subjects factorial design³². As already suggested in Experiment 2, the drag-response format was expected to have a lower risk of primacy effects relative to the grid and drag-item format whereas predictions concerning the drag-item format were still two-sided, possibly showing either a decrease or increase in primacy effects compared to the grid format.

In Experiments 3.1 and 3.2, calculations were based on cases with substantive answers for the eight items of the respective rating scale. Items with a negative-to-positive response option order were reversed, whereby all items were coded in the same direction (1 = completely agree to 5 = completely disagree).

In Experiment 3.1, the analyses of response distributions and item means depending on either a positive-to-negative or negative-to-positive scale arrangement revealed no significant primacy effects for the drag-response format, whereas significant differences were found for 7 out of 8 items in the drag-item format³³ and for 5 out of 8 items in the grid format³⁴ (see Figure 5). In the drag-item format, significant differences as a function of a positive-to-negative and negative-to-positive scale arrangement basically resulted from

³²Since systematic effects of scale sequence on the occurrence of primacy effects would have been theoretically conceivable, two-way ANOVAs were conducted separately for each scale format. However, the factor scale sequence was left out of subsequent analyses of primacy effects because largely no significant interactions between scale sequence and scale arrangement have been found. Exceptions were: item #6 ($F(1, 1,844) = 4.50, p < .05, \eta^2 = .002$) and item #8 ($F(1, 1,844) = 4.88, p < .05, \eta^2 = .003$) in the grid format and item #8 ($F(1, 1,653) = 4.07, p < .05, \eta^2 = .002$) in the drag-response format in Experiment 3.1; item #6 ($F(1, 1,711) = 5.51, p < .05, \eta^2 = .003$) in the drag-item format in Experiment 3.2.

³³Concerning the drag-item format, Pearson's chi-squared tests revealed significant differences for item #2: $\chi^2(4, 1,714) = 16.18, p < .01$; item #3: $\chi^2(4, 1,714) = 21.44, p < .001$; item #4: $\chi^2(4, 1,714) = 30.17, p < .01$; item #5: $\chi^2(4, 1,714) = 29.04, p < .001$; item #6: $\chi^2(4, 1,714) = 16.76, p < .01$; item #7: $\chi^2(4, 1,714) = 33.03, p < .001$; item #8: $\chi^2(4, 1,714) = 15.37, p < .01$.

³⁴Concerning the grid format, Pearson's chi-squared tests revealed significant differences for item #1: $\chi^2(4, 1,844) = 44.54, p < .001$; item #2: $\chi^2(4, 1,844) = 25.61, p < .001$; item #4: $\chi^2(4, 1,844) = 44.01, p < .001$; item #5: $\chi^2(4, 1,844) = 10.23, p < .05$; item #7: $\chi^2(4, 1,844) = 23.14, p < .001$.

respondents selecting the most positive (or most negative) and the moderate positive (or moderate negative) response options more often when beginning with the positive (or negative) scale end. In the grid format, there was rather a preference for the most positive (or most negative) response option when beginning with the positive (or negative) scale end. This systematic preference of response options arranged at the topmost scale end in the drag-item format and at the leftmost scale end in the grid format also resulted in significant mean shifts towards the positive end of the rating scale with a positive-to-negative scale arrangement and towards the negative end with a negative-to-positive scale arrangement for all 8 items in the drag-item format³⁵ and for 5 items in the grid format³⁶. Thus, regarding the drag-response format, prior expectations could be fully confirmed since none of the rating scale items were affected by systematic variations in scale arrangement. On the contrary, the drag-item and grid format were systematically affected by scale arrangement, resulting in primacy effects for the vast majority of rating scale items. Hence, with respect to the drag-item format, the second of the two initial assumptions was confirmed as the risk of primacy effects in the drag-item format was even more pronounced than in the grid format.

³⁵ Concerning the drag-item format, one-way ANOVAs revealed significant differences for item # 1: $F(1, 1,714) = 5.06, p < .05, \eta^2 = .003$; item #2: $F(1, 1,714) = 14.99, p < .001, \eta^2 = .009$; item #3: $F(1, 1,714) = 15.06, p < .001, \eta^2 = .009$; item #4: $F(1, 1,714) = 24.11, p < .001, \eta^2 = .014$; item #5: $F(1, 1,714) = 18.68, p < .001, \eta^2 = .011$; item #6: $F(1, 1,714) = 13.35, p < .001, \eta^2 = .008$; item #7: $F(1, 1,714) = 31.68, p < .001, \eta^2 = .018$; item #8: $F(1, 1,714) = 14.82, p < .001, \eta^2 = .009$.

³⁶ Concerning the grid format, one-way ANOVAs revealed significant differences for item # 1: $F(1, 1,844) = 28.08, p < .001, \eta^2 = .015$; item #2: $F(1, 1,844) = 21.78, p < .001, \eta^2 = .012$; item #4: $F(1, 1,844) = 33.56, p < .001, \eta^2 = .018$; item #5: $F(1, 1,844) = 7.47, p < .01, \eta^2 = .004$; item #7: $F(1, 1,844) = 17.10, p < .001, \eta^2 = .009$.

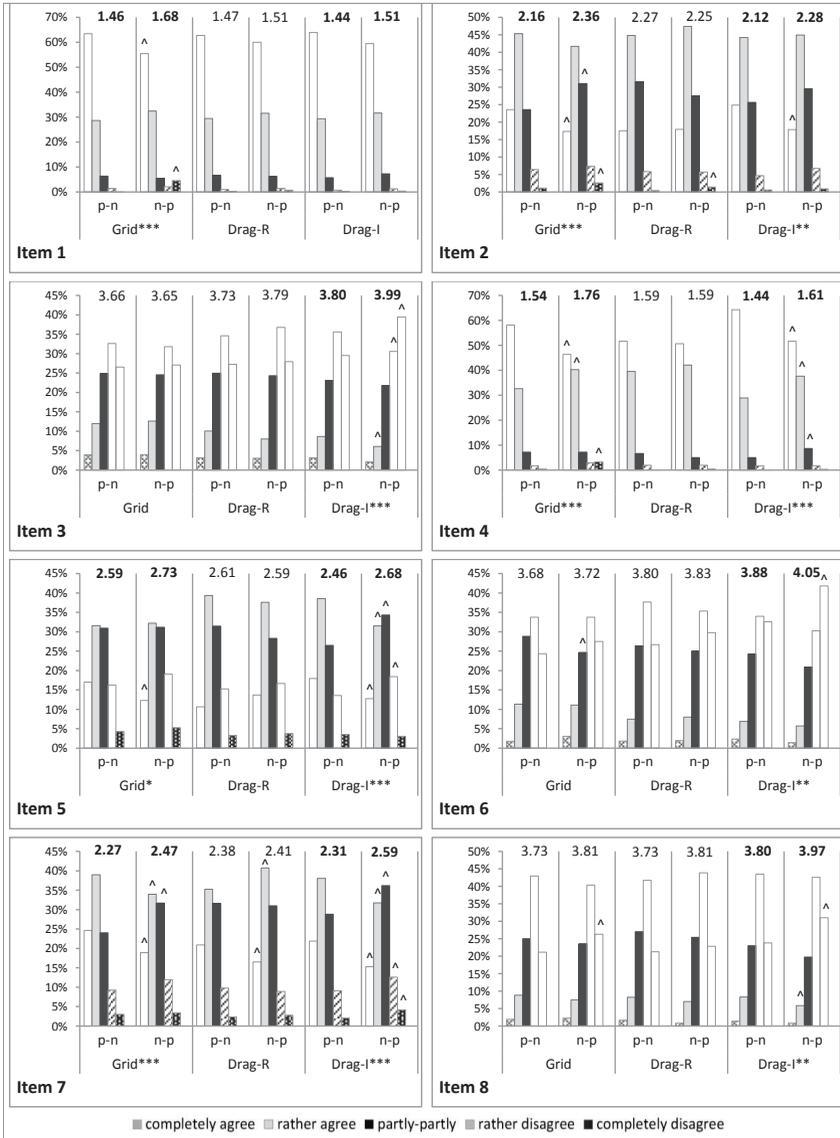


Figure 5: Response distributions and item means depending on a positive-to-negative (p-n) or negative-to-positive (n-p) scale arrangement separately for scale format (Experiment 3.1, $n = 5,211$).

Pearson's chi-squared tests with pair-wise comparisons (Bonferroni correction: $^{\wedge} p < .05$ or less). *** $p < .001$, ** $p < .01$, * $p < .05$. One-way ANOVAs for independent samples: item means in bold $p < .05$ or less.

For Experiment 3.2, the response distributions and the item means of the seven rating scale items³⁷ are depicted in Figure 6 depending on a positive-to-negative or negative-to-positive scale arrangement separately for each of the three scale formats. When the items were presented in the drag-response format, a significant primacy effect was found merely for item #8³⁸. By contrast, 6 out of 7 items were affected by a significant primacy effect when the items were presented in the drag-item format³⁹. These differences between a positive-to-negative and negative-to-positive scale arrangement in the drag-item format were predominantly due to a respondent's tendency to select the most negative response option more often when it was presented first than in the reverse scale arrangement. Contrary to the findings of Experiment 3.1, significant differences were found merely for 2 out of 7 items in the grid format⁴⁰.

Therefore, as expected, the drag-response format was again least affected by primacy effects. Consistent with the findings of Experiment 3.1 but contrary to expectations, the drag-item format was most susceptible to primacy effects. Contrary to the findings of Experiment 3.1, the extent to which the grid format was affected by primacy effects was rather low.

³⁷ Calculations were based on seven rating scale items because item #7 was again excluded due to a negative correlation with the actual factor 'internal locus of control'.

³⁸ Concerning the drag-response format, a one-way ANOVA revealed a significant mean difference for item #8 ($F(1, 1,617) = 4.92, p < .05, \eta^2 = .003$), whereas the Pearson's chi-squared test was non-significant ($\chi^2(4, 1,617) = 8.35, ns$). A Pearson's chi-squared test revealed a significant difference for item #2 ($\chi^2(4, 1,617) = 9.98, p < .05$), however, a one-way ANOVA found no significant difference in item means ($F(1, 1,617) = 0.22, ns$) which is why this was not considered a primacy effect.

³⁹ Concerning the drag-item format, Pearson's chi-squared tests revealed significant differences for item #1: $\chi^2(4, 1,711) = 21.39, p < .001$; item #2: $\chi^2(4, 1,711) = 9.54, p < .05$; item #4: $\chi^2(4, 1,711) = 20.92, p < .001$; item #5: $\chi^2(4, 1,711) = 16.55, p < .01$; item #8: $\chi^2(4, 1,711) = 12.25, p < .05$. Correspondingly, one-way ANOVAs revealed significant mean differences for item #1: $F(1, 1,711) = 13.96, p < .001, \eta^2 = .008$; item #2: $F(1, 1,711) = 8.67, p < .01, \eta^2 = .005$; item #4: $F(1, 1,711) = 17.774, p < .001, \eta^2 = .010$; item #5: $F(1, 1,711) = 4.91, p < .05, \eta^2 = .003$; item #8: $F(1, 1,711) = 7.10, p < .01, \eta^2 = .004$, and additionally, for item #6: $F(1, 1,711) = 5.73, p < .05, \eta^2 = .003$.

⁴⁰ Concerning the grid format, Pearson's chi-squared tests revealed significant differences for item #1: $\chi^2(4, 1,899) = 41.82, p < .001$; item #4: $\chi^2(4, 1,899) = 29.25, p < .001$ which corresponded with the findings of one-way ANOVAs revealing a significant mean difference for item #1: $F(1, 1,899) = 34.11, p < .001, \eta^2 = .018$; item #4: $F(1, 1,899) = 20.71, p < .001, \eta^2 = .011$.

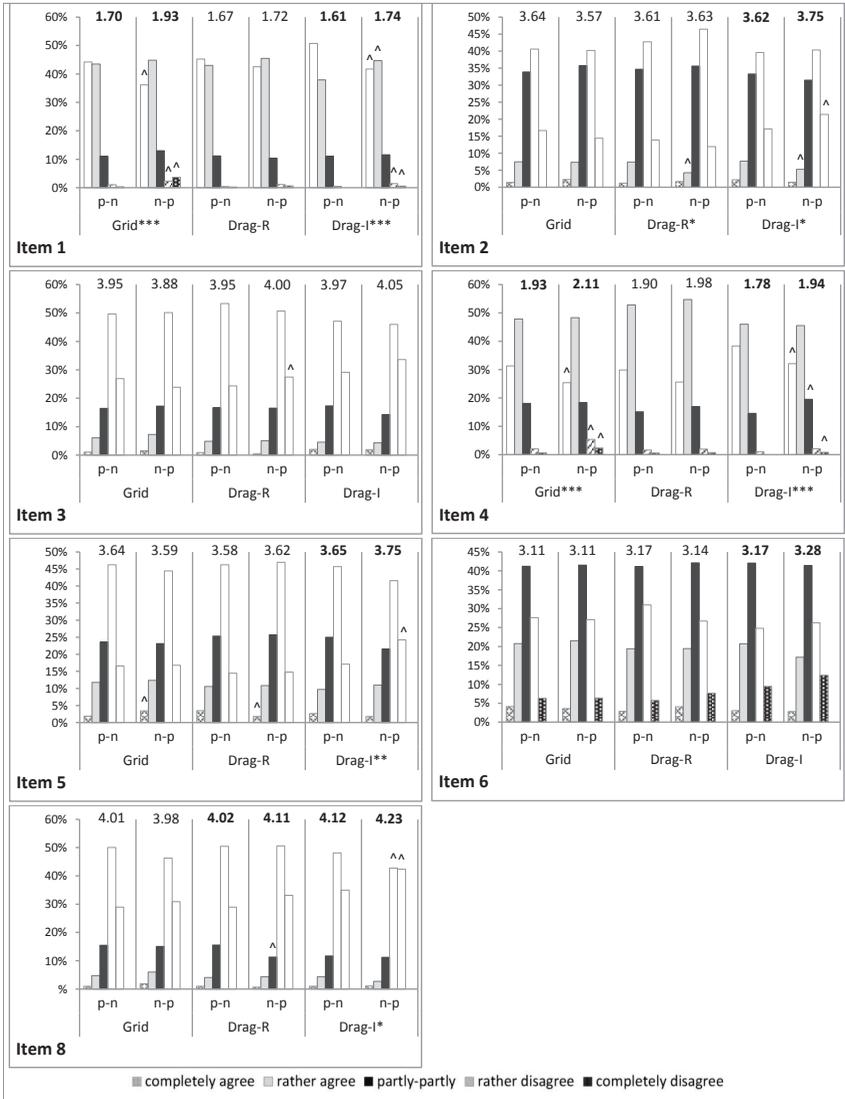


Figure 6: Response distributions and item means depending on a positive-to-negative (p-n) or negative-to-positive (n-p) scale arrangement separately for scale format (Experiment 3.2, $n = 5,227$).

Pearson's chi-squared tests with pair-wise comparisons (Bonferroni correction: $\wedge p < .05$ or less): *** $p < .001$, ** $p < .01$, * $p < .05$. One-way ANOVAs for independent samples: item means in bold $p < .05$ or less.

9.8.3 Summary

The findings of Experiments 3.1 and 3.2 provided clear evidence of primacy effects in rating scales for at least some rating scale designs, whereas in Experiment 2, substantive answers remained virtually unaffected by primacy effects. A comparison between the two drag-and-drop scales and the standard grid in Experiments 3.1 and 3.2 showed that the drag-item format was most susceptible to primacy effects with almost every item of a rating scale (7 out of 8 and 6 out of 7 items) showing significant differences in substantive responses due to varying scale arrangements. While the grid format was also considerably affected by primacy effects (5 out of 8 items) in Experiment 3.1, a systematic tendency to the left side of a rating scale was found to a much lesser extent (2 out of 7 items) in Experiment 3.2. By contrast, when rating scale items were presented in a drag-response format, (virtually) no primacy effects could be observed, neither in Experiment 2 (except for 1 out of 10 items), nor in Experiment 3.1, or in Experiment 3.2 (except for 1 out of 7 items). Thus, findings consistently confirmed prior expectations that the drag-response format was most effective in preventing primacy effects. By contrast, findings concerning the drag-item and grid format were rather mixed: In Experiment 2, no evidence of primacy effects was found for both formats. In Experiments 3.1 and 3.2, however, it has been clearly demonstrated that the drag-item format was most susceptible to primacy effects, whereas findings regarding the incidence of primacy effects in a grid format remained inconclusive revealing either a high (Experiment 3.1) or moderate (Experiment 3.2) susceptibility to primacy effects.

9.9 Semantic-Order Effects

In Experiment 3, the occurrence of semantic-order effects was evaluated depending on scale sequence and scale format in a 2 x 3 between-subjects factorial design⁴¹. Besides the format of a rating scale, the sequence in which

⁴¹ Since systematic effects of scale arrangement on the occurrence of semantic-order effects would have been theoretically conceivable, two-way ANOVAs were conducted separately for each scale format. Factor scale arrangement was left out of subsequent analyses of semantic-order effects because largely no significant interactions between scale sequence and scale arrangement have been found. Exceptions were described in footnote 32.

the items were presented was varied in terms of a forward and backward item order to systematically reveal potential semantic-order effects in a rating scale.

Semantic-order effects in rating scales occur if responses to later items are systematically affected by the meaning of previously processed items of the same rating scale. Commonly, if a set of several rating scale items is related in content, the respective items are intentionally grouped on the same screen in a Web survey in order to encourage respondents to take the various aspects of an underlying theoretical construct into account when answering the items, instead of considering each item independently of one another. The question context is increasingly enriched with each additional item being answered which is deemed necessary if we are to gain a complete understanding of the underlying theoretical construct of a rating scale. Moreover, the question context may, to some extent, facilitate the interpretation of later items, the retrieval of relevant information, and the judgment on later items (Knowles, 1988; Schwarz, Strack, & Mai, 1991).

Semantic-order effects in terms of systematic shifts in mean responses were expected to occur to a comparable extent in all three rating scale formats because each of the three formats would encourage respondents to consider a rating scale item in its entirety instead of processing each item separately. By contrast, predictions concerning systematic shifts in reliability were less obvious. Given that the rating scales used in Experiments 3.1 and 3.2 were multidimensional instead of unidimensional, semantic-order effects might be reduced or would even completely fail to appear when confronting respondents with a randomized sequence of items measuring distinct components of a latent construct. Nevertheless, semantic-order effects were supposed to emerge to a comparable extent in all three rating scale formats in the present two-dimensional rating scales because respective items measuring the same component of a latent construct were clustered to at least a certain extent instead of being completely alternated.

In Experiment 3, calculations were based on cases with substantive answers for all eight rating scale items. Findings of Experiment 3.1 are depicted in Figure 7. Separate one-way ANOVAs clearly showed significant differences in item means as a function of the forward and backward scale sequence. In each of the three scale formats, all rating scale items (except item #3) displayed a significantly higher mean in the forward scale sequence as compared to the backward scale sequence (all comparisons $p < .05$ or less,

respectively). A more detailed assessment of the content of respective rating scale items (items of subscale 1 are marked by a dashed line) suggested that starting with an overall positively assessed item of subscale 1 in the forward scale sequence resulted in significantly more negative responses to each following item than starting with an overall negatively assessed item of subscale 2 in the backward scale sequence, and vice versa. This applied to all three scale formats. An examination of the response distributions (results not presented here) supported these findings concerning mean shifts towards the positive rating scale end in the backward scale sequence since the respondents were significantly more likely to select the extremely positive and moderately positive response option when the items were arranged in the backward scale sequence.

In view of the Fisher's z transformed within-subscale item-total correlations, reliability shifts were most obvious in the grid and drag-item format: Items #6, #7, and #8 of subscale 2 showed significantly stronger within-subscale item-total correlations in the forward compared to the backward scale sequence (all comparisons $p < .05$ or less, respectively). Thus, rating scale items showed increasingly stronger correlations with the respective subscale, the more items have already been answered. By tendency, this also applied to the drag-response format. Conversely, items #1 and #4 of subscale 1 showed significantly stronger within-subscale item-total correlations in the backward compared to the forward scale sequence in all three scale formats (all comparisons $p < .05$ or less, respectively). Thus, item reliabilities increased with progressive scale completion, even in the case of a two-dimensional rating scale.

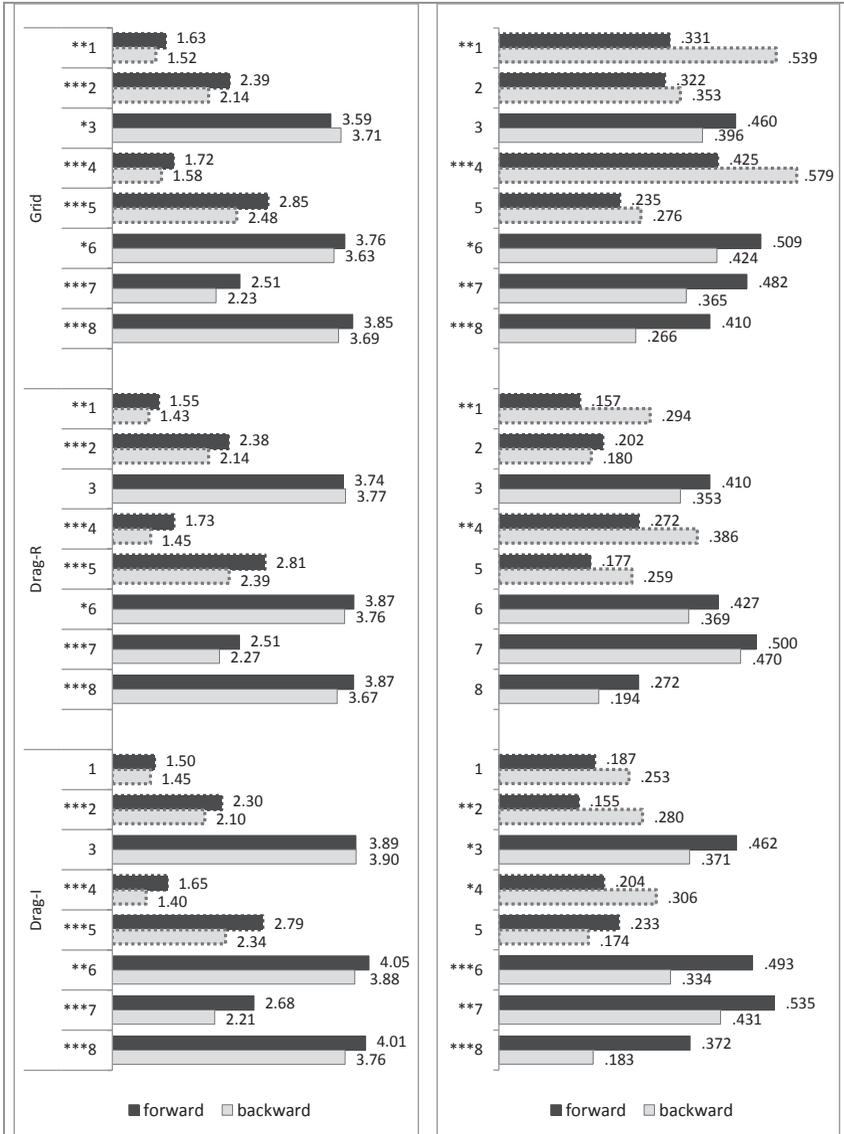


Figure 7: Mean shifts (left) and reliability shifts (right) depending on a forward and backward scale sequence separately for scale format (Experiment 3.1, $n = 5,211$). Note. Items of subscale 1 are marked by a dashed line. Calculations were based on one-way ANOVAs (Bonferroni correction) and multiple z-tests comparing Fisher's z transformed within-subscale item-total correlations (no alpha correction): *** $p < .001$, ** $p < .01$, * $p < .05$.

The findings of Experiment 3.2 are depicted in Figure 8 (items of subscale 1 are marked by a dashed line). Separate one-way ANOVAs for each of the seven items⁴² revealed significant differences in item means as a function of the forward and backward scale sequence for items #2, #5, #6, and #8 in all three scale formats (all comparisons $p < .05$ or less, respectively). Thus, significant mean shifts were found for all the items of subscale 2 (except for item #3) in terms of a lower mean for items #2, #5, and #6 in the forward scale sequence as compared to the backward scale sequence, whereas for item #8 the reverse applied. Thus, as opposed to the findings of Experiment 3.1, starting with an overall positively assessed item of subscale 1 in the forward scale sequence resulted in significantly more positive responses to the following items #2, #5, and #6 than starting with an overall negatively assessed item of subscale 2 in the backward scale sequence, and vice versa. In view of the response distributions (not presented here), this was due to respondents selecting the most negative response option significantly more often in the backward scale sequence compared to the forward scale sequence. On the contrary, for item #1 and #4 of subscale 1, no significant mean shifts were found in any of the three scale formats.

The findings on Fisher's z transformed within-subscale item-total correlations provided little evidence for the existence of reliability shifts. Solely in the drag-response format, significantly higher item reliabilities were found for items #6 and #8 of subscale 2 in the forward compared to the backward scale sequence, whereas in the drag-item format significantly higher item reliabilities were found for items #1 and #4 of subscale 1 in the backward compared to the forward scale sequence (all comparisons $p < .05$ or less, respectively). In the grid format, no significant reliability shifts were found. Thus, while there were clear mean shifts for items of subscale 2 in all three scale formats, evidence of item reliabilities was rather limited in Experiment 3.2.

⁴² Calculations were based on seven rating scale items because item #7 was again excluded due to a negative correlation with the actual factor 'internal locus of control'.

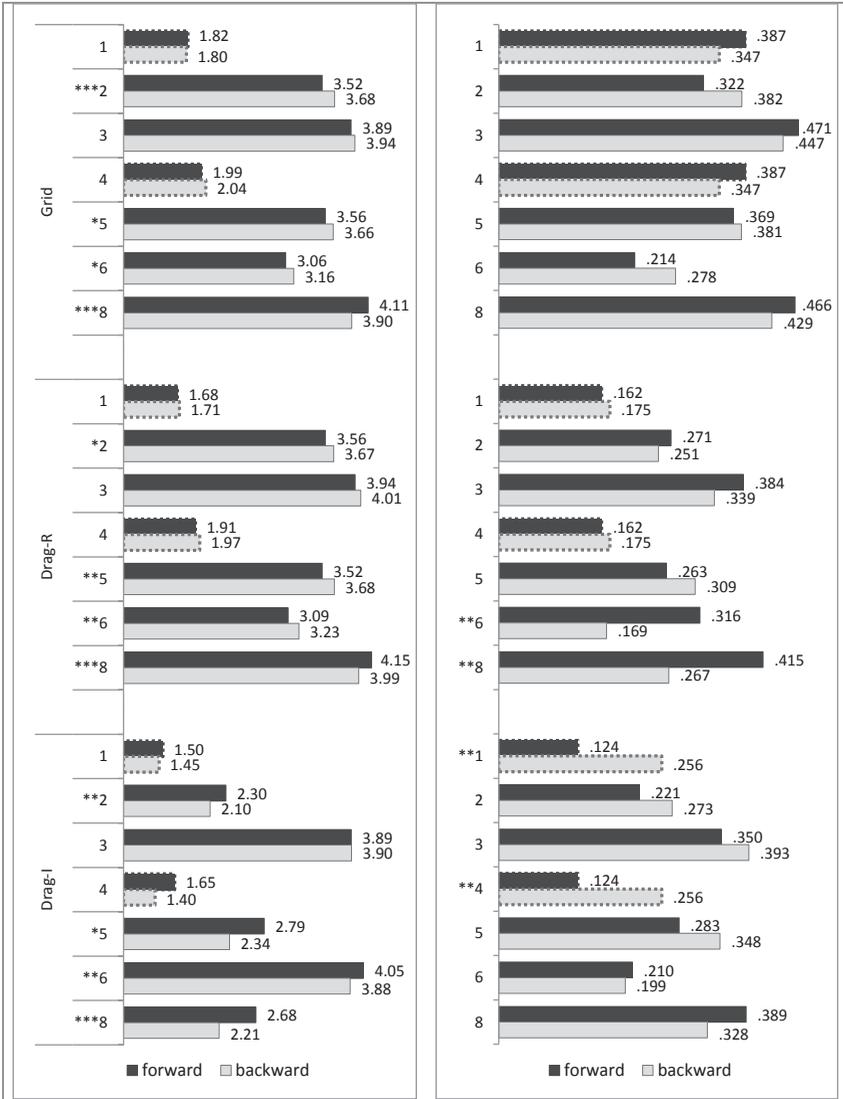


Figure 8: Mean shifts (left) and reliability shifts (right) depending on a forward and backward scale sequence separately for scale format (Experiment 3.2, $n = 5,227$).

Note. Items of subscale 1 are marked by a dashed line. Calculations were based on one-way ANOVAs (Bonferroni correction) and multiple z-tests comparing Fisher's z transformed within-subscale item-total correlations (no alpha correction): *** $p < .001$, ** $p < .01$, * $p < .05$.

In sum, Experiments 3.1 and 3.2 gave clear evidence of semantic-order effects in terms of shifts in item means, depending on a forward or backward sequence of rating scale items. Semantic-order effects occurred to a comparable extent in all three scale formats. Semantic-order effects were also found in terms of reliability shifts, even though findings were less clear: In Experiment 3.1, there was strong evidence of an increase in within-subscale item-total correlations with the number of items of the same underlying construct that were already answered. By contrast, significant reliability shifts were largely nonexistent in Experiment 3.2.

9.10 Response Times

9.10.1 Experiment 1

In Experiment 1, the time taken to complete a rating scale was examined as a function of two (three) different scale lengths and five different scale formats in a 2 x 5 (3 x 5) between-subjects factorial design. Response times were automatically captured by server-side time stamps which enabled the analysis of the total time to complete a set of rating scale items. However, when comparing rating scales of differing length, the total time for scale completion was hardly informative since longer response times with an increasing number of items would be unsurprising. Hence, item-response times were calculated by dividing the total time for scale completion by the total number of items in a rating scale.

Concerning differing scale formats, both drag-and-drop formats were expected to yield longer item-response times compared to the grid format. On the one hand, the drag-and-drop technique was considered more time-consuming in terms of its basic understanding and handling than simply clicking a radio button in well-established rating scale formats. On the other hand, respondents were expected to spend more time on the cognitive processing of the rating scale content. Furthermore, the drag-item format was expected to need longer than the drag-response format since respondents might spend even more time on initial orientation and understanding of its basic functionality. With respect to the number of rating scale items, the average time spent on answering single items was expected to decrease with increasing scale length because of efficiency gains: The question context was increasingly enriched, whereby the cognitive processing of subsequent items

was expected to be facilitated. Accordingly, item-response times would be longest for the 6-item scale, somewhat shorter for the 10-item scale, and shortest for the 16-item scale. Furthermore, an interaction between scale format and scale length was expected in terms of larger efficiency gains in both drag-and-drop formats compared to the grid format. Respondents might become more practiced in using the drag-and-drop technique. Thus, a decline in item-response times with every additional item that had already been answered might be stronger in the drag-and-drop formats compared to the grid format in consequence of a combination of facilitation and learning effects in the former.

In Experiment 1.1, cases with one or more missing values in either the 10-item or 16-item scale⁴³ were excluded from the analysis of server-side response times. In addition, unreasonably high response times were removed by excluding cases that exceeded the session timeout due to interruptions on the target page. Subsequently, cases with an averaged item time ± 2 standard deviations from group mean were excluded, resulting in a total of 673 cases the following analyses were based on. A first examination of server-side response times showed great variations in item-response times ranging from 6.8 up to 11.8 seconds per item as a function of scale format and scale length (see Table 39). Findings of an ANOVA revealed a significant main effect of scale format ($F(4, 673) = 30.97, p < .001, \eta^2 = .157$). As expected, the drag-item format (11.1 sec) showed a significantly longer item-response time compared to the drag-response format (9.6 sec, $p < .01$), as well as compared to the grid (7.0 sec, $p < .001$), one-vertical (8.7 sec, $p < .001$), and one-horizontal format (7.8 sec, $p < .001$). Also, consistent with prior expectations, the drag-response format took significantly longer compared to the grid ($p < .001$) and one-horizontal format ($p < .001$) but no longer than the one-vertical format ($p = .368$). Thus, items presented in a grid format took the least amount of time to be completed, followed by the one-horizontal and one-vertical format, whereas in the drag-response and especially in the drag-item format, it clearly took longest to answer an item. Separate analyses of the

⁴³ In multiple-item-per-screen designs, server-side response times for a 10-item and 16-item scale could only be calculated in its entirety which is why—in contrast to the previously applied procedure—present response time analyses only referred to cases with substantive answers to all items in the respective rating scale. This resulted in a total of 703 cases after exclusion of further 11 cases that have not completed all 16 items of the 16-item scale.

relationship between scale format and item-response time revealed largely the same patterns for the 10-item ($F(4, 673) = 17.19, p < .001, \eta^2 = .167$) and 16-item scale ($F(4, 673) = 17.53, p < .001, \eta^2 = .180$) with further significant differences being indicated in Table 39.

Proven by a non-significant main effect of scale length ($F(1, 673) = 0.09, ns$), the overall assumption of efficiency gains, resulting in decreased item-response times with increasing scale length, could not be confirmed⁴⁴. A significant two-way interaction between scale format and scale length ($F(4, 673) = 2.77, p < .05, \eta^2 = .016$) primarily referred to differences between the grid and both single-item-per-screen formats since the grid format differed from the one-vertical and one-horizontal format solely in the 16-item scale but not in the 10-item scale.

Table 39: Server-side item-response time (mean in seconds) depending on scale format and scale length (Experiment 1.1, $n = 673$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 10	7.2 ^{b,c} (2.5)	9.2 ^{a,c,e} (3.2)	11.8 ^{a,b,d,e} (5.7)	8.8 ^c (4.3)	7.3 ^{b,c,B} (1.6)	8.9 (4.1)
(B) 16	6.8 ^{b,c,d,e} (1.9)	9.9 ^{a,e} (2.6)	10.4 ^{a,d,e} (3.0)	8.6 ^{a,c} (2.7)	8.3 ^{a,b,c,A} (3.1)	8.7 (2.9)
Total	7.0 ^{b,c,d} (2.2)	9.6 ^{a,c,e} (2.9)	11.1 ^{a,b,d,e} (4.7)	8.7 ^{a,c} (3.6)	7.8 ^{b,c} (2.5)	8.8 (3.6)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the 10-item (A) and 16-item scale (B).

⁴⁴ Furthermore, separate analyses of the relationship between scale length and item-response times revealed a significantly longer time for the 16-item scale compared to the 10-item scale when items were presented in a one-horizontal format ($F(1, 137) = 5.86, p < .05, \eta^2 = .042$), whereas differences were non-significant for the grid ($F(1, 139) = 1.03, ns$), drag-response ($F(1, 116) = 1.39, ns$), drag-item ($F(1, 131) = 3.12, ns$), and one-vertical format ($F(1, 150) = 0.14, ns$).

In Experiment 1.2, cases with at least one missing value in either the 6-item scale, the 10-item scale, or the 16-item scale were excluded from analysis of server-side response times⁴⁵. After eliminating cases that exceeded session timeouts on the target page, additional exclusion of cases with an item-response time ± 2 standard deviations from group mean resulted in a total of 4,644 cases. Average item-response times ranged between 7.1 and 14.8 seconds as illustrated in Table 40. The findings of the ANOVA revealed a significant main effect of scale format ($F(4, 4,644) = 182.91, p < .001, \eta^2 = .136$) and scale length ($F(2, 4,644) = 33.05, p < .001, \eta^2 = .014$), as well as a significant two-way interaction between scale format and scale length ($F(8, 4,644) = 3.65, p < .001, \eta^2 = .006$). Concerning the five different scale formats, a Bonferroni post-hoc test showed significant differences between the item-response time of the drag-item format (13.2 sec) compared to the drag-response format (11.0 sec, $p < .001$), as well as in relation to the grid (8.0 sec, $p < .001$), one-vertical (9.4 sec, $p < .001$), and one-horizontal format (9.7 sec, $p < .001$). Hence, consistent with prior expectations, the drag-item format showed the longest item-response time, even longer than the drag-response format. Also as expected, the drag-response format had a significantly longer item-response time compared to the grid ($p < .001$), one-vertical ($p < .001$), and one-horizontal format ($p < .001$). Similar results were found when analyzing differences between the five scale formats separately for the 6-item ($F(4, 1,571) = 69.75, p < .001, \eta^2 = .151$), the 10-item ($F(4, 1,570) = 53.84, p < .001, \eta^2 = .121$), and 16-item scale ($F(4, 1,503) = 62.66, p < .001, \eta^2 = .143$). Concerning scale length, item-response time was significantly higher in the 6-item scale (10.8 sec) as compared to both the 10-item scale (10.1 sec, $p < .001$) and the 16-item scale (9.6 sec, $p < .001$). Furthermore, the 10-item scale showed a significantly longer item-response time than the 16-item scale ($p < .01$). In line with prior expectations, response times decreased with an increasing length of a rating scale. Separate one-way ANOVAs for each of the five scale formats indicated a significant effect of scale length for each scale format, with the exception of the drag-response

⁴⁵ This procedure resulted in an exclusion of further 68 cases that have not completed either all 10 items of the 10-item scale ($n = 24$) or all 16 items of the 16-item scale ($n = 44$) resulting in a total of 4745 cases.

format⁴⁶. Strikingly, the item-response time for the drag-response scale in a 6-item scale was equal to the item-response time in the 16-item scale (11.1 sec, respectively). This finding was also reflected in a significant two-way interaction between scale format and scale length indicating that the drag-response format was the only scale format that yielded constant item-response times irrespective of the number of rating scale items. Thus, whereas in the drag-response format the time spent on answering a single item was unaffected by the total number of rating scale items, item-response times decreased with an increasing number of items in a rating scale in all the other scale formats.

Table 40: Server-side item-response time (mean in seconds) depending on scale format and scale length (Experiment 1.2, $n = 4,644$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 6	8.4 ^{b,c,d,e,C} (3.8)	11.1 ^{a,c} (3.9)	14.8 ^{a,b,d,e,B,C} (6.5)	10.1 ^{a,c,B,C} (4.2)	10.3 ^{a,c,C} (5.4)	10.8 ^{B,C} (5.3)
(B) 10	8.2 ^{b,c,d,e,C} (3.9)	10.7 ^{a,c,d,e} (4.4)	12.8 ^{a,b,d,e,A} (4.6)	9.2 ^{a,b,c,A} (2.9)	9.8 ^{a,b,c} (4.7)	10.1 ^{A,C} (4.4)
(C) 16	7.1 ^{b,c,d,e,A,B} (2.6)	11.1 ^{a,c,d,e} (5.2)	12.2 ^{a,b,d,e,A} (5.9)	8.7 ^{a,b,c,A} (3.5)	9.1 ^{a,b,c,A} (4.1)	9.6 ^{A,B} (4.7)
Total	8.0 ^{b,c,d,e} (3.6)	11.0 ^{a,c,d,e} (4.5)	13.2 ^{a,b,d,e} (5.8)	9.4 ^{a,b,c} (3.6)	9.7 ^{a,b,c} (4.8)	10.2 (4.8)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale lengths, i.e., compared to the 6-item (A), 10-item (B), and 16-item scale (C).

In Experiment 1.3, client-side response times were captured using JavaScript. By means of client-side response times, the processing of a set of several rating scale items presented in a multiple-item-per-screen design could be split up into its single components, whereby the process of scale completion

⁴⁶ One-way ANOVAs indicated a significant effect of scale length in the grid ($F(2, 970) = 12.57, p < .001, \eta^2 = .025$), drag-item ($F(2, 879) = 15.57, p < .001, \eta^2 = .034$), one-vertical ($F(2, 976) = 12.47, p < .001, \eta^2 = .025$), and one-horizontal format ($F(2, 1,017) = 5.61, p < .01, \eta^2 = .011$). The effect was non-significant in the drag-response format ($F(2, 811) = 0.72, ns$).

could be examined in greater detail. More precisely, besides item-response times, the initial-reaction times could be assessed in terms of the time span between page loading and a respondent's first action. Initial-reaction times were expected to be highest in both drag-and-drop scales since the respondents' initial orientation on the screen and the basic understanding of the rating scale might take the longest time. Furthermore, the drag-item format was expected to have a longer initial-reaction time than the drag-response format due to the fact that the respondents might need even more time for the initial orientation and basic understanding when answering the drag-item format than the drag-response format. Moreover, adjusted item-response times were calculated by subtracting the initial-reaction time from the total time for scale completion. This measure was based on the assumption that after an initial orientation on the screen, i.e., after reading and understanding the general question request and after becoming more acquainted with the navigational requirements of the format, the remaining time spent on a rating scale would be indicative of the core time spent on the cognitive processing of the rating scale content. Adjusted item-response times were still expected to be higher in both drag-and-drop scales compared to conventional radio button scales since the respondents would spend more time on the cognitive processing of the rating scale content. Furthermore, the time for mere response execution was believed to be higher in both drag-and-drop scales than in conventional radio button scales. The drag-response and drag-item format were expected to be no longer different when considering adjusted item-response times because in both drag-and-drop scales respondents were expected to spend comparable time on the cognitive processing of rating scale contents.

In Experiment 1.3, cases with at least one missing value in either the 6-item scale, the 10-item scale, or the 16-item scale were excluded from analysis of client-side response times⁴⁷. Additionally, 155 cases had to be excluded because of partial errors in response time records. Unreasonably high response times were removed by excluding cases that exceeded session timeouts due to interruptions on the target page, resulting in a total of 5218 cases the analyses of client-side response times were based on.

⁴⁷ This procedure resulted in an exclusion of further 129 cases that have not completed all 10 items of the 10-item scale ($n = 36$) or all 16 items of the 16-item scale ($n = 93$) resulting in a total of 5400 cases.

After excluding cases with an item-response time ± 2 standard deviations from group mean⁴⁸, it showed that across the five scale formats and three scales lengths, respondents needed on average 9.5 seconds to answer a single item. Average item-response times ranged between 7.2 and 14.1 seconds as illustrated in Table 41. Findings of a two-way ANOVA revealed a significant main effect of scale format ($F(4, 5,075) = 216.00, p < .001, \eta^2 = .146$), a significant main effect of scale length ($F(2, 5,075) = 75.97, p < .001, \eta^2 = .029$), and a significant two-way interaction between scale format and scale length ($F(8, 5,075) = 3.02, p < .01, \eta^2 = .005$). As expected, item-response times were highest in the drag-item format (12.5 sec) which significantly differed from the grid (8.4 sec, $p < .001$), drag-response (11.2 sec, $p < .001$), one-vertical (8.2 sec, $p < .001$), and one-horizontal format (8.0 sec, $p < .001$). Also consistent with the expectations, the drag-response format had a significantly longer item-response time than the grid ($p < .001$), one-vertical ($p < .001$), and one-horizontal format ($p < .001$). As indicated in Table 41, these significant differences were also found in separate analyses of the 6-item ($F(4, 1,719) = 82.23, p < .001, \eta^2 = .161$), the 10-item ($F(4, 1,717) = 63.68, p < .001, \eta^2 = .130$), and 16-item scale ($F(4, 1,639) = 75.20, p < .001, \eta^2 = .155$).

Consistent with prior expectations and in line with findings of Experiment 1.2, item-response times significantly differed depending on scale length with the 6-item scale (10.5 sec) yielding the longest item-response time, followed by the 10-item (9.3 sec, $p < .001$) and 16-item scale (8.6 sec, $p < .001$, respectively). This decrease in response times with an increasing number of rating scale items was also found in separate analyses of the five scale formats⁴⁹. A significant interaction between scale format and scale

⁴⁸ It should be added that the results on client-side item-response times were largely consistent with the analyses of server-side item-response times. Only both single-item-per-screen formats no longer differed from the grid format when taking client-side records as a basis for response time calculations. This was hardly surprising since the present recordings of client-side response times within a Web page stopped with the last item being clicked before the 'Continue' button was pressed, whereby the time between the last answer and the 'Continue' button was not included in client-side response time calculations which considerably reduced item-response times in the one-vertical and one-horizontal format.

⁴⁹ One-way ANOVAs indicated a significant effect of scale length in the grid ($F(2, 1,104) = 6.26, p < .01, \eta^2 = .011$), drag-response ($F(2, 884) = 14.07, p < .001, \eta^2 = .031$), drag-

length suggested that compared to all the other scale formats, the grid format took least time in a 6-item scale, whereas in the 10-item and 16-item scale, both single-item-per-screen formats were answered most rapidly compared to all the other scale formats.

Table 41: Client-side item-response time (mean in seconds) depending on scale format and scale length (Experiment 1.3, $n = 5,075$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 6	8.8 ^{b,c,C} (3.4)	12.3 ^{a,c,d,e,B,C} (5.6)	14.1 ^{a,b,d,e,B,C} (5.8)	9.3 ^{b,c,B,C} (5.1)	8.9 ^{b,c,B,C} (3.6)	10.5 ^{B,C} (5.2)
(B) 10	8.5 ^{b,c} (4.0)	10.6 ^{a,c,d,e,A} (4.0)	12.0 ^{a,b,d,e,A} (6.1)	7.9 ^{b,c,A} (2.9)	8.0 ^{b,c,A,C} (3.3)	9.3 ^{A,C} (4.4)
(C) 16	7.9 ^{b,c,A} (3.6)	10.6 ^{a,d,e,A} (3.3)	11.4 ^{a,d,e,A} (5.6)	7.5 ^{b,c,A} (3.7)	7.2 ^{b,c,A,B} (2.7)	8.6 ^{A,B} (4.1)
Total	8.4 ^{b,c} (3.7)	11.2 ^{a,c,d,e} (4.5)	12.5 ^{a,b,d,e} (5.9)	8.2 ^{b,c} (4.1)	8.0 ^{b,c} (3.3)	9.5 (4.6)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale lengths, i.e., compared to the 6-item (A), 10-item (B), and 16-item scale (C).

Initial-reaction times were examined as a function of scale format and scale length. As indicated in Figure 9, initial-reaction times—after excluding cases with a time ± 2 standard deviations from group mean—considerably differed as a function of scale format. A two-way ANOVA revealed a significant main effect of scale format ($F(4, 5,127) = 162.28, p < .001, \eta^2 = .113$) with a Bonferroni post-hoc test expectably showing that the drag-item format (28.0 sec) was significantly different from the grid (16.0 sec, $p < .001$), one-vertical (15.4 sec, $p < .001$), and one-horizontal format (14.9 sec, $p < .001$). Moreover, the initial-reaction time in the drag-item format was significantly longer than in the drag-response format (20.4 sec, $p < .001$), as expected. Also, in line with the expectations, the drag-response format differed from the grid ($p < .001$), one-vertical ($p < .001$), and one-horizontal format ($p < .001$). Similar differences were found in separate analyses of the 6-item ($F(4, 1,725)$

item ($F(2, 852) = 17.02, p < .001, \eta^2 = .039$), one-vertical ($F(2, 1,128) = 22.05, p < .001, \eta^2 = .038$), and one-horizontal format ($F(2, 1,107) = 22.99, p < .001, \eta^2 = .040$).

= 77.43, $p < .001$, $\eta^2 = .153$), the 10-item ($F(4, 1,746) = 66.83$, $p < .001$, $\eta^2 = .133$), and the 16-item scale ($F(4, 1,656) = 29.98$, $p < .001$, $\eta^2 = .068$).

While the overall main effect of scale length was non-significant ($F(2, 5,127) = 0.64$, *ns*), the interaction between scale format and scale length reached significance ($F(8, 5,127) = 2.25$, $p < .05$, $\eta^2 = .004$): Merely in the drag-item format, initial-reaction times significantly differed depending on scale length ($F(2, 852) = 3.04$, $p < .05$, $\eta^2 = .007$) in terms of a shorter time in the 16-item scale (25.8 sec) compared to the 6-item scale (29.5 sec, $p < .05$), whereas the 10-item scale did not differ from either of those two (28.3 sec, *ns*, respectively). This finding was counter-intuitive because in the drag-item format, the initial visual appearance of the rating scale is exactly the same (except for small figures indicating the total number of items), irrespective of a varying number of rating scale items.

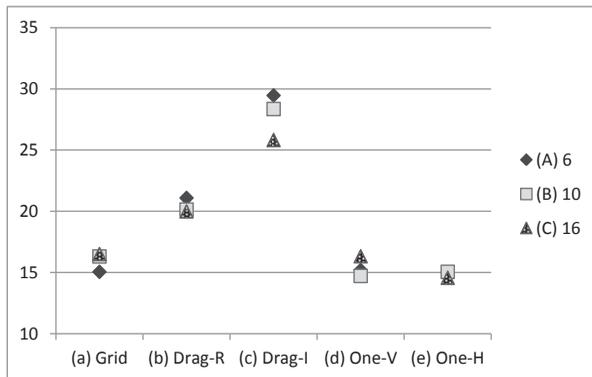


Figure 9: Initial-reaction time (mean in seconds) depending on scale format and scale length (Experiment 1.3, $n = 5,127$)

Concerning the adjusted item-response times leaving initial-reaction times out of consideration, respondents needed on average 7.2 seconds to answer a single rating scale item (cases with an adjusted item-response time ± 2 standard deviations from group mean were excluded) (see Table 42). Thus, the overall adjusted item-response time was reduced by 2.3 seconds compared to the overall item-response time including the time for initial orientation. Findings of an ANOVA revealed a significant main effect of scale format ($F(4, 5,055) = 209.29$, $p < .001$, $\eta^2 = .142$) and scale length ($F(2, 5,055) = 9.36$, $p < .001$, $\eta^2 = .004$), as well as a significant two-way interaction between scale format and scale length ($F(8, 5,055) = 3.26$, $p < .01$, $\eta^2 = .005$).

Although there was a certain convergence due to the proportionally higher initial-reaction times in the drag-and-drop scales, respondents in both the drag-response (8.8 sec) and drag-item format (9.1 sec) still had a significantly longer adjusted item-response time compared to the grid (6.5 sec, $p < .001$), one-vertical (6.2 sec, $p < .001$), and one-horizontal format (6.2 sec, $p < .001$), respectively. This finding was consistent with prior expectations. Also, as expected, the drag-item and drag-response format no longer differed when considering adjusted item-response times. These differences were also found in separate analyses of the relationship between scale format and adjusted item-response times in the 6-item ($F(4, 1,707) = 62.72, p < .001, \eta^2 = .128$), the 10-item ($F(4, 1,707) = 63.32, p < .001, \eta^2 = .130$), and the 16-item scale ($F(4, 1,641) = 86.05, p < .001, \eta^2 = .174$).

Regarding scale length, it became obvious that—as distinct from overall item-response times—the adjusted item-response time was significantly higher in the 16-item scale (7.4 sec) compared to the 6-item scale (7.1 sec, $p < .05$). This inversion of rank order for differing scale lengths was due to the fact that although all three scale lengths had approximately the same initial-reaction times, their share on the total scale completion time carried less weight with increasing scale length. Individual one-way ANOVAs for each of the five scale formats were conducted indicating a significant effect of scale length for the grid ($F(2, 1,099) = 6.34, p < .01, \eta^2 = .011$) and drag-response format ($F(2, 888) = 6.95, p < .01, \eta^2 = .015$), whereas the effect was non-significant for the drag-item ($F(2, 851) = 2.25, ns$), one-vertical ($F(2, 1,121) = 1.53, ns$), and one-horizontal format ($F(2, 1,096) = 2.96, ns$). Thus, even if differences in initial-reaction times were subtracted, the differences between the scale formats persisted with both drag-and-drop formats taking significantly more time compared to a grid and the single-item-per-screen formats.

Table 42: Adjusted item-response time (mean in seconds) depending on scale format and scale length (Experiment 1.3, $n = 5,055$)

Length	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) 6	6.1 ^{b,c,B,C} (2.4)	8.5 ^{a,d,e,C} (4.0)	8.9 ^{a,d,e} (3.4)	6.4 ^{b,c} (3.7)	6.0 ^{b,c} (2.1)	7.1 ^C (3.4)
(B) 10	6.7 ^{b,c,A} (2.8)	8.4 ^{a,d,e,C} (3.1)	9.0 ^{a,d,e} (4.2)	6.1 ^{b,c} (2.0)	6.4 ^{b,c} (2.4)	7.2 (3.1)
(C) 16	6.8 ^{b,c,A} (3.0)	9.4 ^{a,d,e,A,B} (3.0)	9.6 ^{a,d,e} (5.2)	6.1 ^{b,c} (2.3)	6.3 ^{b,c} (2.4)	7.4 ^A (3.5)
Total	6.5 ^{b,c} (2.7)	8.8 ^{a,d,e} (3.5)	9.1 ^{a,d,e} (4.3)	6.2 ^{b,c} (2.8)	6.2 ^{b,c} (2.3)	7.2 (3.4)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale lengths, i.e., compared to the 6-item (A), 10-item (B), and 16-item scale (C).

9.10.2 Experiment 2

In Experiment 2, the impact of scale arrangement and scale format on response times in rating scales was examined in a 2 x 5 between-subjects factorial design. The time taken to complete a rating scale was analyzed in terms of server-side item-response times. As already hypothesized in Experiment 1, both drag-and-drop formats were expected to yield longer item-response times relative to the grid format. Furthermore, the drag-item format was expected to need longer than the drag-response format. Variations in the scale arrangement were expected to result in different response times since the negative-to-positive response option order was expected to take longer to be completed compared to the positive-to-negative response option order. Thus, beginning with the negative response option might yield a longer item-response time than beginning with the positive response option. A two-way interaction between scale format and scale arrangement on item-response time was not expected.

For the analyses of server-side item-response times, cases with one or more missing values in the rating scale were excluded. The additional exclusion of cases that exceeded the session timeout and had an item-response time ± 2 standard deviations from group mean resulted in a total of 703 cases. Item-response times as a function of scale format and scale arrangement are

displayed in Table 43. Average item-response times ranged between 5.0 seconds up to 9.0 seconds. Findings of a two-way ANOVA revealed a significant main effect of scale format ($F(4, 703) = 25.33, p < .001, \eta^2 = .128$). Consistent with prior expectations, a Bonferroni post-hoc test indicated a significantly higher item-response time in the drag-item format (9.0 sec) compared to the grid (5.0 sec, $p < .001$), drag-response (7.4 sec, $p < .001$), one-vertical (6.8 sec, $p < .001$), and one-horizontal format (6.9 sec, $p < .001$). Moreover, the drag-response format revealed a significantly higher item-response time than the grid format ($p < .001$) but did not differ from both single-item-per-screen formats. Thus, again, longest item-response times were found for the drag-item format followed by the drag-response format and conventional radio button scales. These differences were also found with separate analyses of the positive-to-negative ($F(4, 358) = 13.85, p < .001, \eta^2 = .136$) and negative-to-positive response options order ($F(4, 345) = 12.26, p < .001, \eta^2 = .126$).

Contrary to expectations, item-response times did not vary as a function of scale arrangement ($F(1, 703) = 0.06, ns$)⁵⁰. In line with previous expectations, the two-way interaction between scale format and scale arrangement was non-significant as well ($F(4, 703) = 0.70, ns$). Thus, in both drag-and-drop scales, respondents took longest to answer the items with differences to other scale formats arising irrespective of the categorial response option order.

⁵⁰ Separate analyses of the relationship between scale arrangement and item-response times were also non-significant in the grid ($F(1, 136) = 0.11, ns$), drag-response ($F(1, 148) = 1.23, ns$), drag-item ($F(1, 124) = 0.01, ns$), one-vertical ($F(1, 163) = 0.02, ns$), and one-horizontal format ($F(1, 131) = 1.29, ns$).

Table 43: Server-side item-response time (mean in seconds) depending on scale format and scale arrangement (Experiment 2, $n = 703$)

Arrangement	Format					Total
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H	
(A) Pos-Neg	5.0 ^{b,c,d} (2.4)	7.6 ^a (3.5)	8.9 ^{a,d,e} (4.1)	6.9 ^{a,c} (2.9)	6.5 ^c (3.2)	7.2 (3.5)
(B) Neg-Pos	5.1 ^{b,c,d,e} (2.4)	7.1 ^{a,c} (2.5)	9.0 ^{a,b,d,e} (4.4)	6.8 ^{a,c} (2.4)	7.2 ^{a,c} (3.9)	6.9 (3.3)
Total	5.0 ^{b,c,d,e} (2.4)	7.4 ^{a,c} (3.0)	9.0 ^{a,b,d,e} (4.2)	6.8 ^{a,c} (2.7)	6.9 ^{a,c} (3.6)	7.0 (3.4)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e., compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e).

9.10.3 Experiment 3

In Experiment 3, differences in the time spent on answering a rating scale were evaluated depending on scale arrangement and scale format in a 2×3 between-subjects factorial design⁵¹. Just like in Experiment 1.3, client-side response times were recorded by the use of JavaScript. Thus, the item-response time was calculated as the sum of all time measurements within a rating scale divided by the respective number of rating scale items. Furthermore, initial-reaction times and adjusted item-response times could be computed. Concerning the grid, drag-item, and drag-response format, it was again hypothesized that both drag-and-drop formats would yield a longer item-response time compared to the grid format, while the drag-item format was expected to take even longer than the drag-response format. As already suggested in Experiment 2, the negative-to-positive response option order was supposed to yield a longer item-response time as compared to the positive-to-negative response option order. A significant two-way interaction between scale format and scale arrangement was not expected. Also, the drag-item format was expected to yield the longest initial-reaction time, followed by the drag-response format. The grid format was presumed to need significantly

⁵¹ Even though scale sequence was an inherent factor of Experiment 3 that was originally implemented as a $2 \times 2 \times 3$ between-subjects factorial design, the factor scale sequence was left out of subsequent response time analyses since variations in scale sequence were not expected to have a systematic effect on the time respondents spent on answering a rating scale.

less time for initial orientation. By implication, both drag-and-drop scales were expected to have a significantly longer adjusted item-response time than the grid format, whereas the drag-response and drag-item format might show adjusted item-response times at a similar level.

In Experiments 3.1 and 3.2, cases with one or more missing values were excluded from response time analyses. Because of partial errors in the response time records, 212 cases had to be excluded from Experiment 3.1 and 194 cases from Experiment 3.2. In addition, unreasonably high response times were removed by excluding cases that exceeded the session timeouts due to interruptions on the target page, resulting in a total of 4,995 cases in Experiment 3.1 and 5,020 cases in Experiment 3.2.

In Experiment 3.1, respondents needed 8.5 seconds on average to answer a single rating scale item after excluding cases with an item-response time ± 2 standard deviations from group mean⁵². The average item-response times ranged between 6.4 and 10.4 seconds as illustrated in Table 44. Findings of a two-way ANOVA revealed a significant main effect of scale format ($F(2, 4,916) = 286.19, p < .001, \eta^2 = .104$) and a significant main effect of scale arrangement ($F(1, 4,916) = 31.92, p < .001, \eta^2 = .006$). The two-way interaction between scale format and scale arrangement was non-significant ($F(2, 4,916) = 0.19, ns$). As expected, a Bonferroni post-hoc test revealed a significantly higher item-response time in the drag-item format (10.0 sec) compared to the drag-response format (9.1 sec, $p < .001$). Both drag-and-drop scales yielded a significantly higher item-response time than the grid format (6.7 sec, $p < .001$, respectively). These differences were also found in separate analyses of the positive-to-negative ($F(2, 2,463) = 156.95, p < .001, \eta^2 = .113$) and negative-to-positive scale arrangement ($F(2, 2,453) = 132.36, p < .001, \eta^2 = .098$). Also consistent with prior expectations, the negative-to-positive response option order resulted in significantly longer item-response times compared to the positive-to-negative response option order. This equally applied to the grid ($F(1, 1,815) = 14.34, p < .001, \eta^2 = .008$), drag-response ($F(1, 1,597) = 11.11, p < .01, \eta^2 = .007$), and drag-item format ($F(1, 1,504) = 7.95, p < .01, \eta^2 = .005$).

⁵² The average item-response time across all experimental conditions somewhat declined from 9.2 seconds with server-side response times, however, largely the same patterns were found depending on scale format and scale arrangement.

Table 44: Client-side item-response time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.1, $n = 4,916$)

Arrangement	Format			Total
	(a) Grid	(b) Drag-R	(c) Drag-I	
(A) Pos-Neg	6.4 ^{b,c,B} (3.1)	8.7 ^{a,c,B} (3.8)	9.7 ^{a,b,B} (4.7)	8.2 ^B (4.1)
(B) Neg-Pos	7.0 ^{b,c,A} (3.4)	9.5 ^{a,c,A} (5.2)	10.4 ^{a,b,A} (4.5)	8.8 ^A (4.6)
Total	6.7 ^{b,c} (3.3)	9.1 ^{a,c} (4.6)	10.0 ^{a,b} (4.7)	8.5 (4.4)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale formats, i.e., compared to grid (a), drag-response (b), and drag-item format (c). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the positive-to-negative (A) and negative-to-positive (B) scale arrangement.

The initial-reaction time was examined as a function of scale format and scale arrangement, excluding cases that had an initial-reaction time ± 2 standard deviations from group mean. The three scale formats considerably differed from each other (see Figure 10). A two-way ANOVA revealed a significant main effect of scale format ($F(2, 4,946) = 111.50, p < .001, \eta^2 = .043$) with all three formats differing from each other significantly (all comparisons $p < .001$). Thus, as previously expected, the drag-item format featured the highest initial-reaction time (28.0 sec), followed by the drag-response (22.5 sec) and grid format (18.3 sec). This pattern was exactly the same in separate analyses of the positive-to-negative ($F(2, 2,476) = 68.64, p < .001, \eta^2 = .053$) and negative-to-positive scale arrangement ($F(2, 2,470) = 46.85, p < .001, \eta^2 = .037$).

Although a significant main effect of scale arrangement ($F(1, 4,946) = 5.64, p < .05, \eta^2 = .001$) indicated a longer initial-reaction time in the negative-to-positive (23.2 sec) compared to a positive-to-negative response option order (22.0 sec), a significant difference between the positive-to-negative response (21.2 sec) and negative-to-positive scale arrangement (23.8 sec) was merely found for the drag-response format ($F(1, 1,616) = 7.83, p < .01, \eta^2 = .005$). By contrast, respective differences were non-significant in separate analyses of the grid ($F(1, 1,818) = 2.14, ns$) and drag-item format (F

(1, 1,511) = 0.03, *ns*). Furthermore, the interaction between scale format and scale arrangement was non-significant ($F(2, 4,946) = 0.19, ns$).

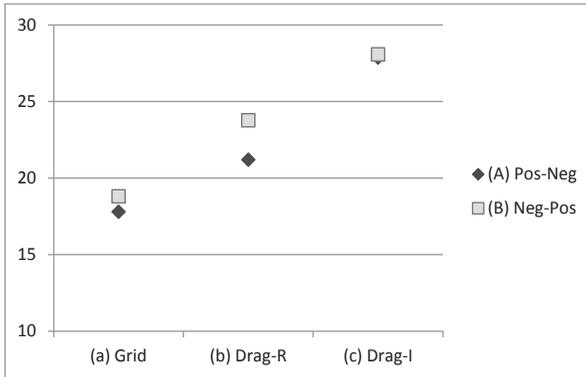


Figure 10: Initial-reaction time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.1, $n = 4,946$)

The analysis of adjusted item-response times showed that respondents needed on average 5.5 seconds to answer a single rating scale item when leaving the initial-reaction time out of account (cases with an adjusted item-response time ± 2 standard deviations from group mean were excluded) (see Table 45). This corresponded to a reduction of 3.0 seconds compared to the overall item-response time including initial-reaction time. Significant main effects of scale format ($F(2, 4,889) = 329.23, p < .001, \eta^2 = .119$) and scale arrangement ($F(1, 4,889) = 37.57, p < .001, \eta^2 = .008$) were found in a two-way ANOVA, whereas the interaction between both factors was non-significant ($F(2, 4,889) = 0.14, ns$). As indicated by a Bonferroni post-hoc test, all three formats significantly differed from each other with the drag-item format yielding the highest adjusted item-response time (6.3 sec), followed by the drag-response format (6.0 sec, $p < .01$ as against drag-item, $p < .001$ as against grid), and the grid format (4.3 sec, $p < .001$, respectively). Thus, as expected, both drag-and-drop scales took significantly longer than the grid format. However, contrary to expectation, the drag-item format took slightly but significantly longer than the drag-response format. Taking account of separate analyses of the positive-to-negative ($F(2, 2,456) = 178.65, p < .001, \eta^2 = .127$) and negative-to-positive scale arrangement ($F(2, 2,433) = 153.25, p < .001, \eta^2 = .112$), differences between the grid and both drag-and-drop formats persisted,

whereas differences between the drag-item and drag-response format were no longer significant (see Table 45).

Concerning the overall effect of scale arrangement, the initial expectations could be confirmed since a significantly longer adjusted item-response time was found in the negative-to-positive (5.7 sec) compared to the positive-to-negative response option order (5.3 sec). This difference equally emerged with separate analysis of the grid ($F(1, 1,821) = 18.35, p < .001, \eta^2 = .010$), drag-response ($F(1, 1,582) = 8.79, p < .01, \eta^2 = .006$), and drag-item format ($F(1, 1,486) = 13.33, p < .001, \eta^2 = .009$).

Table 45: Adjusted item-response time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.1, $n = 4,889$)

Arrangement	Format			Total
	(a) Grid	(b) Drag-R	(c) Drag-I	
(A) Pos-Neg	4.1 ^{b,c,B} (1.8)	5.8 ^{a,B} (2.6)	6.1 ^{a,B} (2.6)	5.3 ^B (2.5)
(B) Neg-Pos	4.5 ^{b,c,A} (2.1)	6.3 ^{a,A} (3.2)	6.6 ^{a,A} (2.5)	5.7 ^A (2.8)
Total	4.3 ^{b,c} (2.0)	6.0 ^{a,c} (2.9)	6.3 ^{a,b} (2.6)	5.5 (2.7)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale formats, i.e., compared to grid (a), drag-response (b), and drag-item format (c). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the positive-to-negative (A) and negative-to-positive (B) scale arrangement.

In Experiment 3.2, respondents needed 9.6 seconds on average to answer a single rating scale item after excluding cases with an item-response time ± 2 standard deviations from group mean⁵³. Average item-response times ranged between 8.5 and 10.5 seconds as illustrated in Table 46. Very similar to the previous findings, an ANOVA revealed a significant main effect of scale format ($F(2, 4,943) = 57.20, p < .001, \eta^2 = .023$) and scale arrangement ($F(1, 4,943) = 8.13, p < .01, \eta^2 = .002$). The two-way interaction between scale format and scale arrangement was non-significant ($F(2, 4,943) = 1.83, ns$).

⁵³ The average item-response time across all experimental conditions somewhat declined from 10.2 seconds with server-side response times, however, largely the same patterns were found depending on scale format and scale arrangement.

As expected, the drag-item format (10.4 sec), once again, yielded a significantly longer item-response time compared to the drag-response format (9.9 sec, $p < .01$), whereas both drag-and-drop scales had a significantly higher item-response time than the grid format (8.8 sec, $p < .001$, respectively). Separate analyses of the relationship between scale format and item-response time found largely the same pattern within the positive-to-negative ($F(2, 2,490) = 45.10, p < .001, \eta^2 = .05$) and negative-to-positive scale arrangement ($F(2, 2,453) = 18.02, p < .001, \eta^2 = .05$).

Also consistent with the expectations, the negative-to-positive response option order resulted in significantly longer item-response times compared to the positive-to-negative response option order. However, in separate analyses, this effect of scale length solely applied to the grid format ($F(1, 1,864) = 11.36, p < .01, \eta^2 = .006$) but was not found in the drag-response ($F(1, 1,560) = 0.36, ns$) or drag-item format ($F(1, 1,519) = 1.26, ns$).

Table 46: Client-side item-response time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.2, $n = 4,943$)

Arrangement	Format			Total
	(a) Grid	(b) Drag-R	(c) Drag-I	
(A) Pos-Neg	8.5 ^{b,c,B} (3.8)	9.8 ^a (4.1)	10.3 ^a (4.5)	9.4 ^B (4.2)
(B) Neg-Pos	9.2 ^{b,c,A} (5.0)	9.9 ^{a,c} (4.2)	10.5 ^{a,b} (5.0)	9.8 ^A (4.8)
Total	8.8 ^{b,c} (4.4)	9.9 ^{a,c} (4.2)	10.4 ^{a,b} (4.8)	9.6 (4.5)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale formats, i.e., compared to grid (a), drag-response (b), and drag-item format (c). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the positive-to-negative (A) and negative-to-positive (B) scale arrangement.

The examination of initial-reaction times revealed that across all experimental conditions, an average period of 17.7 seconds lay in between page loading and a respondent's first action as depicted in Figure 11 (with cases being excluded that had an initial-reaction time ± 2 standard deviations from group mean). In accordance with Experiments 1.3 and 3.1, a significant main effect of scale format on initial-reaction times was found ($F(2, 4,969) = 9.88, p <$

.001, $\eta^2 = .004$). However, differences between the scale formats have been decreased and deviations from previous results were evident. Instead of producing significantly longer initial-reaction times, the drag-response format (16.3 sec) featured a significantly shorter initial-reaction time compared to the grid format (17.9 sec, $p < .05$) as well as compared to the drag-item format (19.0 sec, $p < .001$). While the latter finding was consistent with prior expectations, a shorter initial-reaction time in the drag-response compared to the grid format was striking. Equally surprising was a non-significant difference between the drag-item format and the grid format ($p = .166$).

In view of a non-significant main effect of scale arrangement ($F(1, 4,969) = 2.07$, *ns*) and a significant interaction effect between scale format and scale arrangement ($F(2, 4,969) = 3.26$, $p < .05$, $\eta^2 = .001$), separate analyses of the three scale formats contributed to an explanation of the unexpected differences between both drag-and-drop scales and the grid format: A clearly higher initial-reaction time was found in the grid format with a negative-to-positive (19.0 sec) compared to a positive-to-negative scale arrangement (16.8 sec, $F(1, 1,877) = 5.76$, $p < .05$, $\eta^2 = .003$), whereas no significant effect of scale arrangement was found for the drag-response ($F(1, 1,574) = 1.28$, *ns*) and drag-item format ($F(1, 1,518) = 0.40$, *ns*). Separate analyses revealed a significant effect of scale format in both the positive-to-negative ($F(2, 2,505) = 4.78$, $p < .01$, $\eta^2 = .004$) and negative-to-positive scale arrangement ($F(2, 2,464) = 7.68$, $p < .001$, $\eta^2 = .006$). However, a significant difference between the drag-item (18.7 sec) and grid format (16.8 sec, $p < .05$) was merely found with a positive-to-negative response option order. Besides, a significant difference between the drag-response (16.0 sec) and grid format (19.0 sec, $p < .01$) was solely found in the negative-to-positive scale arrangement. Thus, divergent results were primarily due to a comparatively high initial-reaction time in the grid format when response options were arranged in a negative-to-positive order.

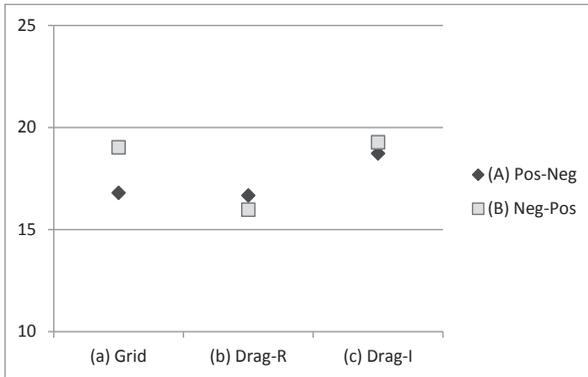


Figure 11: Initial-reaction time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.2, $n = 4,969$)

The examination of adjusted item-response times showed that on average, respondents needed 7.3 seconds to answer a single rating scale item when initial-reaction times were excluded from calculations (cases with an adjusted item-response time ± 2 standard deviations from group mean were excluded) (see Table 47). Thus, the overall adjusted item-response time was again reduced by 2.3 seconds compared to the overall item-response time including time for initial orientation. Findings of an ANOVA revealed a significant main effect of scale format ($F(2, 4,955) = 105.74, p < .001, \eta^2 = .041$) and scale arrangement ($F(1, 4,955) = 5.04, p < .05, \eta^2 = .001$), whereas the two-way interaction between scale format and scale arrangement was non-significant ($F(2, 4,955) = 0.40, ns$). Consistent with prior expectations, both the drag-response (7.7 sec) and drag-item (7.9 sec) format had a significantly longer adjusted item-response time than the grid format (6.4 sec, $p < .001$, respectively). Also, as expected, both drag-and-drop scales did not differ in terms of their adjusted item-response times ($p = .111$). The same patterns were found in separate analyses of the positive-to-negative ($F(2, 2,501) = 58.75, p < .001, \eta^2 = .045$) and negative-to-positive scale arrangement ($F(2, 2,454) = 47.92, p < .001, \eta^2 = .038$). Despite a significant main effect of scale arrangement, no significant differences between a positive-to-negative and negative-to-positive response option order were found in separate analyses of the grid ($F(1, 1,867) = 3.04, ns$), drag-response ($F(1, 1,563) = 3.00, ns$), and drag-item format ($F(1, 1,525) = 0.27, ns$).

Table 47: Adjusted item-response time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.2, $n = 4,995$)

Arrangement	Format			Total
	(a) Grid	(b) Drag-R	(c) Drag-I	
(A) Pos-Neg	6.3 ^{b,c} (2.8)	7.5 ^a (3.1)	7.9 ^a (3.5)	7.2 ^B (3.2)
(B) Neg-Pos	6.6 ^{b,c} (2.9)	7.8 ^a (3.5)	8.0 ^a (3.3)	7.4 ^A (3.3)
Total	6.4 ^{b,c} (2.8)	7.7 ^a (3.3)	7.9 ^a (3.4)	7.3 (3.2)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale formats, i.e., compared to grid (a), drag-response (b), and drag-item format (c). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the positive-to-negative (A) and negative-to-positive (B) scale arrangement.

The examination of item-response times clearly showed that both drag-and-drop scales needed significantly more time to be completed than conventional rating scales using radio buttons as input method. Generally, it needs to be noted that particularly in Experiments 1 and 2, there was a considerable variation in the amount of time spent on answering a rating scale item in the drag-item format, indicated by consistently higher standard deviations, compared to the well-established radio button scales. This also applied to the drag-response format, albeit to a lesser extent. A comparison of both drag-and-drop scales also revealed that the respondents in the drag-item format needed at first sight significantly longer to answer a set of several rating scale items than respondents in the drag-response format. However, by excluding the initial-reaction times with the calculation of the adjusted item-response times, it was shown that this difference between both drag-and-drop scales was primarily due to respondents spending more time on initial orientation and basic understanding in the drag-item format compared to the drag-response format.

9.10.4 Dragging Times

By means of client-side response times, dragging times could be captured in both drag-and-drop scales, indicating the time span between a respondent's selection of a question-element or answer-element, dragging it to the desired position, and dropping it again to indicate an answer. Dragging times might give some further indication of the respondents' processing of the drag-and-drop scales and might also reveal some differences between the drag-response and drag-item format concerning the navigational and cognitive processing of rating scale items. No specific assumptions were made in advance.

In Experiment 1.3, an ANOVA on scale format and scale length revealed a significant main effect of scale format ($F(1, 1,736) = 79.57, p < .001, \eta^2 = .044$), a significant main effect of scale length ($F(2, 1,736) = 7.01, p < .01, \eta^2 = .008$), and a significant interaction between scale format and scale length ($F(2, 1,736) = 8.38, p < .001, \eta^2 = .010$). Accordingly, dragging time was significantly longer in the drag-item format (2.9 sec) compared to the drag-response format (2.4 sec) which also applied to separate analyses of the 6-item ($F(1, 608) = 22.49, p < .001, \eta^2 = .036$), the 10-item ($F(1, 619) = 74.05, p < .001, \eta^2 = .107$), and the 16-item scale ($F(1, 509) = 5.98, p < .05, \eta^2 = .012$) (see Figure 12).

Concerning the main effect of scale length, dragging time was significantly longer in the 16-item scale (2.8 sec) compared to the 10-item scale (2.5 sec, $p < .001$), whereas both did not differ from the 6-item scale (2.7 sec, $p = 2.34$ and $p = .080$, respectively). Separate analyses showed a significant effect of scale length for the drag-response format ($F(2, 895) = 16.99, p < .001, \eta^2 = .037$), whereas the effect of scale length for the drag-item format was non-significant ($F(2, 841) = 0.89, ns$) (see Figure 12). Interestingly, the drag-response format showed a significantly longer dragging time in the 16-item scale (2.7 sec) compared to the 6-item (2.4 sec, $p < .05$) and 10-item scale (2.1 sec, $p < .001$). Furthermore, dragging times were significantly longer in the 6-item scale compared to the 10-item scale ($p < .01$). Thus, in the drag-response format, most time on dragging was spent in the 16-item scale, followed by the 6-item scale, whereas least time was spent in the 10-item scale. By contrast, in the drag-item format, dragging time was unaffected by scale length.

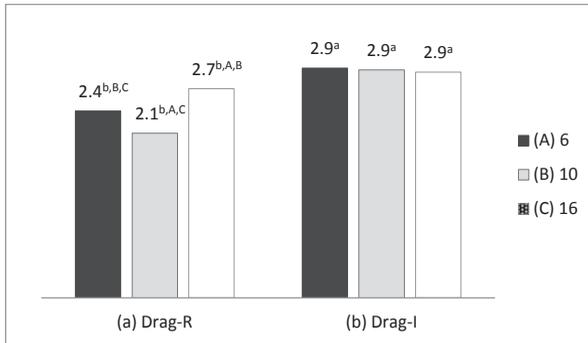


Figure 12: Dragging time (mean in seconds) depending on scale format and scale length (Experiment 1.3, $n = 1,736$)

Calculations were based on a two-way ANOVA with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between the drag-response (a) and drag-item format (b). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale lengths, i.e., compared to the 6-item (A), 10-item (B), and 16-item scale (C).

In Experiment 3.1, an ANOVA on scale format and scale arrangement revealed a significant main effect of scale format ($F(1, 3,072) = 262.45, p < .001, \eta^2 = .079$), a significant main effect of scale arrangement ($F(1, 3,072) = 36.79, p < .001, \eta^2 = .012$), and a significant interaction between scale format and scale arrangement ($F(1, 3,072) = 4.37, p < .05, \eta^2 = .001$). Again, the drag-item format (2.7 sec) showed a significantly longer dragging time compared to the drag-response format (2.0 sec) as indicated by a significant main effect as well as by separate analyses of the positive-to-negative ($F(1, 1,539) = 110.06, p < .001, \eta^2 = .067$) and negative-to-positive scale arrangement ($F(1, 1,533) = 152.57, p < .001, \eta^2 = .091$) (see Figure 13). Furthermore, with respect to the significant main effect of scale arrangement, dragging times were significantly longer with the negative-to-positive (2.4 sec) compared to the positive-to-negative scale arrangement (2.2 sec). This equally applied to the drag-response ($F(1, 1,579) = 10.04, p < .01, \eta^2 = .006$) and drag-item format ($F(1, 1,493) = 26.94, p < .001, \eta^2 = .018$) when analyzing the relationship between scale arrangement and dragging time separately for the two drag-and-drop scales (see Figure 13).

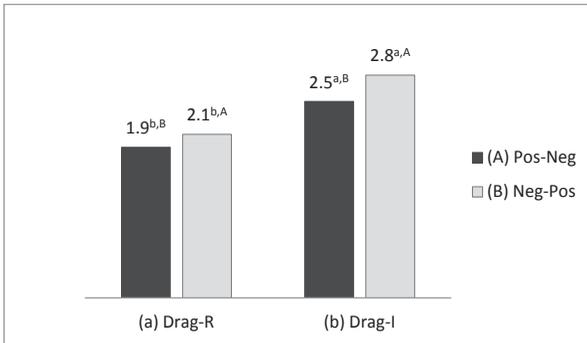


Figure 13: Dragging time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.1, $n = 3,072$)

Calculations were based on a two-way ANOVA with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between the drag-response (a) and drag-item format (b). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the positive-to-negative (A) and negative-to-positive (B) scale arrangement.

The previous findings in Experiments 1.3 and 3.1 were largely confirmed by the findings of Experiment 3.2. Again, there was a significant main effect of scale format ($F(1, 3,024) = 469.41, p < .001, \eta^2 = .135$) and scale arrangement ($F(1, 3,024) = 5.77, p < .05, \eta^2 = .002$), whereas the interaction between scale format and scale arrangement was non-significant ($F(1, 3,024) = 1.10, ns$). Corresponding to the findings of Experiments 1.3 and 3.1, the drag-item format (2.3 sec) showed a significantly longer dragging time compared to the drag-response format (1.7 sec). The same pattern was found in separate analyses of the positive-to-negative ($F(1, 1,526) = 229.63, p < .001, \eta^2 = .131$) and negative-to-positive scale arrangement ($F(1, 1,498) = 243.90, p < .001, \eta^2 = .140$) (see Figure 14). As against the findings of Experiment 3.1, a slightly but significantly shorter dragging time was found in the negative-to-positive (2.0 sec) compared to the positive-to-negative scale arrangement (2.0 sec). This difference also emerged in a separate analysis of the drag-item format ($F(1, 1,469) = 4.48, p < .05, \eta^2 = .003$), but did not apply to the drag-response format ($F(1, 1,555) = 1.30, ns$). However, small effect sizes indicated a relationship of low magnitude.

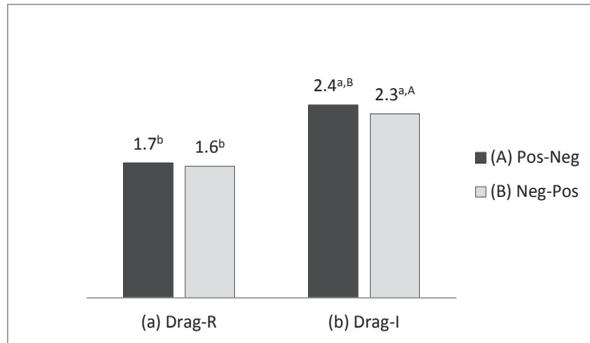


Figure 14: Dragging time (mean in seconds) depending on scale format and scale arrangement (Experiment 3.2, $n = 3,024$) Calculations were based on a two-way ANOVA with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between the drag-response (a) and drag-item format (b). Uppercase superscripts indicate a significant difference ($p < .05$ or less) between the positive-to-negative (A) and negative-to-positive (B) scale arrangement.

The examination of the dragging times of both drag-and-drop scales consistently showed, across all three experiments, that respondents spent significantly more time on dragging the question-elements in the drag-item format than on dragging the answer-elements in the drag-response format. Findings of Experiment 1.3 further revealed that dragging times in the drag-response format were clearly affected by the number of items in a rating scale: Compared to rating scales of moderate length (10 items), dragging times were longest in comparatively short (6 items) and long scales (16 items). By contrast, the drag-item format remained unaffected by scale length. Regarding the effect of scale arrangement, findings of Experiments 3.1 and 3.2 were mixed. In the first experiment, both drag-and-drop scales featured longer dragging times in the negative-to-positive compared to the positive-to-negative response option order. In the second experiment, however, the opposite applied for the drag-item format, whereas the drag-response format remained unaffected by the scale arrangement.

9.10.5 Summary

In the present experiments, the findings regarding variations in item-response times as a function of scale format consistently indicated that, irrespective of differing scale lengths, the drag-response and drag-item format took longer to complete than conventional rating scale formats such as a grid or single-item-per-screen design. This even held true when comparing adjusted item-response times where the time spent on initial orientation on the screen was disregarded. The results relating to varying scale lengths were mixed. Expectations concerning a decrease in item-response times with increasing scale length were confirmed in Experiments 1.2 and 1.3, whereas Experiment 1.1 failed to prove this relationship. Findings concerning differing scale arrangements were mixed as well. The expectations of a longer item-response time in the negative-to-positive compared to a positive-to-negative response option order could be confirmed in Experiments 3.1 and 3.2, whereas Experiment 2 revealed no significant differences between differing scale arrangements. Thus, whereas the findings on scale length and scale arrangement were rather mixed, findings concerning scale format were highly consistent in terms of clearly longer response times and higher need for initial orientation in both drag-and-drop scales.

A closer examination of the differences between the drag-response and the drag-item format showed that longer response times in the drag-item format were predominantly due to a longer time needed for initial orientation in the drag-item format compared to the drag-response format. Thus, the differences between the two drag-and-drop scales concerning overall item-response times disappeared when looking at the adjusted item-response times, disregarding the time spent on initial orientation on a screen. Dragging times reflecting the time span between dragging a question-element or answer-element and dropping it in the desired position were further examined. The findings revealed that it took significantly longer to drag the respective question-element from its initial position to the end position in the drag-item format than performing the same procedure based on the respective answer-element in the drag-response format.

9.11 Respondent Evaluation

Respondents evaluated the survey along the three dimensions of (1) navigation, (2) design, and (3) overall survey perception. A total of 13 adjective pairs were assessed on a 6-point semantic differential with higher values indicating a more positive evaluation. By means of a principal component analysis, the assignment of the adjective pairs to the respective dimensions could be confirmed. The examination of the respondents' evaluation of the various rating scale designs was exploratory; hence, no specific assumptions were made in advance.

The respondents' evaluations regarding navigation-related aspects of the survey are depicted in Figure 15. In Experiment 1.1, a two-way ANOVA indicated a significant main effect of scale format ($F(4, 677) = 7.89, p < .001, \eta^2 = .045$) with a Bonferroni post-hoc test revealing a significantly worse evaluation with respect to navigation capability and ease of use when respondents were assigned to a drag-response format compared to all the other scale formats (all comparisons $p < .05$ or less). A significant main effect of scale length ($F(1, 677) = 9.88, p < .01, \eta^2 = .015$) indicated a significantly worse evaluation when respondents were assigned to the 16-item scale compared to the 10-item scale, whereas the interaction between scale format and scale length was non-significant ($F(4, 677) = 1.57, ns$). Similarly in Experiment 1.2, a significant main effect of scale format ($F(4, 4,622) = 4.49, p < .01, \eta^2 = .004$) and scale length ($F(2, 4,622) = 4.03, p < .05, \eta^2 = .002$), as well as a significant interaction between scale format and scale length ($F(8, 4,622) = 3.44, p < .01, \eta^2 = .006$) clearly showed that the respondents' evaluation of navigational aspects of the survey was significantly worse when answering a drag-response format. This was particularly true if respondents were assigned to a drag-response format consisting of 16 items as compared to a drag-response format consisting of 6 or 10 items. Experiment 1.3 supported these findings as indicated by a significant main effect of scale format ($F(4, 5,185) = 13.50, p < .001, \eta^2 = .010$), a significant main effect of scale length ($F(2, 5,185) = 15.93, p < .001, \eta^2 = .006$), and a significant interaction between scale format and scale length ($F(8, 5,185) = 3.86, p < .001, \eta^2 = .006$). Thus, in the drag-response format, navigational aspects of the survey were evaluated as significantly worse as compared to the other rating scale formats. This worse evaluation of the drag-response format was even more pronounced with the increasing length of a rating scale. By

contrast, in Experiments 2, 3.1, and 3.2, there was neither a significant main effect of scale format or scale arrangement, nor a significant interaction between both factors⁵⁴. Thus, scale arrangement obviously had no effect on the respondents' evaluation of the navigation-related aspects of a survey.

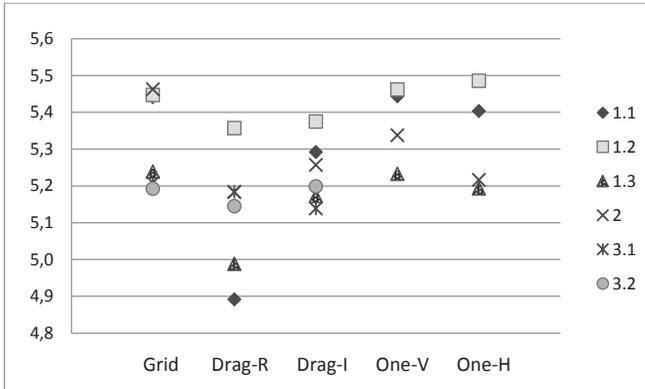


Figure 15: Respondent evaluation regarding navigation-related aspects of the survey depending on scale format, separately for the six experiments

Respondent evaluation was based on a 6-point semantic differential scale with higher values indicating a more positive evaluation.

The respondents' evaluations of design-related aspects of the survey are depicted in Figure 16, with mean ratings being displayed depending on scale format separately for the six experiments. In all six experiments, the findings of two-way ANOVAs consistently showed a significant effect of scale format on the respondents' evaluations of the survey layout in terms of its diversion and innovation⁵⁵. In Experiments 1.2, 3.1, and 3.2, Bonferroni post-hoc tests

⁵⁴ Findings of two-way ANOVAs concerning the respondents' evaluation of navigation-related survey aspects depending on scale format ($F(4, 691) = 2.02, ns$), scale arrangement ($F(1, 691) = 0.27, ns$), and an interaction ($F(4, 691) = 1.87, ns$) in Experiment 2; scale format ($F(2, 4,938) = 2.61, ns$), scale arrangement ($F(1, 4,938) = 0.26, ns$), and an interaction ($F(2, 4,938) = 1.87, ns$) in Experiment 3.1; and scale format ($F(2, 5,056) = 1.73, ns$), scale arrangement ($F(1, 5,056) = 1.44, ns$), and an interaction ($F(2, 5,056) = 2.02, ns$) in Experiment 3.2.

⁵⁵ Findings of two-way ANOVAs concerning the respondents' evaluation of design-related survey aspects for the main effects of scale format: $F(4, 675) = 3.55, p < .01, \eta^2 = .021$ in Experiment 1.1; $F(4, 4,597) = 13.06, p < .001, \eta^2 = .011$ in Experiment 1.2; $F(4, 688) = 3.98, p < .01, \eta^2 = .023$ in Experiment 2; $F(2, 4,936) = 7.44, p < .01, \eta^2 = .003$ in

clearly indicated a significantly better assessment of design-related aspects of the survey when respondents were assigned to either the drag-response format or the drag-item format than when they were assigned to the grid format (all comparisons $p < .05$ or less). In Experiments 1.3 and 2, both drag-and-drop formats were also rated more positively, although Bonferroni post-hoc tests were largely non-significant. By contrast, the respondents' evaluation of design-related aspects of the survey remained unaffected by scale length and scale arrangement since neither of these two main effects nor interactions with scale format were significant⁵⁶.

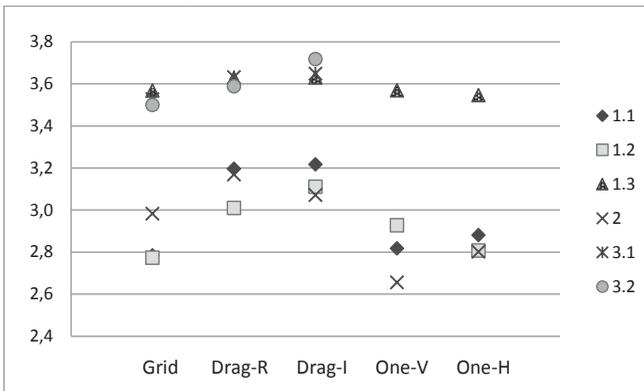


Figure 16: Respondent evaluation regarding design-related aspects of the survey depending on scale format, separately for the six experiments

Respondent evaluation was based on a 6-point semantic differential scale with higher values indicating a more positive evaluation.

Experiment 3.1; $F(2, 5,049) = 21.13, p < .001, \eta^2 = .008$ in Experiment 3.2. The effect of scale format was non-significant in Experiment 1.3: $F(4, 5,180) = 1.41, ns$.

⁵⁶ Findings of two-way ANOVAs concerning the respondents' evaluation of design-related survey aspects for the main effects of scale length ($F(1, 675) = 1.77, ns$) and an interaction between scale format and scale length ($F(4, 675) = 0.49, ns$) in Experiment 1.1; scale length ($F(2, 4,597) = 0.08, ns$) and an interaction ($F(8, 4,597) = 0.49, ns$) in Experiment 1.2; and scale length ($F(2, 5,180) = 1.54, ns$) and an interaction ($F(8, 5,180) = 0.84, ns$) in Experiment 1.3. Findings of a two-way ANOVA regarding scale arrangement ($F(1, 688) = 0.07, ns$) and an interaction between scale format and scale arrangement ($F(4, 688) = 1.50, ns$) in Experiment 2; scale arrangement ($F(1, 4,936) = 2.69, ns$) and an interaction ($F(2, 4,936) = 2.54, ns$) in Experiment 3.1; and scale arrangement ($F(1, 5,049) = 0.31, ns$) and an interaction ($F(2, 5,049) = 1.31, ns$) in Experiment 3.2.

Concerning the third evaluation dimension of overall survey perception (see Figure 17), no significant differences were found depending on scale format in all six experiments⁵⁷. Respondents who answered a rating scale with 16 items consistently reported a less positive overall survey perception than respondent who were assigned to a shorter rating scale with 6 or 10 items in Experiment 1⁵⁸. No major differences were found depending on scale arrangement or interactions between scale format and scale arrangement in Experiments 2 and 3⁵⁹.

⁵⁷ Findings of two-way ANOVAs concerning the respondents' evaluation of overall survey perception for the main effects of scale format: $F(4, 676) = 1.51$, *ns* in Experiment 1.1; $F(4, 4,611) = 2.08$, *ns* in Experiment 1.2; $F(4, 5,192) = 0.49$, *ns* in Experiment 1.3; $F(2, 4,938) = 2.61$, *ns* in Experiment 3.1; $F(2, 5,056) = 1.73$, *ns* in Experiment 3.2. In Experiment 2, the effect of scale format was significant ($F(4, 688) = 2.40$, $p < .05$, $\eta^2 = .014$) due to a by tendency worse rating of the one-horizontal format.

⁵⁸ Findings of two-way ANOVAs concerning the respondents' evaluation of overall survey perception for the main effects of scale length: $F(1, 676) = 7.39$, $p < .01$, $\eta^2 = .011$ in Experiment 1.1; $F(2, 4,611) = 4.09$, $p < .05$, $\eta^2 = .002$ in Experiment 1.2; $F(2, 5,192) = 4.34$, *ns* in Experiment 1.3.

⁵⁹ Findings of two-way ANOVAs concerning the respondents' evaluation of overall survey perception for the main effects of scale arrangement ($F(1, 688) = 1.04$, *ns*) and an interaction between scale format and scale arrangement ($F(4, 688) = 3.28$, $p < .05$, $\eta^2 = .019$) due to a significantly worse evaluation of the one-horizontal format compared to all the other formats when using a positive-to-negative scale arrangement in Experiment 2; scale arrangement ($F(1, 4,942) = 0.00$, *ns*) and an interaction ($F(2, 4,942) = 3.14$, $p < .05$, $\eta^2 = .001$) due to a by tendency worse evaluation of the grid format compared to the drag-response and drag-item format when using a negative-to-positive scale arrangement in Experiment 3.1; scale arrangement ($F(1, 5,059) = 0.92$, *ns*) and an interaction ($F(2, 4,942) = 0.26$, *ns*) in Experiment 3.2.

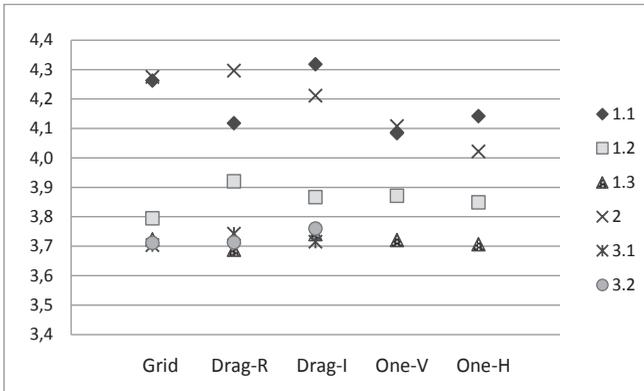


Figure 17: Respondent evaluation regarding overall survey perception depending on scale format, separately for the six experiments. Respondent evaluation was based on a 6-point semantic differential scale with higher values indicating a more positive evaluation.

In sum, the respondent evaluation of the surveys along the three dimensions of navigation, design, and overall survey perception were largely consistent. Whereas the drag-response format was deemed less easy to handle particularly with longer rating scales, the drag-item format did not differ from the grid or the single-item-per-screen formats with respect to navigation capability and ease of use. In all six experiments, both drag-and-drop formats were evaluated more positively related to the survey layout in terms of its diversion and innovation, whereas evaluations concerning a respondent's overall survey perception were unaffected by scale format. Longer rating scales consistently provoked worse evaluations with respect to all three dimensions of navigation, design, and overall survey perception, while scale arrangement largely had no effect on the respondents' evaluation of the surveys.

10. SUMMARY AND CONCLUSIONS

The main objective of this study was to determine whether newly designed drag-and-drop rating procedures can be used in Web surveys as adequate alternatives for conventional grid questions, and moreover, whether they can provide improved data accuracy by promoting a respondent's more attentive and careful processing of rating scales.

Despite the widespread use of grid questions as the presumably most common rating scale format in Web surveys, a general concern which is frequently expressed with regard to data accuracy is a respondent's susceptibility to cognitive shortcuts and the risk of item missing data. Grid questions often evoke systematic response tendencies as respondents rush through a set of several items quickly trying to minimize the extent of effort necessary to answer the rating scale items. This satisfying rather than optimal response behavior can be explained by the increased risk of respondent fatigue in rating scales in general, which results from the monotony of repetitive rating scale items referring to the same content and response options (Alwin & Krosnick, 1985; Gräf, 2002). Moreover, grid questions may carry an increased risk of respondent frustration in consequence of the considerable amount of information presented simultaneously and the necessity to link information from rows with information from columns (Callegaro, Yang, et al., 2009; Couper, et al., 2013; Gräf, 2002; Kaczmirek, 2010). Both aspects of conventional grid questions may be at the expense of the respondents' attentiveness and carefulness towards item idiosyncrasies, resulting, among others, in reduced sensitivity towards item meanings and less differentiated answers compared to using grids with fewer items or single-item-per-screen formats (Callegaro, et al., 2005; Stieger & Reips, 2010; Tourangeau, et al., 2004; Tourangeau, et al., 2009). Furthermore, an increased risk of item nonresponse and survey breakoff in grid questions is commonly explained by the higher respondent burden arising from either the difficulties in processing or the mere visual complexity of a grid question (Beatty & Herrmann, 2002; Couper, et al., 2013; Peytchev, 2006).

In general, high accuracy of survey data requires the respondents' ability and motivation to invest sufficient time and effort in attentive and careful processing of survey questions. In self-administered Web surveys

where no interviewer is present to assist and encourage the respondents to thoroughly complete all survey questions, respondent motivation is considered an even more important determinant of data accuracy than in interviewer-administered surveys (Cannell, et al., 1981; Heerwegh, 2009; Schwarz, Strack, Hippler, et al., 1991). For this reason, even greater importance needs to be attached to the design of a Web survey and the questionnaire applied (Couper, 2000; Ganassali, 2008).

Against this backdrop, two different rating scale procedures using drag-and-drop technique have been designed and tested in the present study with respect to various direct as well as more indirect indicators of data accuracy. The dynamic rating scale designs introduced here asked respondents to drag either the response options towards the rating scale items ('drag-response'), or conversely, the rating scale items towards the response options ('drag-item'). Thus, both rating scale designs make use of a more dynamic drag-and-drop technique as input method, whereas they differ regarding the question component which is draggable. By using these two drag-and-drop rating procedures, respondents might be encouraged to spend the time necessary for attentive and careful processing of rating scales. In this respect, systematic response tendencies may be prevented by (a) counteracting respondent fatigue, (b) arousing attention and decelerating the speed of responding, and (c) strengthening the link between the items and the response options. The use of these two drag-and-drop scales and their effectiveness in promoting attentive and careful processing of rating scales, thus enabling high data accuracy, was assessed on the basis of various kinds of systematic response tendencies typically encountered with regard to the use of grid questions. Furthermore, the occurrence of item missing data in terms of item nonresponse and survey breakoff was assessed. The examination of response times and respondent evaluation provided further insights into the extent of perceived and actual respondent burden and potential difficulties arising with the use of the drag-and-drop scales, which in turn may be closely linked to a respondent's motivation for attentive and careful processing of a rating scale.

The main focus of the present study was on the effects of variations in the format of a rating scale on data accuracy. The drag-response and drag-item scales were compared to a grid question as the prevalent format for rating scales in Web surveys. Moreover, two single-item-per-screen formats with either vertically or horizontally arranged response options were included in this investigation. To outline the most important results of this study, the

account of the main findings distinguishes between the two drag-and-drop scales on the one hand and conventional radio button scales on the other hand, with the former referring to the drag-response and drag-item scale, whereas the latter category comprises the grid and both single-item-per-screen formats. A more detailed differentiation of the rating scale formats is made, if notable differences were found with regard to specific formats within each group. Minor exceptions of the general findings reported here are not explicitly mentioned but reference is made to the respective section of Chapter 9. Furthermore, the results regarding variations in scale length and scale arrangement have also been discussed in greater detail in Chapter 9. Here, these findings are only reported, if striking differences were found as a function of scale length and scale arrangement.

In a first step, the main findings of the six experiments conducted in this study are presented in regards to systematic response tendencies, item missing data, response times, and respondent evaluation. These findings are subject to an initial evaluation. In a second step, the findings are combined and finally, placed in the broader context of the interactive design and administration of survey questions in Web surveys. This study concludes with an outlook on future research.

10.1 Main Findings and Implications

In multidimensional rating scale measures, the replication of their underlying theoretical structure is considered a prerequisite for a meaningful interpretation of substantive rating scale data and a first indication of the basic suitability of a rating scale design. The findings on scale properties of the rating scales assessed (see also section 9.3) consistently supported the basic suitability of both drag-and-drop scales as an alternative to well-established rating scale designs. Both the drag-response and drag-item scale were equally able to replicate the underlying factor structure of multidimensional rating scale measures compared to a grid question and single-item-per-screen designs. Furthermore, a satisfactory internal consistency of the rating scale measures was yielded as well as largely comparable substantive responses to the rating scale items irrespective of their visual presentation (except for Experiment 3, see further below on extreme responding). Thus, the drag-response and drag-item scales can be considered equally appropriate to

measure the underlying theoretical construct of a rating scale measure compared to conventional radio button scales.

Even though the basic suitability of both drag-and-drop scales is deemed proven, a closer inspection of various systematic response tendencies revealed a more differentiated view of data accuracy. In part, considerable differences have been observed between both drag-and-drop scales and conventional radio button scales on the one hand as well as between the drag-response and drag-item scale on the other hand. Although each of these systematic response tendencies may compromise data accuracy of rating scale measures in consequence of a respondent's inattention towards the rating scale content and carelessness towards the response task, underlying mechanisms differ with respect to the four stages of the question-answer process: Careless responding is ascribed to the first step of understanding and interpreting a survey question. Nondifferentiated responding and a respondent's susceptibility to primacy effects are attributable to the second stage of information retrieval and the third stage of judgment. Acquiescent responding and extreme responding are related to the fourth stage of editing and formatting a response. In the following, the main findings of this study on systematic response tendencies are discussed one by one with regard to their specific consequences for the respective stage(s) of the question-answer process.

Careless Responding. Careless responding, understood as a respondent's inattentiveness towards the reverse wording of rating scale items, is attributable to cognitive shortcuts within the first stage of the question-answer process. Respondents rely on the visual proximity of rating scale items to infer the meaning of the current item from surrounding ones, instead of attentively and carefully processing the meaning of each single item or reading an item at all (Weijters, et al., 2013).

Findings on careless responding (see also section 9.4) revealed hardly any significant differences in item-pair correlations between the different rating scale formats, with the few significant differences lacking consistent patterns. Thus, the respondents' attentiveness towards item content is largely unaffected by the respective rating scale format. Relating to both drag-and-drop scales, this suggests that drawing attention to single items and the respective response options brings no notable improvement regarding the

amount of effort respondents spent on comprehending and interpreting a set of rating scale items.

In general, comparable attentiveness towards the reverse wording of items irrespective of the rating scale format does not necessarily mean a sufficiently high extent of respondent attentiveness. However, taking into account the fact that careless responding basically has adverse effects on the internal consistency and factor structures of rating scale measures, the replication of the underlying factor structures with satisfactory percentages of variance explained by the principal components in each of the rating scale formats can be regarded as proof of a sufficiently high level of attentiveness to the reverse wording of items in the present experiments.

Nondifferentiated Responding. Nondifferentiated responding, described as a respondent's inability or unwillingness to sufficiently differentiate between several rating scale items, is mainly attributable to cognitive shortcuts within the second and third stage of the question-answer process. Instead of reassessing and determining their answers for each additional rating scale item again each time, respondents are likely to answer the first item by selecting a response option that seems reasonable and simply adjust subsequent answers to this response option (Krosnick, 1999). Besides this attempt to reduce cognitive effort, respondents may also be tempted to reduce navigational effort in Web surveys by selecting one of the response options easiest to reach with the mouse cursor (Gräf, 2002).

The findings on nondifferentiated responding (see also section 9.5) consistently showed that the drag-item format yielded a higher level of scale differentiation compared to the grid format and, with very few exceptions, also compared to the drag-response format and both single-item-per-screen formats. By contrast, the drag-response format failed to yield a higher scale differentiation than conventional radio button scales. Apparently, bringing special focus on the response options in the drag-response format does not seem to sufficiently encourage the respondents to differentiate more among the rating scale items. By contrast, in the drag-item format—by (a) highlighting the rating scale items and bringing the respective item content into special focus, and by (b) strengthening the link between the respective item and the response options—the respondents are encouraged to reconsider the response options with direct reference to the current item. Consequently, respondents assign each item to this response option that best describes their

judgment for the respective item. Hence, respondents may be more encouraged to make fine distinctions between several rating scale items instead of simply adjusting subsequent answers to previous ones. The potential increases in the extent of navigational effort that comes along with the drag-and-drop scales does not seem to prevent respondents from making use of the full range of response options in the drag-item format.

Acquiescent Responding. Acquiescent responding in terms of a respondent's tendency to agree rather than disagree with rating scale items irrespective of their content is mainly attributable to the fourth stage of the question-answer process. Respondents strive to give the impression of being friendly and polite, and thus, agree rather than disagree with a statement (Krosnick, 1999, p. 554; Podsakoff, et al., 2003).

Aside from very few significant differences, acquiescent responding (see also section 9.6) was unaffected by the format of a rating scale. This finding is in line with the general assumption that acquiescent responding is mainly related to characteristics of the respondent or to the item content rather than its visual presentation (Paulhus, 1991). In the present experiments, the actual values on the acquiescence index indicated highly balanced responses to the rating scales without any systematic tendency to agree or disagree with the rating scale items. Such balanced responses to counterbalanced rating scales also imply a high degree of consistent responding to the item pairs of original and reverse content included in the rating scale. Thus, the present findings on acquiescent responding suggest that irrespective of the rating scale format, the respondents' answers to a set of rating scale items show a high level of response consistency to reverse item pairs at scale level. This result on acquiescent responding supports the present findings on careless responding reported above which indicate a comparable level of response consistency at item level across all the experimental conditions.

Extreme Responding. Extreme responding in terms of a respondent's tendency to select the most extreme response options irrespective of the item content is attributable to the fourth stage of the question-answer process. Once a respondent has reached a judgment, this mental representation needs to be matched to the predefined response options. Hence, a respondent's tendency to use merely the endpoints of a rating scale is primarily due to individual differences in the interpretation of response options and the process of

translating a judgment into one of these response options (Paulhus, 1991; Podsakoff, et al., 2003; Shulman, 1973).

The findings in the present experiments (see also section 9.7) consistently revealed a higher tendency to extreme responding in the drag-item format compared to the drag-response or the radio button scales. With a few exceptions, no differences were found between the remaining scale formats. It is generally assumed that extreme responding has adverse effects on data accuracy. However, this general assumption has to be considered in a more differentiated manner because it is to be expected that a certain tendency to extreme responding has no negative effects on data accuracy, provided that a certain limiting value is not exceeded. Accordingly, a higher level of extremity in the drag-item format needs to be carefully examined. Since extremity values in the drag-item format are still deemed moderate, higher endorsement of extreme response options in the drag-item format is likely to reflect more differentiated responding rather than more extreme responding. In this respect, it is conceivable that the use of a drag-item format encourages those respondents, who would otherwise be reluctant to make an extreme statement, to actually select an extreme response option, if this extreme value best represents their judgment.

Hence, a certain increase in extreme responding may contribute to a higher scale differentiation in the drag-item format rather than affecting data accuracy negatively. This may also explain potential deviations in item means in the drag-item format (Experiment 3, see also section 9.3.2): Although the respondents' answers pointed in the same direction, respondents held a more extreme view in the drag-item scale than in the other scale formats, instead of falling back on moderate response options. In conclusion, this finding is considered to be in line with the general assumption that extreme responding is mainly related to characteristics of the respondent or characteristics of the item content rather than its visual presentation (Paulhus, 1991).

Primacy Effects. Primacy effects in terms of a respondent's systematic preference for one of the response options listed first is attributable to cognitive shortcuts within the second and third stage of the question-answer process. The respondents refrain from carefully assessing the appropriateness of every response option before they make their choice (Galesic, et al., 2008; Krosnick & Presser, 2010). According to an alternative explanation related to Web surveys, the respondents consider themselves satisfied with the response

option that is easiest to reach with the mouse cursor, in order to reduce their navigational effort (Gräf, 2002).

The findings on primacy effects (see also section 9.8) revealed clear differences in a respondent's susceptibility to give preference to the response options listed first depending on the respective rating scale format. The drag-item format was most susceptible because almost all the items of a rating scale were affected by primacy effects. By implication, the drag-item format was even more susceptible to primacy effects compared to a grid format. By contrast, (almost) no primacy effects were found in the drag-response format.

With reference to the layout of both drag-and-drop scales, this suggests that highlighting the response options in the drag-response format and repeatedly drawing the respondents' attention to them each time the respondents are in the act of selecting a response option for the next item, effectively helps prevent respondents from selecting the very first response option that seems reasonable to them. Thus, in the drag-response format, respondents are encouraged to constantly assess the appropriateness of each of the response options, instead of being satisfied with the first available one. In this respect, the extent of navigational effort increasing with the selection of response options that are more distant seems to be rather irrelevant in the drag-response scale.

By applying the same logic to the drag-item format, it can be argued that drawing special attention to the item content was practically ineffective, if not even obstructive, in preventing the occurrence of primacy effects. One possible explanation for this higher susceptibility to primacy effects in the drag-item format might be that respondents are more reluctant to make the effort to select one of the more distant response options. Thus, in order to reduce their navigational effort, respondents are likely to select the response option that is easiest to reach with the mouse cursor. However, this interpretation is contrary to previous findings which revealed higher scale differentiation in the drag-item format irrespective of the navigational effort associated.

Item Nonresponse. In the current experiments, item nonresponse refers to a respondent's failure to provide an answer at all to one or more rating scale items (Beatty & Herrmann, 2002; De Leeuw, et al., 2003). Respondents could proceed to the next Web page without being prompted to provide an answer. Furthermore, the rating scales did not include an explicit nonsubstantive

response option (e.g., ‘don’t know’ option). Thus, item nonresponse could be the result of respondents intentionally skipping one or more items because of their insufficient ability or motivation to complete the items, or the result of skipping one or more items by mistake.

The findings on item nonresponse (see also section 9.1) consistently showed that both drag-and-drop scales suffered from a considerably higher risk of item nonresponse in terms of increased proportions of partially and completely missing values compared to conventional radio button scales. High proportions of partially missing values were particularly likely in the drag-response format, whereby this proportion further increased in longer rating scales. Item nonresponse in the drag-item scale and conventional radio button scales was largely unaffected by the length of a rating scale.

These findings give clear indication of an increase in both perceived and actual respondent burden, resulting in higher proportions of completely and partially missing values in both drag-and-drop scales. In other words, respondents decide to completely skip a rating scale when it is presented in a drag-and-drop format because of the high anticipated effort required to answer the rating scale. Apparently, a considerable number of respondents decide not to spend the effort required for the understanding and processing of the drag-response and drag-item format solely on the basis of their visual appearance. Thus, the mere visual complexity of the drag-response and drag-item format is likely to prevent respondents from even starting the response task. However, one limitation of the present experiments is that although certain devices which were known to be incompatible with the underlying functionality of the designs tested here (like smart phones and tablet PCs) were excluded right at the outset of the experiments, it cannot be completely ruled out here that some respondents actually did not meet the technical requirements to use the JavaScript-based drag-and-drop scales.

In addition, even if the initial respondent decision is in favor of answering a drag-and-drop scale, the extent of the actual effort required for handling and potential difficulties in navigation are likely to exceed the individual threshold of effort respondents are willing to spend when answering a rating scale. Consequently, the likelihood of skipping single items is increased in both drag-and-drop scales but especially in the drag-response scale. Conversely, this means that once a respondent is mainly willing to answer a drag-and-drop scale, he or she is more likely to complete the drag-item scale rather than the drag-response scale. The likelihood that the

respondents accidentally skipped one of the items is considered rather low since—as a result of the visual highlighting in the drag-response and drag-item scale—it can clearly be detected which of the items were not answered yet. And, although an explicit nonsubstantive response option (e.g., ‘don’t know’ option) was not provided, it was nevertheless expected that every respondent actually had an opinion on the issues in question.

Given these large differences in the risk of item missing data, it is important to note that both drag-and-drop scales induce more item nonresponse, but without systematically affecting the sample composition. Thus, item-nonrespondents in the drag-response and drag-item scale did not differ from item-nonrespondents in the radio button scales with respect to respondent characteristics that are considered relevant in this context (esp., computer and Internet literacy, prior survey experience).

Survey Breakoff. Survey breakoff occurs if the substantive responses to all survey questions are missing after a certain point in time as a result of the respondent’s premature termination of the survey. Accordingly, survey breakoff is also considered a more aggravated form of item nonresponse (De Leeuw, et al., 2003).

The findings of the present experiments provided largely consistent evidence that survey breakoff was unaffected by different rating scale formats (see also section 9.2). Thus, the drag-response and drag-item format in principal did not suffer from an increased risk of survey breakoff in the present experiments. However, in view of the considerably increased item nonresponse rates in both drag-and-drop scales, the question arises what would have happened, if respondents had not been allowed to simply skip single rating scale items or even the entire drag-and-drop scale. Those respondents who were not able to complete the drag-and-drop scales would have certainly quit the survey. Of particular interest would have been the respondents who were initially unwilling to spend the effort required for the completion of a response task, but nevertheless provided an answer after being prompted to do so. Concerning the latter, the question arises whether these respondents would have provided substantive responses of appropriate accuracy or whether this initial reluctance of the respondents would have resulted in impaired measurement properties. In either case, it is important to carefully examine the interaction between item nonresponse and survey breakoff on the one hand, and the relationship between the use of forced

prompts and the accuracy of survey responses to the drag-and-drop scales on the other hand.

Response Times. In general, response times are often examined to gain a better understanding of how respondents process a survey question. Commonly, there are two possible interpretations. First, response times may reflect a respondent's difficulties in cognitive processing of a survey question and second, response times allow conclusions to be drawn about a respondent's effort spent on processing a survey question (Callegaro, et al., 2005; Draisma & Dijkstra, 2004; Heerwegh, 2003; Husser & Fernandez, 2013; Stern, 2008; Stieger & Reips, 2010).

The findings on response times (see also section 9.10) clearly and consistently indicated that it took considerably longer to complete the rating scale items that were presented in a drag-response or drag-item scale than in a conventional grid question, and with only a few exceptions, also more time than both single-item-per-screen formats. This seems hardly surprising because navigation and handling in a drag-and-drop scale per se is more time-consuming than simply clicking a radio button in conventional rating scale formats.

In addition, respondents spent more time on the initial orientation on a screen when answering one of the drag-and-drop scales than for a conventional radio button scale. This clearly indicates a respondent's increased time requirements when being confronted with a new kind of rating scale format, i.e., it takes longer to gain a first overview of the drag-and-drop scales and to comprehend their overall structure and navigation. These additional time requirements were even higher in the drag-item format. Thus, since the drag-item scale deviates to a greater extent from the respondents' general expectations derived from the conventional design and construction of a rating scale, the drag-item format is likely to be, at first glance, less intuitive than the drag-response scale.

Leaving the time for initial orientation out of the account, differences in response times between the drag-and-drop scales on the one hand and the conventional radio button scales on the other hand still existed, however, a convergence in response times occurred. Thus, the disproportionate disadvantages in time towards conventional radio button scales need to be considered in a more differentiated manner given the novelty of the drag-and-drop formats and a respondent's initial unfamiliarity with these new kinds of

rating scale formats. This finding suggests that if respondents answer a drag-and-drop scale a second, third, or fourth time, the additional time required for the initial orientation and basic understanding will decrease, and time differences between the drag-and-drop and radio button scales will further be reduced.

In addition, time savings in longer rating scales were found for conventional radio button scales. These time savings are generally assumed to reflect facilitation of the cognitive processing of rating scale items due to an increasingly enriched question context as the number of answered items increases. Thus, contrary to common assumptions, it can be assumed that the question context necessary to gain a complete understanding of the underlying construct of a rating scale was also provided in single-item-per-screen formats. Furthermore, there was no evidence of a respondent's increased susceptibility to speeding and click-through behavior due to fatigue effects in longer grid questions compared to single-item-per-screen formats.

Decreases in response times with increasing scale length were also found in the drag-item scale, whereas evidence of time savings in the drag-response scale was rather mixed. Nevertheless, the question context encouraging a complete understanding and a continuous flow in the cognitive processing of rating scale items also seemed to be provided in the drag-and-drop scales (see also section 9.9 on semantic-order effects). Despite certain time savings in both drag-and-drop scales, these efficiency gains appeared to apply only up to a certain number of items. The stagnation in time savings is likely to be due to increases in the sensorimotor effort for navigation and—particularly in smaller sized browser windows—due to the associated necessity of scrolling in longer drag-and-drop scales. These findings indicate that time savings are likely to occur in both drag-and-drop scales as a result of the facilitation of the cognitive processing as well as due to the learning effects and improvements of the navigational processing of rating scale items. Both may help increase the speed of responding, at least up to a certain number of items in a drag-and-drop scale.

Further examination of dragging times (see also section 9.10.4) consistently revealed that the navigation in the drag-item format was more time-consuming than in the drag-response format. However, it was also found that dragging times in the drag-response scale were clearly affected by the number of rating scale items, whereas dragging times in the drag-item scale were unaffected by the scale length. This finding provides interesting insights

into the cognitive and navigational processing of the drag-and-drop scales. In the drag-response scale, efficiency gains were found in terms of respondents getting more and more familiar with the navigational requirements of the drag-response scale, resulting in decreased average dragging times per item, as more items have already been answered. However, dragging times increased again, if the number of items exceeded a certain limit as the result of increased time for navigation in longer drag-response scales. This finding suggests that dragging times in the drag-response format primarily reflect the time spent on navigational processing. By contrast, dragging times in the drag-item format were unaffected by the number of items in a scale. This in turn suggests that dragging times in the drag-item format involve more than just the time needed for mere navigation. It rather seems as if respondents use the time spent on dragging an item to its desired position—to a sizeable extent—for the cognitive processing of the rating scale content.

Respondent Evaluation. The respondent evaluation of key survey aspects may reflect the extent of the respondents' actual and perceived burden, and thus, their motivation to spend the effort required for complete and accurate responses (Ganassali, 2008; Lozar Manfreda, et al., 2002).

The findings on respondent evaluation (see also section 9.11) clearly showed that navigation-related aspects of a survey in terms of navigation capability and ease of use were evaluated less positively when respondents answered a drag-response scale. This evaluation got even worse with an increasing length of a drag-response scale. By contrast, the respondents' evaluation of navigation-related aspects of a survey remained unaffected by the use of a drag-item scale. At the same time, a survey was rated more positively regarding its design-related issues in terms of diversity and innovation when the respondents were assigned one of the drag-and-drop scales. No differences were found concerning the respondents' overall survey perception. Additionally, with an increase in the length of a rating scale, a survey was consistently evaluated worse concerning all three dimensions: navigation, design, and overall survey perception.

Obviously, the respondents consider the drag-response scale less easy to handle than the drag-item or conventional radio button scales. This is particularly the case when the drag-response format is used for longer rating scales. By contrast, the drag-item format does not differ from a conventional radio button scale with respect to its ease of use. Rating scales using the drag-

and-drop technique are generally deemed a more challenging task than simply clicking a radio button in a conventional rating scale (Couper, et al., 2006; Funke, et al., 2011). Thus, while it is hardly surprising that the drag-response scale is considered less easy to use compared to a radio button scale, it is all the more surprising that the drag-item scale does not differ from radio button scales with respect to its ease of use. In principle, as far as the layout and structure of a drag-response and drag-item scale is concerned, one might assume that the drag-item scale is even more challenging and burdensome than the drag-response format, especially due to the fact that it gradually increases in size with each additional item being answered. However, this is obviously not the case from a respondent's point of view. According to the satisficing theory, a high level of task difficulty and the associated increased risk of the respondents' actual and perceived burden may also carry an increased risk of respondent frustration, which in turn may interfere with the respondents' ability and motivation to provide complete and accurate responses. Following this argumentation, the drag-response scale may be at a higher risk of resulting in respondent frustration compared to the drag-item scale and conventional radio button scales, especially in case of exceeding a certain number of rating scale items.

Relating to design diversion and innovation, the drag-response and drag-item format were assessed more positively compared to conventional radio button scales. Thus, both drag-and-drop scales are considered more diversified, innovative, and inventive compared to the conventional radio button scales. This finding is in line with previous findings on various kinds of drag-and-drop scales (Downes-Le Guin, et al., 2012; Sikkel, et al., 2014; Sleep & Puleston, 2009; Stanley & Jenkins, 2007). As opposed to repetitively and monotonously asking the same type of survey questions, new question formats are likely to be evaluated more positively because they may offer a new experience and provide variety to the surveys. In this case, new question formats may promote higher respondent involvement and counteract respondent fatigue in survey responding. As a result, respondents are more likely to engage in attentive and careful processing of survey questions, which in turn may promote more complete and accurate responses. However, the question remains whether there is a kind of 'novelty effect', according to which the positive effects related to increased respondent involvement in drag-and-drop scales would be likely to wear out rather quickly. This would imply that the drag-response and drag-item scales should only be used rarely.

Another issue arising in view of the previous findings on the respondents' susceptibility to systematic response tendencies and item nonresponse, is that a respondent's positive evaluation of design-related aspects of a novel rating scale format seems to be no sufficient indicator of how much burden a respondent perceives during survey completion.

In view of these direct and more indirect indicators of data accuracy in rating scales, the present findings are rather mixed and in part contradictory, in regards to the effectiveness of the drag-and-drop scales in promoting attentive and careful processing of rating scales, thus enabling high data accuracy. Hence, it is even more important to adopt a holistic view of the present findings in order to get a clearer understanding of the effect of both drag-and-drop scales on complete and accurate responses in rating scales, rather than to consider each indicator separately.

10.2 Overall Assessment

Generally, the respondents' susceptibility to cognitive shortcuts in terms of systematic response tendencies is primarily explained by levels of respondent motivation and task difficulty. Lowered motivation may tempt respondents to fall back on systematic responding in order to decrease the cognitive and navigational effort required for complete and accurate responses. Difficult response tasks are assumed to imply increased effort, which in turn may further increase the respondents' burden and encourage them to rely on some kind of cognitive shortcuts in order to reduce this effort (Krosnick, 1991). However, the present findings revealed a more differentiated view of the relationship between the task difficulty on the one hand and the respondents' motivation to provide complete and accurate responses in rating scales on the other hand.

Both drag-and-drop scales evidently imply an increased effort because of the higher complexity of the response task. This was indicated by a considerably increased incidence of item nonresponse as well as by clearly longer response times in both drag-and-drop scales. A notable number of respondents were deterred by this increased perceived and actual respondent burden, and thus, skipped the entire set of rating scale items or, at least, some of them. A detailed examination of response times also revealed that a

considerable amount of time spent on completing the rating scale was attributed to the increased navigational effort in the drag-and-drop scales. Thus, this increased perceived and actual respondent burden is likely to prevent respondents from even starting the question-answer processing. Instead, they simply skip one or more rating scale items. Consequently, the completeness of rating scale measures is compromised in both drag-and-drop scales.

Following the aforementioned assumption on the relationship between task difficulty, respondent burden and respondent motivation, the higher level of task difficulty and the increased actual and perceived respondent burden in both drag-and-drop scales are expected to decrease the respondents' motivation for attentive and careful processing, which in turn might impair measurement in both drag-and-drop scales. Related to a respondent's susceptibility to cognitive shortcuts in rating scales, the present findings revealed that the second and third stage of the question-answer process, namely information retrieval and judgment, are clearly influenced by differences in the rating scale format. However, opposed to the generally accepted assumption that increased task difficulty impairs data accuracy, it was found that—despite the increased level of task difficulty—both drag-and-drop scales offer advantages that positively affect the attentive and careful processing of a rating scale.

Nondifferentiated responding and primacy effects are two kinds of cognitive shortcuts that are typically observed in rating scales. Both have shown to be decisively affected by the drag-response and drag-item format. However, in view of the present findings, the drag-response and drag-item scales led to different outcomes. In general, the two drag-and-drop formats have different implications for both the navigational and cognitive processing. The differences in the navigational processing are obvious: In the drag-response scale, the response options are assigned to the items, whereas in the drag-item scale, the items are assigned to the response options. Both scales also differ concerning the visual highlighting of the respective question components: In the drag-response scale, the primary focus is on the response option meanings, while the opposite is the case with the drag-item scale, where special attention is drawn to the item content. These apparently trivial differences between the two drag-and-drop scales have crucial implications for the cognitive processing of rating scale items. And obviously, the drag-

response scale and drag-item scale have varying effects on respective kinds of cognitive shortcuts.

Concerning nondifferentiated responding, the present findings showed that special focus on the item content is essential for promoting higher scale differentiation. Bringing the item content into focus and strengthening the link between the respective item and possible response options encourages respondents to repeatedly reassess the appropriateness of a respective response option for a current item. Concerning the risk of primacy effects, the present findings indicated that a special focus on response options is central to the prevention of a respondent's tendency to select the very first answer that seems reasonable. Thus, bringing the response options into focus encourages respondents to assess the appropriateness of a response options for every new item being answered.

Hence, beneficial factors influencing the respondents' attentiveness and carefulness in rating scales are the visual highlighting of one of the key components of a rating scale, as well as the dynamic strengthening of the link between an item and the response options. These two characteristics of the drag-and-drop scales evidently encourage the respondents' repeated (re)evaluation of the appropriateness of a response option, with the aim of achieving the optimal matching between the item and a selected response option. In both cases, the increased navigational effort associated with the selection of the response options that are possibly not nearest, and thus, requiring more effort to be reached with the mouse cursor, does not impede respondents from optimal responding. Both higher scale differentiation and less primacy effects endure in moderate or even long rating scale, which means that there is no wear-out effect or fatigue effect over the course of rating scale completion.

However, there are also restrictions involved in this explanation, in view of the fact that the drag-item scale yielded highest scale differentiation but was also most susceptible to primacy effects. Obviously, besides the visual highlighting and strengthened linking, there seems to be another mechanism accounting for higher scale differentiation and, simultaneously, higher primacy effects in the drag-item scale. These—at first glance—contradictory results might be explained by the resemblance between a drag-item scale and a rank-order scale. Accordingly, the structure of a drag-item scale seems to implicate a kind of 'visual rank order' which may promote higher differentiation among the rating scale items. Thus, respondents are

encouraged to make use of the full range of response options and therefore, produce less nondifferentiation. However, the higher level of navigational burden for reaching more distant response options in the drag-item scale remains. Therefore, respondents are still tempted—at least to some extent—to prefer one of the nearer response options; hence, primacy effects can be observed.

The findings on nondifferentiated responding and primacy effects showed that the second and third stage of the question-answer process can be decisively affected by variations in the rating scale format. Nevertheless, further research is still needed to better understand how differences in the respondents' focus of attention and in the navigational processing sequence of items and response options induce differences in the cognitive processing of a rating scale. Additionally, the present results also showed that the first stage of question comprehension and the fourth stage of response formatting and editing are largely unaffected by the format of a rating scale. Presumably, the systematic response tendencies arising from these two stages are primarily affected by characteristics related to the respondent or to the content of a rating scale, rather than to its visual presentation. For instance, the personal importance of the question topic may affect the respondents' carefulness in processing the item content as well as their susceptibility to acquiescent or extreme responding (Chang & Krosnick, 2009; Herzog & Bachman, 1981).

Furthermore, the occurrence of systematic response tendencies and item missing data were largely unaffected by the length of a rating scale, whereas an effect of scale length on the time required to complete a set of rating scale items has been noted. In conclusion, in order to keep the respondents' effort within manageable limits, the number of rating scale items in both drag-and-drop scales must not exceed a certain limit. Thus, despite the functionality of auto-scrolling, up to 10 rating scale items may be considered to be the maximum length of both drag-and-drop scales.

Another important finding was related to the visual grouping of rating scale items in terms of presenting them either in a grid—as a kind of multiple-item-per-screen design—or in a single-item-per-screen design, which mostly yielded no differences in the indicators of data accuracy examined in the present study. Thus, with a few exceptions, there is no evidence that because of the visual proximity of items in grid questions, respondents would be more susceptible to cognitive shortcuts than when the items are presented separately.

In consideration of the present findings, it can be concluded that the drag-response and drag-item scales measure the same theoretical construct as conventional radio button scales. Also, the question context necessary for a comprehensive understanding of the underlying theoretical construct and a fluent cognitive processing of a set of rating scale items is provided in both drag-and-drop scales. Thus, even though a higher respondent focus on the single components of a rating scale is promoted in both drag-and-drop scales and particularly in the drag-item scale, respondents are not at risk of losing sight of the 'big picture' in a rating scale. Furthermore, contrary to conventional radio button scales, visual highlighting of the item content and response option meanings as well as dynamic strengthening of the link between the items and response options in the drag-and-drop scales both positively affect the respondent's attentiveness and carefulness in processing a set of rating scale items. Hence, visual enhancement and greater interactivity encourage the respondents' attention to the key components of a rating scale and counteract the repetitive nature of responding in traditional rating scales. Conversely, these findings also suggest that the respondents' attention needs to be drawn to the key components of a survey question directly and repeatedly in order to prevent them from falling back on cognitive shortcuts. Moreover, the use of a more challenging data input method in rating scales may prevent the risk of respondent fatigue in an otherwise rather monotonous and uniform response task. On the contrary, however, the increased extent of respondent burden perceived already before, or while processing the drag-and-drop scales, carries an increased risk of skipping the entire rating scale or omitting some of the items. Thus, the present findings emphasize the necessity of determining a proper level of task difficulty which seems to be a balancing act between the respondents' fatigue appearing if the level of cognitive load is low and probably too low to ensure respondent involvement, and on the other side, the respondents' frustration occurring when the level of cognitive load is high and then again, probably too high to accomplish the response task within reasonable time and with adequate effort.

10.3 General Discussion

In survey research, there is an ongoing discussion about the proper complexity of the response task to promote, rather than prevent, the respondents' thorough processing of survey questions. On the one hand, making the response task easier for the respondents may compromise data accuracy by causing fatigue and a decline in respondent motivation to attentively and carefully process all the relevant components of a survey question. On the other hand, high cognitive and navigational demands may result in frustration and distraction from the actual response task, which in turn may find its expression in the respondents' reduced motivation to spend sufficient effort to answer survey questions completely and accurately. In this regard, the possible applications of visual and dynamic questionnaire features are often discussed since a higher visual and navigational complexity may distract respondents from the actual response task, rather than encouraging them to engage in attentive and careful responding, which in turn may hamper complete and accurate responses.

Against the backdrop of the present study, respondents are likely to be discouraged from even starting the response task because of the mere visual complexity of a survey question and the high perceived burden associated with the response task. Furthermore, even if respondents are willing to invest this increased effort and start processing a survey question, they are more likely to prematurely abandon the processing because of the high actual burden accompanied with a survey question. Thus, in order to receive complete responses—as a prerequisite for accurate answers—the perceived and actual respondent burden associated with a response task has to be kept within a certain limit. This will be of even greater importance, if survey researchers decide to implement forced prompts requiring respondents to unavoidably answer each single survey question before they can proceed to the next question. With respect to the use of visual and dynamic questionnaire features, this means that respondents should ideally be able to comprehend the structure and navigational requirements of a survey question at first glance without being required to spend too much effort already at this initial stage of visual perception and pre-attentive processing of a survey question.

However, simply lowering the difficulty of a response task with the aim of reducing respondent burden and promoting complete and accurate survey responses seems to oversimplify matters. Particularly in Web surveys, in view

of the absence of an interviewer but, at the same time, in the presence of a lot of distracting events causing respondents to lose interest in the response task rather quickly, full respondent attention has to be drawn to the response task. One possible way to do this is to make the response task more challenging in order to attain higher respondent involvement and evoke high respondent motivation to invest the effort required for complete and accurate responses.

Survey researchers have plenty of visual and dynamic features at their disposal to improve the design and administration of Web questionnaires and positively affect various response decisions, determining complete and accurate respondents' answers. Visual and dynamic features can be used to make the response task more diverse and engaging and, at the same time, to draw the respondents' attention to the respective response task. Related to the design and administration of rating scales as one question format that often suffers from a respondent's inattention and fatigue, visual highlighting and the use of dynamic drag-and-drop techniques can efficiently be applied to directly and repeatedly draw the respondents' attention to the key components of a rating scale. Furthermore, the repetitive and monotonous nature of a rating scale can be disrupted, thus increasing data accuracy as both aspects contribute to prevent respondents from falling back on cognitive shortcuts in the processing of rating scales.

By making use of visual enhancement and greater interactivity in the design and administration of survey questions, respondent attention can be promoted, whereas respondent fatigue may be counteracted, and routines in the respondents' survey-taking behavior disrupted. Nevertheless, the use of advanced dynamic features for constructing survey questions can also enhance respondent burden by imposing higher cognitive and navigational requirements at least on those respondents who are less capable to manage the increased requirements. Hence, it seems to be a balancing act between the extent of respondent burden and a certain degree of cognitive load that counteracts respondent fatigue, or positively worded, promotes respondent involvement and the motivation to spend the effort on attentive and careful processing of survey questions. Although respondents seem to accept or even appreciate a more challenging and engaging response task to a greater extent than previously assumed, the level of respondent burden has to be kept within manageable limits. This is of great importance to encourage, rather than frustrate or even discourage respondents from responding. This seems to be a

promising approach aiming at complete and accurate responses in Web surveys.

10.4 Limitations and Further Research

Respondent-Related Characteristics

The main focus of the present study was on aspects related to the design and administration of survey questions which can be directly influenced by the survey researcher, with the aim of positively impacting the respondents' survey-taking behavior. However, it must not be forgotten that certain respondent characteristics such as their knowledge in dealing with computers and the Internet are further decisive factors influencing the likelihood of complete and accurate survey data. This is particularly the case when it comes to the implementation of new Web survey question formats that require at least a basic understanding of the functionality of computers and the Internet. More specifically, the respondents' computer and Internet literacy is considered an influencing factor determining the ease of use and adequate handling of the drag-and-drop rating scales since their basic functionality is adopted from the drag-and-drop interactions used in many standard software programs. Although the consideration of respondent-related characteristics to an adequate extent would have been beyond the scope of the present study, it is an important next step towards assessing possible applications of drag-and-drop rating scales in Web surveys.

Also, it has to be taken into account that the surveys the present experiments have been embedded in targeted a relatively homogeneous and, at the same time, young and highly educated sample of university students and university applicants who are generally believed to be fairly versed in using computer technologies and the Internet. This holds true for the given samples that are highly homogeneous with regard to the respondent's age and education. Moreover, the samples are characterized by a high level of computer and Internet literacy since the vast majority of respondents ascribed themselves advanced or expert skills in dealing with the computer (ranging between 75.5% and 86.9%); and, an even higher percentage of about 90% of the respondents described their Internet skills at an advanced or expert level (ranging between 87.8% and 92.2%). In this regard, computer and Internet use can be considered a part of the daily routine for many of the university

students and university applicants in the samples of this study. Given the fact that the present studies addressed a special target population, it is necessary to replicate the present findings based on a sample of a more general population, featuring a greater level of heterogeneity concerning key socio-demographic characteristics such as age and education (as proxy measures for the level of cognitive ability and cognitive sophistication) as well as further key characteristics determining a respondent's ability and motivation to provide complete and accurate responses in Web surveys, such as the level of computer and Internet literacy and prior Web survey experience.

In general, a respondent's prior survey experience is considered another relevant factor influencing the accuracy of survey data. The extent of prior survey experience may affect both the respondent's ability and motivation and may have a twofold effect on the respondent's survey-taking behavior. For experienced respondents who are more familiar with and experienced in the question-answer process, it may be easier to process survey questions and provide complete and accurate answers, compared to rather inexperienced respondents. On the contrary, however, those respondents who are versed in answering survey questions may also be less motivated to provide complete and accurate answers since they are presumably also more experienced in reducing the effort required to answer survey questions compared to respondents with little or no survey experience (Chang & Krosnick, 2009; Toepoel, et al., 2008; Yan & Tourangeau, 2008). This impact of a respondent's prior survey experience may be even more pronounced in Web surveys using visual and advanced dynamic input controls such as rating scales based on drag-and-drop techniques. And even if a respondent has a certain level of knowledge in dealing with computers and the Internet, the routines and other requirements applied in a Web survey environment may nevertheless be completely different and 'new territory' for rather inexperienced Web survey respondents.

The amount of survey experience acquired through the number of surveys in which a respondent has participated before gives some indication of a respondent's practice in survey responding (Toepoel, et al., 2008; Yan & Tourangeau, 2008). In the present study, respondents were asked to indicate the number of Web surveys they participated in within the last 12 months. The participants in the Panel Survey 2012 were, not surprisingly, most experienced with an average of 3.3 Web survey participations within the last year. Nevertheless, the difference between the panel members and the

participants of the University Applicants Survey 2012 and 2013 was smaller than anticipated, with an average of 1.8 and 2.3 Web survey participations within the last 12 months, respectively. In all three samples, there was a great variance in the respondents' prior Web survey experience. Therefore, further research on the effectiveness of the present drag-and-drop rating scales should be conducted with due regard to the differences in the respondents' Web survey experience.

Albeit differences in the average prior Web survey experience between the three samples were rather small, it is noticeable that for most of the systematic response tendencies assessed in the present study, only small, or no effects of experimental manipulations were found in Experiments 1.1 and 2 which were implemented in the Panel Survey 2012 and conducted among opt-in panel members studying at the TU Darmstadt. One possible explanation for the lack of significance in Experiments 1.1 and 2 might be the smaller sample size in these two experiments compared to the other experiments. Another reason might refer to the differences in sample composition and the underlying motivation to take part in the survey. In this context, panel conditioning is considered a further source of measurement error in surveys conducted among panel members. Panel conditioning describes the fact that, because of the respondents' increasing survey experience, "their responses may increasingly begin to differ from the responses given by people answering the same survey for the first time" (Couper, 2000, p. 476). Discussing the consequences of this assumption would go way beyond the scope of the study at hand. Nevertheless, future research needs to investigate the drag-and-drop rating scales based on a sample of less survey-affine respondents to have a stronger effect on the results.

Context-Related Characteristics of Rating Scales

A systematic examination of serial-order effects will be useful to gain deeper insights into the interplay of respondent learning and respondent fatigue. By means of the systematic variation of the position of a drag-and-drop rating scale within a questionnaire, the occurrence of fatigue effects can be assessed, potentially resulting from increases in respondent burden over the course of questionnaire completion. Generally, it is assumed that respondent burden increases with the number of survey questions already answered, whereby the respondent's motivation to spend the effort on attentive and careful

processing is likely to decrease (Galesic & Bosnjak, 2009; Peytchev, 2009). Thus, it can be examined whether the effectiveness of the drag-and-drop rating scales in preventing the respondent's susceptibility to cognitive shortcuts is decreased in consequence of the increased cumulative burden experienced by the respondent, or whether the use of drag-and-drop rating scales may even mitigate this negative effect of increasing burden over the course of questionnaire completion.

In addition, by means of dependent randomization and the assignment of respondents to a drag-and-drop rating scale several times during questionnaire completion, effects related to learning and fatigue can be examined. Learning effects may arise since respondents become more practiced and familiar with the drag-and-drop formats. These learning effects may shorten response times and decrease respondent burden (Callegaro, et al., 2005; Krosnick & Presser, 2010; Scherpenzeel & Saris, 1997). On the contrary, fatigue effects may occur again as the respondents are willing to spend the increased effort probably once or twice, but not three times. Conceivably, wear-out effects would further promote respondent fatigue since the initially positive effect of diversion and innovation of the drag-and-drop rating scales fails to appear when answering the same scale several times. This would suggest using the drag-and-drop rating scales rather sparingly.

Additional Insights from Eye Tracking Data

Eye tracking data can provide a more direct assessment of the respondents' attention to and their processing of the drag-response and drag-item scale. The recording of eye movements allows for the detection of the extent of the respondents' attention to the single components of a rating scale by recording the number and duration of eye fixations. Furthermore, the recording of eye movements helps identify the sequence in which respondents process these rating scale components (Jacob & Karn, 2003; Poole & Ball, 2005). This information would be beneficial in several respects as outlined below.

First, it can be detected more precisely to what extent respondents actually concentrate their attention on the rating scale content, and more specifically, how much time they spend on the cognitive processing of the items and response options, rather than on the mere navigation of the draggable question-elements and answer-elements. Thus, eye fixations provide some indication of 'productive' and 'non-productive' processing time in the drag-and-drop rating scales. In this respect, eye tracking data may help

clarify whether, in more challenging response tasks, additional time is used for deeper cognitive processing or whether the increased response times are mainly due to the respondents' higher navigational effort. Moreover, eye tracking data may disclose whether dragging operations in the drag-item scale actually involve more time that is spent on cognitive processing of the rating scale content than in the drag-response scale.

Second, it can be determined in which order respondents process the items and response options, and how often their eye movements switch backwards and forwards between a respective item and the response options. In this regard, the sequence of eye movements may disclose the process of assessing and reassessing the appropriateness of a response option with the aim of achieving the optimal match between an item and a response option. The present findings clearly show that the second and third stage of the question-answer process are decisively affected by the drag-and-drop rating scales, however, obviously in different ways because the drag-response format averts primacy effects, whereas the drag-item scale promotes higher scale differentiation. Thus, exploring the sequence of eye movements may help explain the differences between the drag-response and drag-item scales with regard to the cognitive processing of the rating scale contents.

Third, both the respondents' eye fixations and eye movements in the course of processing the various components of a rating scale may give some indication of the respondents' straightforward processing of the rating scale, or otherwise, their confusion about the response task. Many short and unfocused eye movements between the varying components of a rating scale might be indicative of the latter. In this regard, eye tracking data may help gain a better understanding how the complexity of a response task actually affects the respondents' cognitive processing in terms of either (a) drawing the respondents' undivided attention to the key components of a survey question and encouraging focused and careful processing, or (b) causing the respondents' distraction and impeding the cognitive processing of a survey question. This may help determine the appropriate difficulty level of a response task with the aim of designing survey questions that implicate a certain level of cognitive load which in turn may increase the respondents' attention and motivation but also keeps the extent of cognitive load within reasonable limits to still enable a straightforward completion of the survey question without overburdening the respondents.

The various aspects discussed in this section reveal multiple possibilities for future research. Even further, these aspects briefly highlighted here can be considered the next steps that need to be systematically implemented to ascertain the range of possible applications of these types of drag-and-drop rating scales. The objective of this study was not first and foremost to present a ready-made version of new drag-and-drop rating scales but rather to gain a better understanding of the respondents' cognitive processing of survey questions and to make use of this knowledge to gradually move away from the sheer adaption of static paper-based questionnaire layouts towards the use of the potential offered by visual and dynamic questionnaire features in Web surveys. Finally, the use of the drag-and-drop rating scales introduced in this study represents a first and promising attempt to take advantage of the potentials of visual enhancement and greater interactivity as unique features of Web surveys, giving evidence of the implications of questionnaire design and administration for the respondents' attentive and careful cognitive processing of survey questions and increased data accuracy in Web surveys.

REFERENCES

- AAPOR (2011). *Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys*. Deerfield, IL: AAPOR.
- ADM (2013). *Jahresbericht 2013*. Frankfurt a.M.: Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.
- Alwin, D. F. (1989). Problems in the Estimation and Interpretation of the Reliability of Survey Data. *Quality and Quantity*, 23(3-4), 277-331.
- Alwin, D. F. (1991). Research on Survey Quality. *Sociological Methods & Research*, 20(1), 3-29.
- Alwin, D. F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken, NJ: Wiley.
- Alwin, D. F. (2010). How Good Is Survey Measurement? Assessing the Reliability and Validity of Survey Measures. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (pp. 405-434). Bingley: Emerald.
- Alwin, D. F., & Krosnick, J. A. (1985). The Measurement of Values in Surveys: A Comparison of Ratings and Rankings. *The Public Opinion Quarterly*, 49(4), 535-552.
- Alwin, D. F., & Krosnick, J. A. (1991). The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes. *Sociological Methods & Research*, 20(1), 139-181.
- Andrews, F. M. (1984). Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach. *Public Opinion Quarterly*, 48(2), 409-442.
- Barge, S., & Gehlbach, H. (2012). Using the Theory of Satisficing to Evaluate the Quality of Survey Data. *Research in Higher Education*, 53(2), 182-200.
- Barnette, J. J. (2000). Effects of Stem and Likert Response Option Reversals on Survey Internal Consistency: If You Feel the Need, There is a Better Alternative to Using those Negatively Worded Stems. *Educational and Psychological Measurement*, 60(3), 361-370.
- 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(3), 331-346.
- Bassili, J. N., & Scott, B. S. (1996). Response Latency as a Signal to Question Problems in Survey Research. *Public Opinion Quarterly*, 60(3), 390-399.
- Bauman, S. L., Jobity, N., Airey, J., & Atak, H. (2000). *Invites, Intros and Incentives: Lessons from a Web Survey*. Paper presented at the 55th Annual Conference of the American Association for Public Opinion Research (AAPOR), May 18-21, 2000, Miami Beach, FL.
- Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research*, 38(2), 143-156.
- Baumgartner, H., & Steenkamp, J.-B. E. M. (2006). Response Biases in Marketing Research. In R. Grover & M. Vriens (Eds.), *The Handbook of Marketing Research. Uses, Misuses, and Future Advances* (pp. 95-109). Thousand Oaks: Sage.
- Beatty, P., & Herrmann, D. (1995). A Framework for Evaluating Don't Know Responses in Surveys. *JSM Proceedings of the Survey Research Methods Section. Alexandria, VA: American Statistical Association*, 1005-1010.
- Beatty, P., & Herrmann, D. (2002). To Answer or Not to Answer: Decision Processes Related to Survey Item Nonresponse. In R. M. Groves, D. A. Dillmann, J. L.

- Eltinge & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 71-86). New York: Wiley.
- Belson, W. A. (1981). *The Design and Understanding of Survey Questions*. Aldershot: Gower.
- Bergkvist, L., & Rossiter, J. R. (2007). The Predictive Validity of Multiple-Item Versus Single-Item Measures of the Same Constructs. *Journal of Marketing Research*, *44*(2), 175-184.
- Bescherer, C., & Spannagel, C. (2011). CUSE-D-r. Fragebogen zur computerbezogenen Selbstwirksamkeit - reduziert. Retrieved from <http://www.ph-heidelberg.de/wp/spannagel/cuse/CUSE-D-r.doc> (27.06.2012)
- Best, S. J., & Krueger, B. S. (2004). *Internet Data Collection*. Thousand Oaks: Sage.
- Bethlehem, J. G., & Biffignandi, S. (2012). *Handbook of Web Surveys*. Hoboken, New Jersey: Wiley.
- Bethlehem, J. G., Cobben, F., & Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. Hoboken, New Jersey: Wiley.
- Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, *74*(5), 817-848.
- Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to Survey Quality*. Hoboken, NJ: Wiley.
- Bishop, G. F. (1990). Issue Involvement and Response Effects in Public Opinion Surveys. *Public Opinion Quarterly*, *54*(2), 209-218.
- Bishop, G. F. (2008). Item Order Randomization. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (pp. 398). Thousand Oaks: Sage.
- Blumenstiel, J. E., & Roßmann, J. (2013). *Identifying and Mitigating Satisficing in Web Surveys: Some Experimental Evidence*. Paper presented at the General Online Research Conference (GOR), March 4-6, Mannheim, Germany.
- Bosnjak, M., & Tuten, T. L. (2001). Classifying Response Behaviors in Web-based Surveys. *Journal of Computer-Mediated Communication*, *6*(3), 0-0.
- Bosnjak, M., & Tuten, T. L. (2003). Prepaid and Promised Incentives in Web Surveys: An Experiment. *Social Science Computer Review*, *21*(2), 208-217.
- Bowling, A. (2005). Mode of Questionnaire Administration Can Have Serious Effects on Data Quality. *Journal of Public Health*, *27*(3), 281-291.
- Bradburn, N. M. (2004). Understanding the Question-Answer Process. *Survey Methodology*, *30*(1), 5-15.
- Bradlow, E. T., & Fitzsimons, G. J. (2001). Subscale Distance and Item Clustering Effects in Self-Administered Surveys: A New Metric. *Journal of Marketing Research*, *38*(2), 254-261.
- Brill, J. E. (2008). Likert Scale. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (pp. 427-429). Thousand Oaks: Sage.
- Statistisches Bundesamt (2014). Private Haushalte in der Informationsgesellschaft (IKT). *Fachserie*, *15*(4).
- Buxton, W. (1986). Chunking and Phrasing and the Design of Human-Computer Dialogues. *Proceedings of the IFIP World Computer Congress, Dublin, Ireland*, 475-480.
- Callegaro, M., Shand-Lubbers, J., & Dennis, J. M. (2009). *Presentation of a Single Item versus a Grid: Effects on the Vitality and Mental Health Scales of the SF-36v2 Health Survey*. Paper presented at the 64th Annual Conference of the American Association for Public Opinion Research (AAPOR), May 14-17, 2009, Hollywood, FL.
- Callegaro, M., Wells, T., & Kruse, Y. (2008). *Effects of Precoding Response Options for Five Point Satisfaction Scales in Web Surveys*. Paper presented at the Pacific

- Chapter of American Association of Public Opinion Research (AAPOR), December 11-12 2008, San Francisco, CA.
- Callegaro, M., Yang, Y., Bhola, D. S., & Dillman, D. A. (2005). Response Latency as an Indicator of Optimization: A Study Comparing Job Applicants' and Job Incumbents' Response Time On a Web Survey. *Working Paper Series of the Program in Survey Research and Methodology, University of Nebraska, Lincoln, No.11*.
- 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*(1), 5-25.
- Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on Interviewing Techniques. *Sociological Methodology, 12*, 389-437.
- Chaiken, S., & Stangor, C. (1987). Attitudes and Attitude Change. *Annual Review of Psychology, 38*(1), 575-630.
- Chan, J. C. (1991). Response-Order Effects in Likert-Type Scales. *Educational and Psychological Measurement, 51*(3), 531-540.
- Chang, L., & Krosnick, J. A. (2009). National Surveys via RDD Telephone Interviewing versus the Internet. *Public Opinion Quarterly, 73*(4), 641-678.
- Christian, L. M., & Dillman, D. A. (2004). The Influence of Graphical and Symbolic Language Manipulations on Responses to Self-Administered Questions. *Public Opinion Quarterly, 68*(1), 57-80.
- Christian, L. M., Dillman, D. A., & Smyth, J. D. (2008). The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. De Leeuw, L. Japac, P. J. Lavrakas, M. W. Link & R. L. Sangster (Eds.), *Advances in Telephone Survey Methodology* (pp. 250-275). Hoboken, NJ: Wiley.
- Christian, L. M., Parsons, N. L., & Dillman, D. A. (2009). Designing Scalar Questions for Web Surveys. *Sociological Methods Research, 37*(3), 393-425.
- Coen, T., Lorch, J., & Piekarski, L. (2005). *The Effects of Survey Frequency on Panelists' Responses*. Paper presented at the ESOMAR, 17-19 April, 2005, Budapest, Hungary.
- Cole, J., McCormick, A. C., & Bowers, A. (2012). *Straight-Lining and Survey Reluctance: Prevalence and Implications*. Paper presented at the 67th Annual Conference of the American Association for Public Opinion Research (AAPOR), May 17-20, 2012, Orlando, FL.
- Conrad, F. G., Couper, M. P., Tourangeau, R., & Galesic, M. (2005). *Interactive Feedback Can Improve the Quality of Responses in Web Surveys*. Paper presented at the 60th Annual Conference of the American Association for Public Opinion Research (AAPOR), May 12-15, 2005, Miami Beach, FL.
- Conrad, F. G., Couper, M. P., Tourangeau, R., & Peytchev, A. (2003). *Effectiveness of Progress Indicators in Web Surveys: It's What's Up Front That Counts*. Paper presented at the ASC's 4th International Conference on Survey and Statistical Computing, September 17-19, 2003, Warwick University, UK.
- Conrad, F. G., Schober, M., Jans, M., Orlovski, R., Nielsen, D., & Levenstein, R. (2008). *Features of animacy in virtual interviews*. Paper presented at the Annual conference of the American Association of Public Opinion Research.
- Conrad, F. G., & Schober, M. F. (2000). Clarifying Question Meaning in a Household Telephone Survey. *Public Opinion Quarterly, 64*(1), 1-28.
- Conrad, F. G., Schober, M. F., & Coiner, T. (2007). Bringing Features of Human Dialogue to Web Surveys. *Applied Cognitive Psychology, 21*(2), 165-187.

- Cook, C., Heath, F., & Thompson, R. L. (2000). A Meta-Analysis of Response Rates in Web- or Internet-Based Surveys. *Educational and Psychological Measurement, 60*(6), 821-836.
- Coote, L. V. (2011). Measurement Properties of Rankings and Ratings. *Journal of Business Research, 64*(12), 1296-1302.
- Couch, A., & Keniston, K. (1960). Yeasayers and Naysayers: Agreeing Response Set as a Personality Variable. *Journal of Abnormal and Social Psychology, 60*(2), 151-174.
- Couper, M. P. (2000). Web Surveys. A Review of Issues and Approaches. *Public Opinion Quarterly, 64*(4), 464-494.
- Couper, M. P. (2008). *Designing Effective Web Surveys*. New York: Cambridge University Press.
- Couper, M. P. (2011). The Future of Modes of Data Collection. *Public Opinion Quarterly, 75*(5), 889-908.
- Couper, M. P., & Bosnjak, M. (2010). Internet Surveys. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (pp. 527-550). Bingley: Emerald.
- Couper, M. P., Tourangeau, R., & Conrad, F. G. (2007). Visual Context Effects in Web Surveys. *Public Opinion Quarterly, 71*(4), 623-634.
- 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*(1), 111-127.
- Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the Effectiveness of Visual Analog Scales. *Social Science Computer Review, 24*(2), 227-245.
- Couper, M. P., Tourangeau, R., Conrad, F. G., & Zhang, C. (2013). The Design of Grids in Web Surveys. *Social Science Computer Review, 31*(3), 322-345.
- Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web Survey Design and Administration. *Public Opinion Quarterly, 65*(2), 230-253.
- Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001). Web Surveys: Perceptions of Burden. *Social Science Computer Review, 19*(2), 146-162.
- Cronbach, L. J. (1946). Response Sets and Test Validity. *Educational and Psychological Measurement, 6*(4), 475-494.
- De Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation. *Journal of Marketing Research, 45*(1), 104-115.
- De Leeuw, E. D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics, 21*(2), 233-255.
- De Leeuw, E. D., Hox, J., & Dillman, D. A. (2008). The Cornerstones of Survey Research. In E. D. De Leeuw, J. Hox & D. A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 1-17). New York: Lawrence Erlbaum.
- De Leeuw, E. D., Hox, J., & Huisman, M. (2003). Prevention and Treatment of Item Nonresponse. *Journal of Official Statistics, 19*(2), 153-176.
- Derouvray, C., & Couper, M. P. (2002). Designing a Strategy for Reducing "No Opinion" Responses in Web-Based Surveys. *Social Science Computer Review, 20*(1), 3-9.
- Deutskens, B., Ruyter, K. D., Wetzels, M., & Oosterfeld, P. (2004). Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study. *Marketing Letters, 15*(1), 21-36.
- Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for Choosing Between Multi-Item and Single-Item Scales for Construct Measurement: A Predictive Validity Perspective. *Journal of the Academy of Marketing Science, 40*(3), 434- 449.

- Dillman, D. A. (2000). *Mail and Internet Surveys. The Tailored Design Method*. New York: Wiley.
- Dillman, D. A., & Bowker, D. K. (2001). The Web Questionnaire Challenge to Survey Methodologists. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet Science* (pp. 159-178). Lengerich: Pabst Science Publishers.
- Dillman, D. A., Sinclair, M. D., & Clark, J. R. (1993). Effects of Questionnaire Length, Respondent-Friendly Design, and a Difficult Question on Response Rates for Occupant-Addressed Census Mail Surveys. *Public Opinion Quarterly*, 57(3), 289-304.
- Dillman, D. A., Smith, J. D., & Christian, L. M. (2009). *Internet, Mail and Mixed-Mode Surveys. The Tailored Design Method*. Hoboken, NJ: Wiley.
- Dillman, D. A., Tortora, R. D., & Bowker, D. (1998). *Principles for Constructing Web Surveys (Technical Report No. 98-50)*. Pullman: Washington State University Social and Economic Sciences Research Center.
- Dillmann, D. A., Eltinge, J. L., Groves, R. M., & Little, R. J. A. (2002). Survey Nonresponse in Design, Data Collection, and Analysis. In R. M. Groves, D. A. Dillmann, J. L. Eltinge & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 3-26). New York: Wiley.
- Dixon, J., & Tucker, C. (2010). Survey Nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (pp. 593-630). Bingley: Emerald.
- Dolnicar, S., Grün, B., & Yanamandram, V. (2013). Dynamic, Interactive Survey Questions Can Increase Survey Data Quality. *Journal of Travel & Tourism Marketing*, 30(7), 690-699.
- Downes-Le Guin, T., Baker, R., Mechling, J., & Ruylea, E. (2012). Myths and Realities of Respondent Engagement in Online Surveys. *International Journal of Market Research*, 54(5), 1-21.
- Draisma, S., & Dijkstra, W. (2004). Response Latency and (Para)Linguistic Expressions as Indicators of Response Error. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 131-147). Hoboken, NJ: Wiley.
- Drolet, A. L., & Morrison, D. G. (2001). Do We Really Need Multiple-Item Measures in Service Research? *Journal of Service Research*, 3(3), 196-204.
- Emde, M., & Fuchs, M. (2012). Using Adaptive Questionnaire Design in Open-Ended Questions: A Field Experiment. *JSM Proceedings of the Survey Research Methods Section, Alexandria, VA: American Statistical Association*.
- Fowler, F. J. (1992). How Unclear Terms Affect Survey Data. *Public Opinion Quarterly*, 56(2), 218-231.
- Fricker, R. D. (2008). Sampling Methods for Web and E-mail Surveys. In F. N. G., R. M. Lee & G. Blank (Eds.), *The SAGE Handbook of Online Research Methods* (pp. 195-216). Thousand Oaks: Sage.
- Fricker, S. S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An Experimental Comparison of Web and Telephone Surveys. *Public Opinion Quarterly*, 69(3), 370-392.
- Friedman, H. H., & Amoo, T. (1999). Rating the Rating Scales. *Journal of Marketing Management*, 9(3), 114-123.
- Fuchs, M., & Funke, F. (2007). Video Web Survey. Results of an Experimental Comparison with a Text-Based Web Survey. In M. Trotman, T. Burrell, L. Gerrard, K. Anderton, G. Basi, M. Couper, K. Morris, D. Birks, A. Johnson, R. Baker, M. Rigg, S. Taylor & A. Westlake (Eds.), *Challenges of a Changing World. Proceedings of the Fifth International Conference of the Association for Survey Computing* (pp. 63-80). Southampton, UK: Association for Survey Computing.

- Funke, F., & Reips, U.-D. (2009). *Twisting Rating Scales: Horizontal versus Vertical Visual Analogue Scales versus Categorical Scales in Web-Based Research*. Paper presented at the 3rd Conference of the European Survey Research Association (ESRA), June 29-July 3, Warsaw, Poland.
- Funke, F., & Reips, U.-D. (2012). Why Semantic Differentials in Web-Based Research Should Be Made from Visual Analogue Scales and Not from 5-Point Scales. *Field Methods*, 24(3), 310-327.
- Funke, F., Reips, U.-D., & Thomas, R. K. (2011). Sliders for the Smart: Type of Rating Scale on the Web Interacts with Educational Level. *Social Science Computer Review*, 29(2), 221-231.
- Galesic, M. (2006). Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey. *Journal of Official Statistics*, 22(2), 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(2), 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(5), 892-913.
- Galesic, M., & Yan, T. (2010). Use of Eye Tracking for Studying Survey Response Processes. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and behavioral research and the internet* (pp. 349-370). New York: Routledge.
- Ganassali, S. (2008). The Influence of the Design of Web Survey Questionnaires on the Quality of Responses. *Survey Research Methods*, 2(1), 21-32.
- Gerlitz, J.-Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. Dokumentation der Instrumententwicklung BFI-S auf Basis des SOEP-Pretests 2005. *DIW Research Notes* 4.
- Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A Very Brief Measure of the Big-Five Personality Domains. *Journal of Research in Personality*, 37(6), 504-528.
- Graesser, A. C., Cai, Z., Louwse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID). *Public Opinion Quarterly*, 70(1), 3-22.
- Gräf, L. (2002). Assessing Internet Questionnaires: The Online Pretest Lab. In B. Batinic, U.-D. Reips & A. Werner (Eds.), *Online Social Sciences* (pp. 49-68). Göttingen: Hogrefe.
- Grandjean, B. D., Nelson, N. M., & Taylor, P. A. (2009). *Comparing an Internet Panel Survey to Mail and Phone Surveys on Willingness to Pay for Environmental Quality: A National Mode Test*. Paper presented at the 64th Annual Conference of the American Association for Public Opinion Research (AAPOR), May 14-17, 2009, Hollywood, Florida.
- Grandmont, J., Graff, B., Goetzinger, L., & Dorbecker, K. (2010). Grappling with Grids: How Does Question Format Affect Data Quality and Respondent Engagement? *Proceedings of the American Statistical Association, Survey Research Methods Section*, 5949-5958.
- Greenleaf, E. A. (1992). Measuring Extreme Response Style. *Public Opinion Quarterly*, 56(3), 328-351.
- Groves, R. M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology*. Hoboken, NJ: Wiley.
- Groves, R. M., Presser, S., & Dipko, S. (2004). The Role of Topic Interest in Survey Participation Decisions. *Public Opinion Quarterly*, 68(1), 2-31.

- Harrison, D. A., & McLaughlin, M. E. (1993). Cognitive Processes in Self-Report Responses: Tests of Item Context Effects in Work Attitude Measures. *Journal of Applied Psychology, 78*(1), 129-140.
- Healey, B. (2007). Drop Downs and Scroll Mice: The Effect of Response Option Format and Input Mechanism Employed on Data Quality in Web Surveys. *Social Science Computer Review, 25*(1), 111-128.
- Heerwegh, D. (2003). Explaining Response Latencies and Changing Answers Using Client-Side Paradata from a Web Survey. *Social Science Computer Review, 21*(3), 360-373.
- Heerwegh, D. (2009). Mode Differences Between Face-to-Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects. *International Journal of Public Opinion Research, 21*(1), 111-121.
- Heerwegh, D., & Loosveldt, G. (2006). An Experimental Study on the Effects of Personalization, Survey Length Statements, Progress Indicators, and Survey Sponsor Logos in Web Surveys. *Journal of Official Statistics, 22*(2), 191-210.
- Heerwegh, D., & Loosveldt, G. (2008). Face-to-Face versus Web Surveying in a High-Internet-Coverage Population. *Public Opinion Quarterly, 72*(5), 836-846.
- Heerwegh, D., & Loosveldt, G. (2002). An Evaluation of the Effect of Response Formats on Data Quality in Web Surveys. *Social Science Computer Review, 20*(4), 471-484.
- Herzog, A. R., & Bachman, J. G. (1981). Effects of Questionnaire Length on Response Quality. *Public Opinion Quarterly, 45*(4), 549-559.
- Hinkin, T. R. (1995). A Review of Scale Development Practices in the Study of Organizations. *Journal of Management, 21*(5), 967-988.
- Hofmans, J., Theuns, P., Baekelandt, S., Mairesse, O., Schillewaert, N., & Cools, W. (2007). Bias and Changes in Perceived Intensity of Verbal Qualifiers Effected by Scale Orientation. *Survey Research Methods, 1*(2), 97-108.
- Holbrook, A. L. (2008). Response Order Effects. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (pp. 754-755). Thousand Oaks: Sage.
- Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias. *Public Opinion Quarterly, 67*(1), 79-125.
- Holbrook, A. L., Krosnick, J. A., Carson, R. T., & Mitchell, R. C. (2000). Violating Conversational Conventions Disrupts Cognitive Processing of Attitude Questions. *Journal of Experimental Social Psychology, 36*(5), 465-494.
- Hui, C. H., & Triadis, H. C. (1985). The Instability of Response Sets. *Public Opinion Quarterly, 49*(2), 253-260.
- Husser, J. A., & Fernandez, K. E. (2013). To Click, Type, or Drag? Evaluating Speed of Survey Data Input Methods. *Survey Practice, 6*(2).
- Jackson, D. N. (1967). Acquiescence Response Styles: Problems of Identification and Control. In I. A. Berg (Ed.), *Response Set in Personality Assessment* (pp. 71-114). Chicago, IL: Aldine.
- Jackson, D. N., & Messick, S. J. (1958). Content and Style in Personality Assessment. *Psychological Bulletin, 55*(4), 243-252.
- Jackson, D. N., & Pacine, L. (1961). Response Styles and Academic Achievement. *Educational and Psychological Measurement, 21*(4), 1015-1028.
- Jacob, R. J. K., & Karn, K. S. (2003). Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. In J. Hyönä, R. Radach & H. Deubel (Eds.), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research* (pp. 573-605). Amsterdam: Elsevier.

- Jenkins, C. R., & Dillman, D. A. (1997). Towards a Theory of Self-Administered Questionnaire Design. In L. E. Lyberg, P. P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwarz & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 165-198). New York: Wiley.
- Kaczmirek, L. (2010). Attention and Usability in Internet Surveys: Effects of Visual Feedback in Grid Questions. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet* (pp. 191-214). New York: Routledge.
- Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice Hall.
- Kalton, G., & Schuman, H. (1982). The Effect of the Question on Survey Responses: A Review. *Journal of the Royal Statistical Society. Series A (General)*, 145(1), 42-73.
- Kaminska, O., McCutcheon, A. L., & Billiet, J. (2010). Satisficing Among Reluctant Respondents in a Cross-National Context. *Public Opinion Quarterly*, 74(5), 956-984.
- Kaplowitz, M. D., Hadlock, T. D., & Levine, R. (2004). A Comparison of Web and Mail Survey Response Rates. *Public Opinion Quarterly*, 68(1), 94-101.
- Kays, K., Gathercoal, K., & Buhrow, W. (2012). Does Survey Format Influence Self-Disclosure on Sensitive Question Items? *Computers in Human Behavior*, 28(1), 251-256.
- Kelley, K., Clark, B., Brown, V., & Sitzia, J. (2003). Good Practice in the Conduct and Reporting of Survey Research. *International Journal for Quality in Health Care*, 15(3), 261-266.
- Kennedy, C. (2008). Bipolar Scale. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (pp. 64-65). Thousand Oaks: Sage.
- Kennedy, G. E. (2004). Promoting Cognition in Multimedia Interactivity Research. *Journal of Interactive Learning Research*, 15(1), 43-61.
- Keusch, F. (2012). *The Direction of Rating Scales and Its Influence on Response Behavior in Web Surveys*. Paper presented at the 67th Annual Conference of the American Association for Public Opinion Research (AAPOR), May 17-20, 2012, Orlando, FL, Orlando, Florida.
- Kieruj, N. D., & Moors, G. (2013). Response Style Behavior: Question Format Dependent or Personal Style? *Quality & Quantity*, 47(1), 193-211.
- Knapp, F., & Heidingsfelder, M. (1999). *Drop-Out-Analyse: Wirkungen des Untersuchungsdesigns*. Paper presented at the General Online Research Conference (GOR), October 28-29, Nürnberg, Germany.
- Knäuper, B. (1999). The Impact of Age and Education on Response Order Effects in Attitude Measurement. *Public Opinion Quarterly*, 63(3), 347-370.
- Knäuper, B., Belli, R. F., Hill, D. H., & Herzog, A. R. (1997). Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality. *Journal of Official Statistics*, 13(7), 181-199.
- Knowles, E. S. (1988). Item Context Effects on Personality Scales: Measuring Changes the Measure. *Journal of Personality and Social Psychology*, 55(2), 312-320.
- Knowles, E. S., & Byers, B. (1996). Reliability Shifts in Measurement Reactivity: Driven by Content Engagement or Self-Engagement? *Journal of Personality and Social Psychology*, 70(5), 1080-1090.
- Knowles, E. S., Coker, M. C., Cook, D. A., Diercks, S. R., Irwin, M. E., Lundeen, E. J., et al. (1992). Order Effects within Personality Measures. In N. Schwarz & S. Sudman (Eds.), *Context Effects in Social and Psychological Research* (pp. 221-247). New York: Springer.
- Knowles, E. S., & Condon, C. A. (1999). Why People Say 'Yes': A Dual-Process Theory of Acquiescence. *Journal of Personality and Social Psychology*, 77(2), 379-386.

- Krebs, D., & Hoffmeyer-Zlotnik, J. H. P. (2009). *Bipolar versus Unipolar Scale Format in Fully versus Endpoint Verbalized Scales*. Paper presented at the 3rd Conference of the European Survey Research Association, ESRA, 29. Juni - 3. Juli, Warsaw.
- Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly*, 72(5), 847-865.
- Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, 5(3), 213-236.
- Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology*, 50(1), 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(2), 201-219.
- Krosnick, J. A., & Alwin, D. F. (1988). A Test of the Form-Resistant Correlation Hypothesis: Ratings, Rankings, and the Measurement of Values. *Public Opinion Quarterly*, 52(4), 526-538.
- Krosnick, J. A., & Berent, M. K. (1993). Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format. *American Journal of Political Science*, 37(3), 941-964.
- Krosnick, J. A., & Fabrigar, L. R. (1997). Designing Rating Scales for Effective Measurements in Surveys. In L. E. Lyberg, P. P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwarz & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 141-164). New York: Wiley.
- Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2005). The Measurement of Attitudes. In D. Albarracín, B. T. Johnson & M. P. Zanna (Eds.), *The Handbook of Attitudes* (pp. 21-76). Mahwah, NJ: Erlbaum.
- Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in Surveys: Initial Evidence. *New Directions for Evaluation*, 1996(70), 29-44.
- 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: Emerald.
- Krosnick, J. A., & Schuman, H. (1988). Attitude Intensity, Importance, and Certainty and Susceptibility to Response Effects. *Journal of Personality and Social Psychology*, 54(6), 940-952.
- Kunz, T., & Fuchs, M. (2012). Positioning of Clarification Features in Web Surveys: Evidence from Eye Tracking Data. *JSM Proceedings of the Survey Research Methods Section*. Alexandria, VA: American Statistical Association.
- Kunz, T., & Fuchs, M. (2014). *Instant Interactive Feedback in Grid Questions: Reminding Web Survey Respondents of Speeding and Nondifferentiation*. Paper presented at the 69th Annual Conference of the American Association for Public Opinion Research (AAPOR), May 15-18, 2014, Anaheim, CA.
- Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive Burden of Survey Questions and Response Times: A Psycholinguistic Experiment. *Applied Cognitive Psychology*, 24(7), 1003-1020.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Hoboken, NJ: Wiley.
- Lozar Manfreda, K., Batagelj, Z., & Vehovar, V. (2002). Design of Web Survey Questionnaires: Three Basic Experiments. *Journal of Computer-Mediated Communication*, 7(3).
- Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web Surveys versus Other Survey Modes: A Meta-Analysis Comparing Response Rates. *International Journal of Market Research*, 50(1), 79-104.

- Lozar Manfreda, K., & Vehovar, V. (2002). *Survey Design Features Influencing Response Rates in Web Surveys*. Paper presented at the International Conference on Improving Surveys (ICIS), August 29-31, Copenhagen, Denmark.
- Lubian, D. (2010). *Measuring Attitudes: Using Branching and Numerical Scales*. Paper presented at the 65th Annual Conference of the American Association for Public Opinion Research (AAPOR), May 13-16, 2010, Chicago, IL.
- Malhotra, N. (2009). Order Effects in Complex and Simple Tasks. *Public Opinion Quarterly*, 73(1), 180-198.
- McCarty, J. A., & Shrum, L. J. (1997). Measuring the Importance of Positive Constructs: A Test of Alternative Rating Procedures. *Marketing Letters*, 8(2), 239-250.
- McCarty, J. A., & Shrum, L. J. (2000). The Measurement of Personal Values in Survey Research: A Test of Alternative Rating Procedures. *Public Opinion Quarterly*, 64(3), 271-298.
- McGee, R. K. (1967). Response Set in Relation to Personality: An Orientation. In I. A. Berg (Ed.), *Response Set in Personality Assessment* (pp. 1-31). Chicago, IL: Aldine.
- Meisenberg, G., & Williams, A. (2008). Are Acquiescent and Extreme Response Styles Related to Low Intelligence and Education? *Personality and Individual Differences*, 44(7), 1539-1550.
- Messer, B. L., Edwards, M. L., & Dillman, D. A. (2012). Determinants of Item Nonresponse to Web and Mail Respondents in Three Address-Based Mixed-Mode Surveys of the General Public. *Survey Practice*, 5(2).
- Messick, S. J. (1967). The Psychology of Acquiescence: An Interpretation of Research Evidence. In I. A. Berg (Ed.), *Response Set in Personality Assessment* (pp. 115-145). Chicago, IL: Aldine.
- Millar, M. M., & Dillman, D. A. (2012). Do Mail and Internet Surveys Produce Different Item Nonresponse Rates? An Experiment Using Random Mode Assignment. *Survey Practice*, 5(2), 1-5.
- Modick, H.-E. (1977). Ein dreiskaliger Fragebogen zur Erfassung des Leistungsmotivs: Bericht über eine deutschsprachige Weiterentwicklung des Prestatie Motivatie Test. *Diagnostica*, 23(4), 298-321.
- Moore, J. C., Stinson, L. L., & Welniak, E. J. (1997). Income Reporting in Surveys: Cognitive Issues and Measurement Error. In M. G. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur & R. Tourangeau (Eds.), *Cognition and Survey Research* (pp. 155-174). New York: Wiley.
- Narayan, S., & Krosnick, J. A. (1996). Education Moderates Some Response Effects in Attitude Measurement. *Public Opinion Quarterly*, 60(1), 58-88.
- Norman, K. L., Friedman, Z., Norman, K., & Stevenson, R. (2001). Navigational Issues in the Design of Online Self-Administered Questionnaires. *Behaviour & Information Technology*, 20(1), 37-45.
- O'Muircheartaigh, C., Krosnick, J. A., & Helic, A. (2000). Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data. *The Harris School Working Paper Series*, 01.03.
- O'Neil, K., Penrod, S., & Bornstein, B. (2003). Web-Based Research: Methodological Variables' Effects On Dropout And Sample Characteristics. *Behavior research methods, instruments, & computers*, 35(2), 217-226.
- Olson, K. (2006). Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias. *Public Opinion Quarterly*, 70(5), 737-758.
- Olson, K., & Parkhurst, B. (2013). Collecting Paradata for Measurement Error Evaluations. In F. Kreuter (Ed.), *Improving Surveys with Paradata. Analytic Uses of Process Information* (pp. 43-72). Hoboken, NJ: Wiley.

- Otto, J. H., Döring-Seipel, E., Grebe, M., & Lantermann, E.-D. (2001). Entwicklung eines Fragebogens zur Erfassung der wahrgenommenen emotionalen Intelligenz. Aufmerksamkeit auf Klarheit und Beeinflussbarkeit von Emotionen. *Diagnostica*, 47(4), 178-187.
- Paulhus, D. L. (1991). Measurement and Control of Response Bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17-59). New York: Academic Press.
- Peabody, D. (1961). Attitude Content and Agreement Set in Scales of Authoritarianism, Dogmatism, Anti-Semitism, and Economic Conservatism. *Journal of Abnormal and Social Psychology*, 63(1), 1-11.
- Peabody, D. (1962). Two Components in Bipolar Scales: Direction and Extremeness. *Psychological Review*, 69(2), 65-73.
- Peytchev, A. (2004). *Web Survey Design: Effect of Layout on Measurement Error*. Paper presented at the Sixth International Conference on Social Science Methodology (RC33), August 17-20, 2004, Amsterdam, The Netherlands.
- Peytchev, A. (2006). *Web Survey Design: Effect of Layout on Measurement Error*. Paper presented at the Sixth International Conference on Social Science Methodology (RC33), August 17-20, 2004, Amsterdam, The Netherlands.
- Peytchev, A. (2007). *Participation Decisions and Measurement Error in Web Surveys (Unpublished Doctoral Dissertation)*. University of Michigan, Ann Arbor.
- Peytchev, A. (2009). Survey Breakoff. *Public Opinion Quarterly*, 73(1), 74-97.
- Peytchev, A. (2011). Breakoff and Unit Nonresponse Across Web Surveys. *Journal of Official Statistics*, 27(1), 33-47.
- Peytchev, A., Couper, M. P., McCabe, S. E., & Crawford, S. D. (2006). Web Survey Design. Paging Versus Scrolling. *Public Opinion Quarterly*, 70(4), 596-607.
- Peytchev, A., & Crawford, S. D. (2005). A Typology of Real-Time Validations in Web-Based Surveys. *Social Science Computer Review*, 23(2), 235-249.
- Podsakoff, P. M., MacKenzie, S. B., Jeong-Yeon, L., & Podsakoff, N. P. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology*, 88(5), 879.
- Poole, A., & Ball, L. J. (2005). Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future. In C. Ghaoui (Ed.), *Encyclopedia of Human Computer Interaction*. Pennsylvania: Idea Group, Inc.
- Preston, C. C., & Colman, A. M. (2000). Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences. *Acta Psychologica*, 104(1), 1-15.
- Rammstedt, B., & Krebs, D. (2007). Does Response Scale Format Affect the Answering of Personality Scales?: Assessing the Big Five Dimensions of Personality with Different Response Scales in a Dependent Sample. *European Journal of Psychological Assessment*, 23(1), 32-38.
- Redline, C. D., & Dillman, D. A. (2002). The Influence of Alternative Visual Designs on Respondents' Performance with Branching Instructions in Self-Administered Questionnaires. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 179-193). New York: Wiley.
- Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-Ended vs. Close-Ended Questions in Web Questionnaires. *Metodološki zvezki*, 19, 159-177.
- Rorer, L. G., & Goldberg, L. R. (1965). Acquiescence in the MMPI? *Educational and Psychological Measurement*, 25(3), 801-817.
- Sakshaug, J. W., & Crawford, S. D. (2010). The Impact of Textual Messages of Encouragement on Web Survey Breakoffs: An Experiment. *International Journal of Internet Science*, 4(1), 50-60.

- Sakshaug, J. W., Yan, T., & Tourangeau, R. (2010). Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi-Mode Survey of Sensitive and Non-Sensitive Items. *Public Opinion Quarterly*, 74(5), 907-933.
- Salzberger, T., & Koller, M. (2013). Towards a New Paradigm of Measurement in Marketing. *Journal of Business Research*, 66(9), 1307-1317.
- 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(1), 61-79.
- Sarstedt, M., & Wilczynski, P. (2009). More for Less? A Comparison of Single-Item and Multi-Item Measures. *DBW - Die Betriebswirtschaft*, 69(2), 211-227.
- Schaeffer, N. C., & Presser, S. (2003). The Science of Asking Questions. *Annual Review of Sociology*, 29, 65-88.
- Scherpenzeel, A. C., & Saris, W. E. (1997). The Validity and Reliability of Survey Questions: A Meta-Analysis of MTMM Studies. *Sociological Methods & Research*, 25(3), 341-383.
- Schober, M. F., & Conrad, F. G. (1997). Does Conversational Interviewing Reduce Survey Measurement Error? *Public Opinion Quarterly*, 61(4), 576-602.
- Schonlau, M., Fricker, R. D., & Elliott, M. N. (2002). *Conducting Research Surveys via E-Mail and the Web*. Santa Monica: Rand.
- Schriesheim, C. A., & Hill, K. D. (1981). Controlling Acquiescence Response Bias by Item Reversals: The Effect on Questionnaire Validity. *Educational and Psychological Measurement*, 41(4), 1101-1114.
- Schuman, H., & Presser, S. (1996). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Thousand Oaks: Sage.
- Schwarz, N. (1990). What Respondents Learn from Scales: The Informative Functions of Response Alternatives. *International Journal of Public Opinion Research*, 2(3), 274-285.
- Schwarz, N. (1994). Judgment in a Social Context: Biases, Shortcomings, and the Logic of Conversation. *Advances in Experimental Social Psychology*, 26, 123-162.
- Schwarz, N., & Bless, H. (1992). Constructing Reality and Its Alternatives: Assimilation and Contrast Effects in Social Judgment. In L. L. Martin & A. Tesser (Eds.), *The Construction of Social Judgment* (pp. 217-245). Hillsdale, NJ: Erlbaum.
- Schwarz, N., Grayson, C. E., & Knäuper, B. (1998). Formal Features of Rating Scales and the Interpretation of Question Meaning. *International Journal of Public Opinion Research*, 10(2), 177-183.
- Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating Scales: Numeric Values May Change the Meaning of Scale Labels. *Public Opinion Quarterly*, 55(4), 570-582.
- Schwarz, N., Strack, F., Hippler, H.-J., & Bishop, G. (1991). The Impact of Administration Mode on Response Effects in Survey Measurement. *Applied Cognitive Psychology*, 5(3), 193-212.
- Schwarz, N., Strack, F., & Mai, H.-P. (1991). Assimilation and Contrast Effects in Part-Whole Question Sequences: A Conversational Logic Analysis. *Public Opinion Quarterly*, 55(1), 3-23.
- Shih, T.-H., & Xitao Fan. (2008). Comparing Response Rates from Web and Mail Surveys: A Meta-Analysis. *Field Methods*, 20(3), 249-271.
- Shoemaker, P. J., Eichholz, M., & Skewes, E. A. (2002). Item Nonresponse: Distinguishing Between Don't Know and Refuse. *International Journal of Public Opinion Research*, 14(2), 193-201.
- Shulman, A. (1973). A Comparison of Two Scales on Extremity Response Bias. *Public Opinion Quarterly*, 37(3), 407-412.

- Shulruf, B., Hattie, J., & Dixon, R. (2008). Factors Affecting Responses to Likert Type Questionnaires: Introduction of the ImpExp, a New Comprehensive Model. *Social Psychology of Education, 11*(1), 59-78.
- Sikkel, D., Steenbergen, R., & Gras, S. (2014). Clicking vs. Dragging: Different Uses of the Mouse and Their Implications for Online Surveys. *Public Opinion Quarterly, 78*(1), 177-190.
- Siminski, P. (2008). Order Effects in Batteries of Questions. *Quality & Quantity, 42*(4), 477-490.
- Sleep, D., & Puleston, J. (2009). *Panel Quality: Leveraging Interactive Techniques to Engage Online Respondents*. Paper presented at the ARF Convention & EXPO, March 30-April 1, New York City, NY.
- Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (2009). Open-Ended Questions in Web Surveys. Can Increasing the Size of Answer Boxes and Providing Extra Verbal Instructions Improve Response Quality? *Public Opinion Quarterly, 73*(2), 325-337.
- Smyth, J. D., Dillman, D. A., Christian, L. M., & O'Neill, A. C. (2010). Using the Internet to Survey Small Towns and Communities: Limitations and Possibilities in the Early 21st Century. *American Behavioral Scientist, 53*(9), 1423-1448.
- Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Effects of Using Visual Design Principles to Group Response Options in Web Surveys. *International Journal of Internet Science, 1*(1), 1-16.
- Snyder, M., & Ickes, W. (1985). Personality and Social Behavior. In L. Gardner & E. Aronson (Eds.), *Handbook of Social Psychology* (Vol. 2, pp. 883-948). New York: Random House.
- Stanley, N., & Jenkins, S. (2007). Watch What I Do: Using Graphical Input Controls in Web Surveys. In M. Trotman, T. Burrell, L. Gerrard, K. Anderton, G. Basi, M. Couper, K. Morris, D. Birks, A. Johnson, R. Baker, M. Rigg, S. Taylor & A. Westlake (Eds.), *Challenges of a Changing World. Proceedings of the Fifth International Conference of the Association for Survey Computing* (pp. 81-92). Southampton, UK: Association for Survey Computing.
- 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*(4), 377-398.
- Stieger, S., & Reips, U.-D. (2010). What Are Participants Doing While Filling In an Online Questionnaire: A Paradata Collection Tool and an Empirical Study. *Computers in Human Behavior, 26*(6), 1488-1495.
- Strack, F. (1992). "Order Effects" in Survey Research: Activation and Information Functions of Preceding Questions. In N. Schwarz & S. Sudman (Eds.), *Context Effects in Social and Psychological Research* (pp. 23-34). New York: Springer.
- Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Josey-Bass Publishers.
- Suessbrick, A., Schober, M. F., & Conrad, F. G. (2000). Different Respondents Interpret Ordinary Questions Quite Differently. *JSM Proceedings of the Survey Research Methods Section, Alexandria, VA: American Statistical Association*, 907-912.
- Taylor, E. (2006). *Non-Differentiation and Web-Based Survey Methods: An Experiment*. Paper presented at the Pacific Chapter of American Association of Public Opinion Research (PAPOR), December 7-8, 2006, San Francisco, CA.
- Thomas, R. K. (2011). *A Comparison of Visual Analog and Graphic Rating Scales in Web-Based Surveys*. Paper presented at the FedCASIC 2011 Workshop, Washington, DC, March 22-.24.,2011.

- Thorndike, F. P., Carlbring, P., Smyth, F. L., Magee, J. C., Gonder-Frederick, L., Ost, L.-G., et al. (2009). Web-Based Measurement: Effect of Completing Single or Multiple Items per Webpage. *Computers in Human Behavior*, 25(2), 393-401.
- Toepoel, V., & Couper, M. P. (2011). Can Verbal Instructions Counteract Visual Context Effects in Web Surveys? *Public Opinion Quarterly*, 75(1), 1-18.
- Toepoel, V., Das, M., & Van Soest, A. (2008). Effects of Design in Web Surveys. Comparing Trained and Fresh Respondents *Public Opinion Quarterly*, 72(5), 985-1007.
- Toepoel, V., Das, M., & Van Soest, A. (2009a). Design of Web Questionnaires: The Effect of Layout in Rating Scales. *Journal of Official Statistics*, 25(4), 509-528.
- Toepoel, V., Das, M., & Van Soest, A. (2009b). Design of Web Questionnaires: The Effects of the Number of Items per Screen. *Field Methods*, 21(2), 200-213.
- Toepoel, V., & Dillman, D. A. (2008). *Words, Numbers and Visual Heuristics in Web Surveys: Is There a Hierarchy of Importance?* : CentER Discussion Paper No. 2008-92, CentER, Tilburg University, The Netherlands.
- Toepoel, V., & Dillman, D. A. (2010). How Visual Design Affects the Interpretability of Survey Questions. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet* (pp. 165-190). New York: Routledge.
- 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 Design: Building a Bridge Between Disciplines* (pp. 73-100). Washington, DC: National Academy Press.
- Tourangeau, R. (1992). Context Effects on Responses to Attitude Questions. In N. Schwarz & S. Sudman (Eds.), *Context Effects in Social and Psychological Research* (pp. 35-47). New York: Springer.
- Tourangeau, R., Conrad, F. G., Arens, Z., Fricker, S., Lee, S., & Smith, E. (2006). Everyday Concepts and Classification Errors: Judgments of Disability and Residence. *Journal of Official Statistics*, 22(3), 385-418.
- Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The Science of Web Surveys*. New York: Oxford University Press.
- Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, Position, and Order. Interpretative Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly*, 68(3), 368-393.
- Tourangeau, R., Couper, M. P., & Conrad, F. G. (2007a). Color, Labels, and Interpretive Heuristics for Response Scales. *Public Opinion Quarterly*, 71(1), 91-112.
- Tourangeau, R., Couper, M. P., & Conrad, F. G. (2007b). *The Impact of the Visible: The Design of Web Surveys*. Paper presented at the Workshop on Internet Survey Methodology, September 17-19, Lillehammer (Norway).
- Tourangeau, R., Groves, R. M., Kennedy, C., & Yan, T. (2009). The Presentation of a Web Survey, Nonresponse and Measurement Error among Members of Web Panel. *Journal of Official Statistics*, 25(3), 299-321.
- Tourangeau, R., Groves, R. M., & Redline, C. D. (2010). Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error. *Public Opinion Quarterly*, 74(3), 413-432.
- Tourangeau, R., & Rasinski, K. A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin*, 103(3), 299-314.
- Tourangeau, R., Rasinski, K. A., & D'Andrade, R. (1991). Attitude Structure and Belief Accessibility. *Journal of Experimental Social Psychology*, 27(1), 48-75.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.

- Tourangeau, R., & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin*, *133*(5), 859-883.
- Tress, F. (2012). *Bad Boy Matrix Question – Whatcha Gonna Do When They Come for You?* Paper presented at the General Online Research (GOR), March 5-7 2012, Mannheim, Germany.
- Trouteaud, A. R. (2004). How You Ask Counts: A Test of Internet-Related Components of Response Rates to a Web-Based Survey. *Social Science Computer Review*, *22*(3), 385-392.
- Van Dijk, T. K., Datema, F., Piggen, A.-L. J. H. F., Welten, S. C. M., & Van de Vijver, F. J. R. (2009). Acquiescence and Extremity in Cross-National Surveys: Domain Dependence and Country-Level Correlates *Quod Erat Demonstrandum: From Herodotus? Ethnographic Journeys to Cross-Cultural Research* (pp. 149-158). Athens: Pedio Books Publishing.
- Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response Styles in Rating Scales. *Journal of Cross-Cultural Psychology*, *35*(3), 346-360.
- 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*(2), 195-217.
- Vehovar, V., Batagelj, Z., Lozar Manfreda, K., & Zaletel, M. (2002). Nonresponse in Web Surveys. In R. M. Groves, D. A. Dillmann, J. L. Eltinge & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 229-242). New York: Wiley.
- Vicente, P., & Reis, E. (2010). Using Questionnaire Design to Fight Nonresponse Bias in Web Surveys. *Social Science Computer Review*, *28*(2), 251-267.
- Viswanathan, M. (2005). *Measurement Error and Research Design*. Thousand Oaks: Sage.
- Weijters, B., & Baumgartner, H. (2012). Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research*, *49*(5), 737-747.
- Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed Item Bias: An Integrative Model. *Psychological Methods*, *18*(3), 320-334.
- Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The Effect of Rating Scale Format on Response Styles: The Number of Response Categories and Response Category Labels. *International Journal of Research in Marketing*, *27*(3), 236-247.
- Weijters, B., Geuens, M., & Schillewaert, N. (2009). The Proximity Effect: The Role of Inter-Item Distance on Reverse-Item Bias. *International Journal of Research in Marketing*, *26*(1), 2-12.
- Weinhardt, M., & Schupp, J. (2011). *Multi-Itemskalen im SOEP Jugendfragebogen*. Data Documentation (60), Berlin: Deutsches Institut für Wirtschaftsforschung (DIW).
- Weisberg, H. F. (2005). *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago: 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*(4), 378-390.
- Weng, L.-J., & Cheng, C.-P. (2000). Effects of Response Order on Likert-Type Scales. *Educational and Psychological Measurement*, *60*(6), 908-924.
- Wright, K. B. (2005). Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services. *Journal of Computer-Mediated Communication*, *10*(3).
- Yan, T. (2005). *Gricean Effects in Self-Administered Surveys (Unpublished Doctoral Dissertation)*. University of Maryland, Maryland
- Yan, T., Conrad, F. G., Tourangeau, R., & Couper, M. P. (2011). Should I Stay or Should I Go: The Effects of Progress Feedback, Promised Task Duration, and Length of

- Questionnaire on Completing Web Surveys. *International Journal of Public Opinion Research*, 23(2), 131-147.
- Yan, T., & Curtin, R. (2010). The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective. *International Journal of Public Opinion Research*, 22(4), 535-551.
- 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(1), 51-68.
- Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., et al. (2011). Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples. *Public Opinion Quarterly*, 75(4), 709-747.
- Zhang, C., & Conrad, F. (2013). Speeding in Web Surveys: The Tendency to Answer Very Fast and its Association with Straightlining. *Survey Research Methods*, 8(2), 127-135.
- Ziniel, S. (2008). Split-Half. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (pp. 834-835). Thousand Oaks: Sage.

APPENDIX A: SCALE AND ITEM CHARACTERISTICS

Table 48: Otto and colleagues' (2001) three-dimensional scale on perceived emotional intelligence (Experiment 1.1)

item#	item# reduced	Factor	Item wording (original in German)	Item wording (own translation into English)	10-item (4/4/2)	16-item (6/6/4)
1	1	1 _a	Ich kimmere mich gewöhnlich wenig darum, was ich gerade fühle.	Usually, I am hardly concerned with my current feelings.	X	X
2	2	2 _a	Manchmal kann ich gar nicht sagen, was meine Gefühle sind.	Sometimes it's hard for me to say what my feelings are about.	X	X
3		2 _c	Ich bin selten darüber im Unklaren, wie ich mich fühle. (-)	I am rarely uncertain about my feelings.		X
4	3	3 _a	Obwohl ich manchmal traurig bin, schaue ich meistens optimistisch in die Zukunft. (-)	Although sometimes I'm sad, I'm optimistic about the future in general.	X	X
5		3 _b	Wenn ich emotional aufgewühlt bin, wird mir klar, dass die "guten Dinge im Leben"- Illusionen sind.	When I'm troubled emotionally, I realize "the good things in life" are just illusions.		X
6		1 _c	Ich glaube daran, beim Handeln das Herz sprechen zu lassen. (-)	I believe in letting the heart speak in actions.		X
7		3 _b	Wenn ich aus der Fassung gerate, rufe ich mir all die angenehmen Seiten des Lebens ins Gedächtnis. (-)	When I lose countenance, I remind myself of the enjoyable facets of life.		X
8	4	2 _b	Gewöhnlich bin ich im Unklaren darüber, wie ich mich fühle.	Usually I'm uncertain about my feelings.	X	X
9		1 _c	Man sollte sich niemals von seinen Gefühlen leiten lassen.	You should never allow your feelings to guide you.		X
10	5	3 _a	Wenn ich auch manchmal glücklich bin, schaue ich meistens pessimistisch in die Zukunft.	Although sometimes I'm happy, I'm pessimistic about the future in general.	X	X
11	6	1 _a	Ich achte sehr darauf, wie ich mich fühle. (-)	Usually, I am very concerned with my current feelings.	X	X
12		2 _c	Ich kann mir keinen Reim auf meine Gefühle machen.	It's hard to make sense of my feelings.		X
13	7	1 _b	Ich widme meinen Gefühlen nicht viel Aufmerksamkeit.	I don't pay much attention to my feelings.	X	X
14	8	1 _c	Ich denke oft über meine Gefühle nach. (-)	I often think about my feelings.	X	X
15	9	2 _b	Ich bin mir gewöhnlich über meine Gefühle sehr im Klaren. (-)	Usually I'm quite certain about my feelings.	X	X
16	10	2 _a	Ich weiß fast immer genau, wie ich mich fühle. (-)	I always know how I feel.	X	X

Table 49: Modick's (1977) three-dimensional scale on achievement motive (Experiment 1.2/ 1.3)

item#	item# reduced	Factor	Item wording (original in German)	Item wording (own translation into English)	6-item (2/2/2)	10-item (4/4/2)	16-item (6/6/4)
1	1	2 _a	Auch in kritischen Situationen behalte ich einen kühlen Kopf. (*)	Even in critical situations I keep my cool.	X	X	X
2	2	1 _a	Ich plane nicht gern, sondern lasse lieber alles auf mich zukommen. (*)	I don't like making plans, I prefer letting things happen.	X	X	X
3	3	3 _a	Ich kann besser denken, wenn ich ein leichtes Gefühl ängstlicher-Spannung habe. (-)	A feeling of anxious tension makes me see matters more clearly.	X	X	X
4	4	1 _a	Ich halte es für wichtig, meine Zukunft vorzustrukturieren. (-)	I think scheduling my future is important.	X	X	X
5		2 _b	Wenn ich während einer Prüfung Angst habe, lässt mich mein Gedächtnis oft im Stich. (-)	When I'm afraid during an exam, often my memory lets me down.		X	X
6	5	3 _a	Wenn mich etwas in Spannung versetzt, kann ich weniger gut arbeiten als sonst.	If something pressures me, I work less effectively than usual.	X	X	X
7		2 _c	Ich bin sehr aufgeregt, wenn ich mich einer Prüfung unterziehen muss. (-)	When taking an exam, I am very excited.			X
8		2 _b	Auf mein Denkvermögen ist auch bei leichten Angstgefühlen immer Verlass. (*)	My memory is very reliable, even when I'm a little bit anxious.		X	X
9		1 _c	Für mich ist es nicht wichtig, mehr zu leisten als andere. (*)	It's not important for me to work harder than other people.			X
10		3 _b	Ich komme meistens zu besseren Leistungen, wenn ich etwas angespannt bin. (-)	Usually I achieve better results when I'm tensed.			X
11		1 _b	Ich versuche, mein Leben über einen längeren Zeitraum hinweg zu planen. (-)	I try to make plans for my life over a longer period of time.		X	X
12		2 _c	Prüfungen schaue ich gelassen entgegen. (*)	I face exams untroubled.			X
13		1 _b	Im Allgemeinen bin ich wenig auf die Zukunft ausgerichtet.	Generally, I don't aim my action to the future too much.		X	X
14		3 _b	Das Gefühl von Spannung ist für meine Leistung oft ungünstig. (*)	The feeling of tension often compromises my performance.			X
15		1 _c	Ein gewisses Maß an Wettbewerb kann nicht schaden. (-)	There is no harm in a little competition.			X
16	6	2 _a	Mir passiert es oft, dass ich in kritischen Situationen Fehler mache. (-)	I often make errors in critical situations.	X	X	X

Table 50: Gerlitz and Schupp's (2005) Ten-Item Personality Inventory (TIPI) (Experiment 2)

Item order	Factor	Item wording (original in German)	Item wording (original in English)
1	1	Ich bin zuverlässig. (-)	I see myself as dependable, self-disciplined.
2	2	Ich bin leicht aufzuregen. (-)	I see myself as anxious, easily upset.
3	3	Ich bin offen für neue Erfahrungen. (-)	I see myself as open to new experiences, complex.
4	4	Ich bin zurückhaltend.	I see myself as reserved, quiet.
5	5	Ich bin mitfühlend, warmherzig. (-)	I see myself as sympathetic, warm.
6	1	Ich bin unachtsam.	I see myself as disorganized, careless.
7	5	Ich bin kritisch.	I see myself as critical, quarrelsome.
8	2	Ich bin gefühlsmäßig stabil.	I see myself as calm, emotionally stable.
9	3	Ich bin konventionell.	I see myself as conventional, uncreative.
10	4	Ich bin extrovertiert. (-)	I see myself as extraverted, enthusiastic.

Table 51: Scale on reasons for social advancement published in Weinhardt and Schupp (2011) (Experiment 3.1)

Item order	Factor	Item wording (original in German)	Item wording (own translation into English)
1	1	Man muss sich anstrengen und fleißig sein.	You have to make an effort and work hard.
2	1	Man muss begabt und intelligent sein.	You have to be talented and intelligent.
3	2	Man muss aus der richtigen Familie stammen.	You have to descend from the proper family.
4	1	Man muss gute Fachkenntnisse auf seinem Spezialgebiet haben.	You have to have expert knowledge on your discipline.
5	1	Man muss einen möglichst guten Schulabschluss haben.	You have to receive a good graduation certificate.
6	2	Man muss rücksichtslos und hart sein.	You have to be ruthless and hard.
7	2	Man muss Beziehungen zu den richtigen Leuten haben.	You have to make connections with the right people.
8	2	Man muss sich auf der richtigen Seite politisch engagieren.	You have to act politically on the right side.

Table 52: Scale on locus of control published in Weinhardt and Schupp (2011) (Experiment 3.2)

Item order	Factor	Item wording (original in German)	Item wording (own translation into English)
1	1	Wie mein Leben verläuft, hängt von mir selbst ab.	It depends on me how my life goes.
2	2	Was man im Leben erreicht, ist in erster Linie eine Frage von Schicksal oder Glück.	What you achieve in life primarily depends on destiny and luck.
3	2	Ich mache häufig die Erfahrung, dass andere über mein Leben bestimmen.	I often experience others taking control of my life.
4	1	Erfolg muss man sich hart erarbeiten.	You have to work hard to succeed.
5	2	Wenn ich im Leben auf Schwierigkeiten stoße, zweifle ich oft an meinen Fähigkeiten.	When facing difficulties in my life, I often doubt my abilities.
6	2	Welche Möglichkeiten ich im Leben habe, wird von den sozialen Umständen bestimmt.	My possibilities in life are determined by the social circumstances.
7	1	Wichtiger als alle Anstrengungen sind die Fähigkeiten, die man mitbringt.	The abilities you are equipped with are more important than the effort you take.
8	2	Ich habe wenig Kontrolle über die Dinge, die in meinem Leben passieren.	I have little control over the things that happen in my life.

APPENDIX B: SCALE PROPERTIES

Table 53: Principal components analysis (Experiment 1.1.1, $n = 714$)

item# reduced	Grid			Drag-R			Drag-I			One-V			One-H		
	C1	C2	C3												
1	.728	.285	-.060	.827	.166	.017	.843	.095	.142	.871	.136	.045	.770	.085	.095
2	.099	.808	.002	.274	.760	.024	.086	.836	.111	.076	.857	.016	-.036	.749	.148
3 (-)	.064	.108	.882	-.039	-.022	.896	.010	.209	.760	.082	.144	.867	-.037	.078	.902
4	.076	.780	.225	.246	.705	.196	.124	.666	.369	.164	.822	.201	.140	.823	.189
5	-.056	.167	.857	-.010	.170	.860	.009	.029	.905	.026	.105	.898	.084	.149	.899
6 (-)	.753	.225	.065	.806	.204	-.036	.821	.130	.056	.851	.133	.176	.803	.181	-.011
7	.872	.138	-.025	.881	.214	.001	.873	.117	.066	.866	.234	.106	.827	.277	.013
8 (-)	.806	-.063	.039	.735	.095	-.033	.764	.135	-.278	.823	.065	-.117	.796	-.155	-.033
9 (-)	.219	.816	.121	.137	.857	.040	.099	.853	-.025	.151	.868	.128	.141	.869	-.012
10 (-)	.160	.866	.102	.071	.885	-.011	.180	.834	.060	.171	.878	.066	.123	.858	.020
% ¹⁾	70.7			71.4			71.2			77.4			72.1		

Note. Items marked with a minus (-) were recoded prior to analysis. Salient loading values indicating assignment to a principal component (C1, C2, and C3) are printed in bold. 1) The percentage of variance explained by the extracted components.

Table 54: Principal components analysis (Experiment 1.2, $n = 4,813$)

item# reduced	Grid			Drag-R			Drag-I			One-V			One-H		
	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
1	.051	.707	-.004	-.025	.761	.019	.043	.804	-.003	.027	.796	.042	-.038	.757	.058
2	.826	-.082	-.152	.833	-.018	-.123	.822	-.040	-.015	.821	-.053	-.072	.793	-.049	-.194
3 (-)	-.030	.046	.939	-.018	.006	.946	-.008	.057	.918	-.014	.081	.927	-.012	.048	.928
4 (-)	.833	.097	.100	.848	.010	.081	.820	.041	-.015	.818	-.009	.061	.828	-.001	.155
5	-.031	-.618	.549	-.062	-.651	.451	-.040	-.493	.629	.011	-.576	.579	-.049	-.637	.475
6 (-)	-.053	.837	.039	-.016	.794	.053	-.052	.792	-.097	-.100	.820	-.028	-.043	.820	.049
% ¹⁾	70.0			69.6			68.8			70.1			68.9		

Note: Items marked with a minus (-) were recoded prior to analysis. Salient loading values indicating assignment to a principal component (C1, C2, and C3) are printed in bold. 1) The percentage of variance explained by the extracted components.

Table 55: Principal components analysis (Experiment 1.3, $n = 5,529$)

item# reduced	Grid			Drag-R			Drag-I			One-V			One-H		
	C1	C2	C3												
1	.145	.681	-.050	-.003	.739	.045	.134	.785	-.018	.096	.783	.025	.024	.801	.037
2	.835	-.069	-.114	.825	-.067	-.044	.842	-.038	-.051	.828	-.007	-.103	.831	-.060	-.101
3 (-)	-.058	.053	.931	-.027	.111	.921	-.028	.105	.903	-.079	.117	.883	-.009	.119	.891
4 (-)	.797	.088	.081	.835	.041	-.015	.817	.049	.065	.820	.009	.029	.835	.054	.085
5	.099	-600	.522	-.067	-.573	.587	.062	-.441	.681	.008	-.417	.694	-.009	-.443	.675
6 (-)	-.093	.817	.115	-.045	.787	-.009	-.113	.801	-.080	-.095	.762	-.127	-.031	.796	-.120
% ¹⁾		67.6		68.2		69.6		67.5		69.4					

Note. Items marked with a minus (-) were recoded prior to analysis. Salient loading values indicating assignment to a principal component (C1, C2, and C3) are printed in bold. 1) The percentage of variance explained by the extracted components.

Table 56: Principal components analysis (Experiment 2, $n = 727$)

item#	Grid					Drag-R					Drag-I					One-V					One-H				
	C1	C2	C3	C1	C2	C3	C4	C5	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	
1 (-)	-.276	.689	-.105	-.107	.115	.670	.204	-.205	.080	.709	-.077	-.042	-.108	.067	.788	.005	-.045	.080	.708	-.066	-.045	.080	.708	-.066	
2 (-)	.010	-.152	.764	.013	.833	.043	-.059	-.254	.039	.068	.840	-.125	.001	.724	.229	-.154	-.048	.780	-.074	-.005	-.048	.780	-.074	-.005	
3 (-)	.532	.508	-.057	.122	-.402	-.090	.688	-.262	.574	-.304	-.347	-.019	.706	-.200	-.119	.144	.541	-.257	-.235	.430	.541	-.257	-.235	.430	
4	.718	-.069	-.196	.893	-.076	-.037	.070	-.039	.757	.117	-.021	.037	.778	.025	.053	-.271	.850	-.110	.049	-.018	.850	-.110	.049	-.018	
5 (-)	.269	.582	-.095	.134	.207	.115	.805	.238	.161	.186	.183	.803	.152	.137	.173	.815	.124	.340	.582	.000	.124	.340	.582	.000	
6	.016	.554	-.405	.258	-.150	.712	-.138	.213	.004	.642	.031	.118	.069	-.036	.776	.013	-.096	-.288	.765	.072	-.096	-.288	.765	.072	
7	.038	.506	-.131	-.081	.007	.018	.042	.885	-.182	-.162	-.351	.674	-.166	-.150	-.128	.639	-.297	.028	.649	.010	-.297	.028	.649	.010	
8	-.031	.069	.838	-.115	.760	-.139	.068	.306	-.409	-.276	.647	.209	-.101	.837	-.134	.076	-.033	.769	.133	.114	-.033	.769	.133	.114	
9	.523	-.030	.334	.306	.056	-. 604	.021	-.042	.120	-. 694	-.042	.138	.448	.397	-.251	.224	.221	.117	.734	.013	.221	.117	.734	.013	
10 (-)	.810	.016	.029	.879	-.027	-.101	.132	-.059	.877	-.050	.038	-.003	.848	.039	.070	-.006	.871	.121	-.016	-.059	.871	.121	-.016	-.059	
% ¹⁾		51.9			70.7					61.8					62.4									61.3	

Note. Items marked with a minus (-) were recoded prior to analysis. Salient loading values indicating assignment to a principal component (C1, C2, C3, C4, and C5) are printed in bold. 1) The percentage of variance explained by the extracted components.

Table 57: Principal components analysis (Experiment 3.1, $n = 5,211$)

item#	Grid		Drag-R		Drag-I	
	C1	C2	C1	C2	C1	C2
1	.721	-.307	.518	-.370	.577	-.200
2	.641	.059	.588	.160	.598	.126
2	-.117	.707	-.058	.690	-.041	.703
1	.788	-.187	.718	-.154	.665	-.095
2	.522	.146	.633	.095	.591	.165
2	-.027	.748	-.005	.710	-.032	.711
2	.297	.679	.044	.752	.048	.759
% ¹⁾	49.7		43.5		43.0	

Note. Salient loading values indicating assignment to a principal component (C1 and C2) are printed in bold. 1) The percentage of variance explained by the extracted components.

Table 58: Principal components analysis (Experiment 3.2, $n = 5,227$)

item#	Grid		Drag-R		Drag-I	
	C1	C2	C1	C2	C1	C2
1	.687	-.380	.531	-.400	.595	-.404
2	-.314	.503	-.185	.484	-.172	.470
2	-.176	.698	-.097	.679	-.029	.704
1	.885	.045	.852	.135	.850	.136
2	.022	.704	.142	.655	.130	.665
2	-.031	.471	-.374	.346	-.268	.341
2	-.294	.624	-.226	.592	-.186	.619
% ¹⁾	49.5		43.4		43.8	

Note. Salient loading values indicating assignment to a principal component (C1 and C2) are printed in bold. Inclusion of item #7 originally assigned to subscale 1 "internal locus of control" resulted in a three-dimensional scale structure with item #7 being negatively correlated with the actual subscale 1, which is why this item was excluded from subsequent analyses. 1) The percentage of variance explained by the extracted components.

Table 59: Internal consistency reliability and item means depending on scale format (Experiment 1.1, $n = 714$)

	Format				
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H
Overall: Cronbach's α	.795	.794	.779	.832	.773
Subscale 1: Cronbach's α	.816	.847	.855	.886	.818
1	3.53 (1.06)	3.57 (1.07)	3.68 (1.05)	3.44 (1.05)	3.64 (1.06)
6 (-)	3.25 (0.95)	3.36 (0.96)	3.44 (0.99)	3.17 (0.99)	3.34 (0.95)
7	3.62 (1.00)	3.66 (0.98)	3.74 ^d (1.08)	3.36 ^c (1.07)	3.53 (1.03)
8 (-)	3.25 ^c (1.08)	3.30 ^c (1.12)	3.69 ^{a,b,d} (1.00)	3.07 ^c (1.19)	3.37 (1.08)
Subscale 2: Cronbach's α	.857	.839	.834	.889	.849
2	2.89 (1.07)	2.98 (1.13)	3.09 (1.02)	2.92 (1.23)	3.06 (1.11)
4	3.74 (0.96)	3.86 (0.81)	3.66 (1.01)	3.64 (0.95)	3.72 (0.92)
9 (-)	3.33 (0.90)	3.46 (0.92)	3.49 (0.87)	3.32 (0.92)	3.51 (0.87)
10 (-)	3.26 (0.96)	3.35 (0.91)	3.36 (1.07)	3.38 (0.92)	3.42 (0.88)
Subscale 3: Cronbach's α	.719	.720	.674	.756	.795
3 (-)	3.72 (1.04)	3.74 (1.04)	3.65 (1.17)	3.91 (0.93)	3.92 (0.89)
5	4.06 (1.05)	4.10 (0.95)	3.93 (1.27)	4.01 (0.99)	4.06 (0.93)

Note. Standard deviations are indicated in parentheses. Items marked with a minus (-) were recoded prior to analysis. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e. compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e).

Table 60: Internal consistency reliability and item means depending on scale format (Experiment 1.2, $n = 4,813$)

	Format				
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H
Overall: Cronbach's α	.072	.045	.073	.087	.039
Subscale 1: Cronbach's α	.548	.588	.518	.517	.483
2	3.61 (1.02)	3.59 (0.99)	3.58 (1.02)	3.54 (1.02)	3.61 (0.99)
4 (-)	3.69 (0.91)	3.77 (0.94)	3.76 (1.01)	3.79 (0.95)	3.74 (0.94)
Subscale 2: Cronbach's α	.495	.511	.524	.578	.558
1	1.97 ^{b,c} (0.71)	1.86 ^a (0.72)	1.84 ^{b,e} (0.71)	1.91 (0.69)	1.94 ^c (0.70)
6 (-)	2.38 (0.82)	2.33 (0.80)	2.38 (0.86)	2.39 (0.80)	2.34 (0.75)
Subscale 3: Cronbach's α	.479	.386	.481	.442	.367
3 (-)	2.71 (1.04)	2.73 (1.00)	2.76 (1.13)	2.75 (1.01)	2.72 (1.01)
5	3.49 (0.97)	3.55 (0.89)	3.53 (1.04)	3.46 (0.93)	3.53 (0.90)

Note. Standard deviations are indicated in parentheses. Items marked with a minus (-) were recoded prior to analysis. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the five scale formats, i.e. compared to grid (a), drag-response (b), drag-item (c), one-vertical (d), and one-horizontal format (e).

Table 61: Internal consistency reliability and item means depending on scale format (Experiment 1.3, $n = 5,529$)

	Format				
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H
Overall: Cronbach's α	.131	.054	.200	.081	.160
Subscale 1: Cronbach's α	.524	.552	.564	.541	.557
2	3.62 (1.05)	3.62 (1.00)	3.61 (1.10)	3.58 (1.02)	3.62 (1.03)
4 (-)	3.63 (0.95)	3.65 (0.98)	3.66 (1.02)	3.69 (0.95)	3.68 (0.95)
Subscale 2: Cronbach's α	.438	.447	.497	.428	.528
1	2.01 (0.73)	1.97 (0.73)	1.93 (0.73)	1.95 (0.73)	1.98 (0.72)
6 (-)	2.41 (0.81)	2.36 (0.79)	2.41 (0.90)	2.35 (0.74)	2.39 (0.79)
Subscale 3: Cronbach's α	.406	.420	.495	.473	.467
3 (-)	2.67 (1.05)	2.67 (1.03)	2.67 (1.08)	2.75 (1.06)	2.67 (1.00)
5	3.42 (0.96)	3.45 (0.92)	3.42 (1.00)	3.41 (0.96)	3.39 (0.96)

Note. Standard deviations are indicated in parentheses. Items marked with a minus (-) were recoded prior to analysis.

Table 62: Internal consistency reliability and item means depending on scale format (Experiment 2, $n = 727$)

	Format				
	(a) Grid	(b) Drag-R	(c) Drag-I	(d) One-V	(e) One-H
Overall: Cronbach's α	.346	.292	.131	.416	.378
Subscale 1: Cronbach's α	.507	.273	.388	.459	.457
1 (-)	4.34 (0.71)	4.31 (0.76)	4.46 (0.70)	4.40 (0.61)	4.33 (0.71)
6	3.60 (1.02)	3.71 (0.81)	3.88 (0.99)	3.82 (0.84)	3.86 (0.78)
Subscale 2: Cronbach's α	.619	.548	.474	.517	.510
2 (-)	2.95 (1.11)	2.81 (1.00)	2.80 (1.09)	2.80 (0.98)	2.76 (1.02)
8	2.44 (1.02)	2.35 (1.01)	2.34 (1.01)	2.30 (0.97)	2.36 (0.95)
Subscale 3: Cronbach's α	.339	.079	.342	.372	.387
3 (-)	3.91 (0.86)	3.87 (0.86)	4.01 (0.96)	3.98 (0.82)	3.91 (0.80)
9	3.03 (0.91)	3.11 (0.94)	3.15 (0.95)	2.99 (0.93)	2.96 (0.84)
Subscale 4: Cronbach's α	.634	.819	.672	.765	.777
4	2.82 (1.03)	2.99 (1.10)	2.84 (1.14)	2.81 (1.01)	2.92 (1.06)
10 (-)	2.74 (1.03)	2.70 (1.12)	2.62 (1.14)	2.70 (1.06)	2.70 (1.15)
Subscale 5: Cronbach's α	-.268	.192	.291	.297	.031
5 (-)	3.84 (1.01)	3.83 (0.88)	3.85 (0.95)	3.89 (0.83)	3.92 (0.89)
7	2.23 (0.89)	2.22 (0.87)	2.23 (0.90)	2.11 (0.87)	2.13 (0.78)

Note. Standard deviations are indicated in parentheses. Items marked with a minus (-) were recoded prior to analysis. Although none of the five scale formats could exactly replicate the five-dimensional structure of the rating scale, calculations of subscale internal consistency reliability and designation of item means refer to the ideal factor structure.

Table 63: Internal consistency reliability and item means depending on scale format (Experiment 3.1, $n = 5,211$)

	Format		
	(a) Grid	(b) Drag-R	(c) Drag-I
Overall: Cronbach's α	.466	.404	.488
Subscale 1: Cronbach's α	.605	.452	.422
1	1.57 ^{b,c} (0.87)	1.49 ^a (0.71)	1.48 ^a (0.69)
2	2.26 (0.92)	2.26 (0.85)	2.20 (0.86)
4	1.65 ^c (0.85)	1.59 ^c (0.71)	1.53 ^{a,b} (0.71)
5	2.66 ^c (1.08)	2.60 (1.01)	2.57 ^a (1.03)
Subscale 2: Cronbach's α	.636	.593	.626
3	3.66 ^{b,c} (1.12)	3.76 ^{a,c} (1.05)	3.89 ^{a,b} (1.04)
6	3.70 ^{b,c} (1.05)	3.81 ^{a,c} (0.99)	3.96 ^{a,b} (1.01)
7	2.37 (1.04)	2.39 (0.98)	2.45 (1.01)
8	3.77 ^c (0.97)	3.77 ^c (0.92)	3.89 ^{a,b} (0.93)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale formats, i.e. compared to grid (a), drag-response (b), and drag-item format (c).

Table 64: Internal consistency reliability and item means depending on scale format (Experiment 3.2, $n = 5,227$)

	Format		
	(a) Grid	(b) Drag-R	(c) Drag-I
Overall: Cronbach's α	.222	.239	.278
Subscale 1: Cronbach's α	.535	.289	.318
1	1.81 ^{b,c} (0.85)	1.70 ^a (0.72)	1.68 ^a (0.73)
4	2.02 ^{b,c} (0.87)	1.94 ^{a,c} (0.75)	1.86 ^{a,b} (0.78)
Subscale 2: Cronbach's α	.617	.528	.525
2	3.60 ^c (0.90)	3.62 (0.84)	3.68 ^a (0.92)
3	3.92 ^c (0.89)	3.98 (0.82)	4.01 ^a (0.91)
5	3.61 ^c (0.98)	3.60 ^c (0.95)	3.70 ^{a,b} (0.98)
6	3.11 ^c (0.94)	3.16 (0.93)	3.23 ^a (0.97)
8	4.00 ^{b,c} (0.89)	4.07 ^{a,c} (0.83)	4.17 ^{a,b} (0.84)

Note. Standard deviations are indicated in parentheses. Calculations were based on ANOVAs with pair-wise comparisons (Bonferroni correction): Lowercase superscripts indicate a significant difference ($p < .05$ or less) between any two of the three scale formats, i.e. compared to grid (a), drag-response (b), and drag-item format (c).

APPENDIX C: STATEMENT OF ACADEMIC HONESTY

Hiermit erkläre ich, dass ich die beigefügte Dissertation selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel genutzt habe. Alle wörtlich oder inhaltlich übernommenen Stellen habe ich als solche gekennzeichnet.

Ich versichere außerdem, dass ich die beigefügte Dissertation nur in diesem und keinem anderen Promotionsverfahren eingereicht habe und dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind.

Darmstadt, den 19.11.2015

APPENDIX D: CURRICULUM VITAE

Seit 01/2010	Technische Universität Darmstadt Wissenschaftliche Mitarbeiterin im Fachbereich Gesellschafts- und Geschichtswissenschaften, Institut Empirische Sozialforschung
10/2003-05/2009	Universität Mannheim Diplom Sozialwissenschaften