UNIVERSITEIT GENT

FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

DEPARTMENT OF MARKETING

Response Styles in Consumer Research

Bert Weijters

2006

Supervisor: Prof Dr Maggie Geuens
Co-supervisor: Prof Dr Niels Schillewaert

Dissertation submitted to
the Faculty of Economics and Business Administration,
Ghent University,
in fulfillment of the requirements for the degree of
Doctor in Applied Economic Sciences

This doctoral research was funded by ICM
(Interuniversitair Centrum voor Managementwetenschappen)

Vlerick Leuven Gent Management School

the Autonomous Management School of
Ghent University and Katholieke Universiteit Leuven

Competence Centre
Marketing

Response styles in consumer research

Bert Weijters

2006

Dissertation submitted to
the Faculty of Economics and Business Administration,
Ghent University,
in fulfillment of the requirements for the degree of
Doctor in Applied Economic Sciences

This doctoral research was funded by ICM
(Interuniversitair Centrum voor Managementwetenschappen)

Supervisors:

Supervisor:
Prof Dr Maggie Geuens
(Universiteit Gent; Vlerick Leuven Gent Management School)

Co-supervisor:
Prof Dr Niels Schillewaert
(Vlerick Leuven Gent Management School)

Members of the doctoral jury:

Prof Dr Hans Baumgartner
(Smeal College of Business, The Pennsylvania State University)

Prof Dr Jaak Billiet
(Katholieke Universiteit Leuven)

Prof Dr Alain De Beuckelaer
(Radboud University Nijmegen)

Prof Dr Patrick De Pelsmacker
(Universiteit Antwerpen)

Prof Dr Eddy Omey
(Universiteit Gent)

Dean Prof Dr Roland Paemeleire
(Universiteit Gent)

Prof Dr Patrick Van Kenhove
(Universiteit Gent)

ACKNOWLEDGEMENTS

Writing a doctoral dissertation may be hazardous for your mental health. I realized that the day I started imagining scatter plots in fluffy cloud formations. But let me assure you: It is worth the risk. The process provides a unique learning experience, not only in academic terms. It also made me realize the importance of social support. I want to thank all of you who were available in some way and made this whole process possible. And to those who I do not mention by name in this acknowledgement, I hope you understand this is due to some transient error in the retrieval and reporting process (which obviously merits further research).

First of all, let me thank the person who to a large extent introduced me to the realm of academic research, conferences, the publication process, and the doctoral process itself: my supervisor Maggie Geuens. Maggie combines empathy and reliability with a great feel for theory and an eye for detail, a unique blend. Fast and thorough feedback guaranteed. Someone else who was closely involved in this endeavor is my co-supervisor Niels Schillewaert. Niels is a great sparring partner, always willing to put me with my feet back on the ground when at risk to loose touch with reality, but also willing to creatively think along new lines together. I am very grateful to both!

Patrick Van Kenhove, as a member of the supervising committee, provided thoughtful feedback, grounded in his in-depth knowledge of the area of marketing research. Alain De Beuckelaer also willingly shared his methodological expertise. It was great to get links, references and ideas that broadened my scope in such a way.

Further, I would like to thank the additional members of the reading committee Hans Baumgartner and Jaak Billiet for their interest and time. It is wonderful to get input from the experts in the field.

Doing this doctoral research was also made possible by the support of four organizations. My appreciation goes to ICM (Intercollegiate Centre for Management Sciences) for providing a three year scholarship. Competence Centre Marketing of the Vlerick Leuven Gent Management School provided its support throughout. I would like to express my gratitude to Maggie as the initiator of the doctoral process, Kristof De Wulf as the former CCM chairman who engaged in this commitment, and Steve Muylle as the current CCM chairman who continued it. I am also grateful to all CCM people (former and present colleagues) for the fun, the lunch and coffee breaks, and the social and professional backing in general. Also, I would like to thank the marketing department of the faculty of Applied Economics of Ghent University, and Patrick Van Kenhove as its chairman, for enabling this PhD process. Insites kindly provided much of the data for the studies reported in this dissertation. Thanks a lot, Niels, for making this possible.

Several people provided feedback to previous drafts of some of the studies reported here or concerning the process in general. I am grateful to Deva Rangarajan, Marion Debruyne, Koen Dewettink, my father and many others for their input.

Finally, much appreciation goes to my parents, in that they have always supported me in word and deed in all aspects of my professional and non-professional life. It’s good to always feel welcome. Let me conclude with Ilse and Luna, who have shared most of their married life or life as such (respectively) with a doctoral student. Thanks for the unwavering support, the laughs and the love.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ..... **I**

SAMENVATTING ..... **VII**

SUMMARY ..... **XI**

CHAPTER 1: INTRODUCTION **1**

CHAPTER OUTLINE 1

IMPORTANCE OF CONSUMER SURVEY RESEARCH 2

COMPONENTS OF ERROR 3

GENERAL OBJECTIVE 5

OUTLINE OF THE DISSERTATION 5

NOTE: HOW TO READ THIS DISSERTATION 7

NOTE: WHAT THIS DISSERTATION IS NOT ABOUT 7

CHAPTER 2: CONCEPTUAL BACKGROUND **9**

CHAPTER OUTLINE 9

A PROPOSED GENERAL FRAMEWORK 10

THE PSYCHOLOGY OF SURVEY RESPONSE 12

MEASUREMENT MODELS 13

RESPONSE STYLES 16

CHAPTER 3: MEASURES OF RESPONSE STYLES **27**

CHAPTER OUTLINE 27

INTRODUCTION 28

TWO SOURCES OF RESPONSE STYLES 29

BASIC FORMULAS FOR RESPONSE STYLE MEASURES 31

A TYPOLOGY OF RESPONSE STYLE OPERATIONALIZATIONS 33

CHAPTER 4: RESPONDENTS' UNDERSTANDING OF REVERSED ITEMS IN QUESTIONNAIRES: THE INTERACTION BETWEEN ITEM CONTENT AND ITEM LOCATION (EMPIRICAL STUDY 1) 45

CHAPTER OUTLINE 45

INTRODUCTION 46

CONCEPTUAL BACKGROUND AND HYPOTHESIS DEVELOPMENT 48

METHODOLOGY 56

RESULTS 60

DISCUSSION 64

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH 69

APPENDIX 4-1: STATISTICAL DISCUSSION OF THE PEARSON CORRELATION 73

APPENDIX 4-2: REPLICATION USING POLYCHORIC CORRELATION COEFFICIENTS 75

CHAPTER 5: THE SHORT TERM STABILITY OF RESPONSE STYLES (EMPIRICAL STUDY 2) 77

CHAPTER OUTLINE 77

INTRODUCTION 78

CONCEPTUAL FRAMEWORK 79

METHODOLOGY 85

RESULTS 90

DISCUSSION 104

CONCLUSION 109

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH 109

CHAPTER 6: THE LONG TERM STABILITY OF INDIVIDUAL RESPONSE STYLES (EMPIRICAL STUDY 3) 111

CHAPTER OUTLINE 111

INTRODUCTION 112

CONCEPTUAL FRAMEWORK 113

METHODOLOGY 120

ANALYSES AND RESULTS 123

Discussion 132

Limitations and suggestions for future research 137

CHAPTER 7: ASSESSING RESPONSE STYLES ACROSS MODES OF DATA COLLECTION (EMPIRICAL STUDY 4) 139

Chapter outline 139

Introduction 140

Theoretical framework 143

hypotheses 149

Empirical study 151

Part 1: Diagnosis of Cross-mode Differences in Response Styles 151

Discussion response style comparison 161

Part 2: Impact of response styles on a substantive construct 163

Conclusion 175

Limitations and directions for future research 176

CHAPTER 8: RESPONSE STYLES AS SATISFICING STRATEGIES (EMPIRICAL STUDY 5) 179

Chapter outline 179

Introduction 180

Conceptual background 181

Methodology 190

Results 195

Discussion 222

Limitations and directions for future research 225

Appendix 8-1: Further operational details of the DIFF and OPTIM measures 227

CHAPTER 9: CONCLUSION 231

Chapter outline 231

Recapitulation 232

implications 235

Impact of response styles 236

MEANING OF RESPONSE STYLES 239

THE CURE 241

LIMITATIONS AND FUTURE RESEARCH 246

APPENDIX A: DEFINITIONS AND ABBREVIATIONS 249

A-1 LIST OF ABBREVIATIONS 249

APPENDIX A-2: DEFINITIONS OF SOME KEY CONCEPTS 251

APPENDIX B: ITEMS 253

APPENDIX B – 1: ITEM SET 1 253

APPENDIX B – 2: ITEM SET 2 261

REFERENCES 269

SAMENVATTING

Vragenlijsten zijn een onmisbare bron van informatie voor onderzoekers die een beter inzicht willen verwerven in consumentengedrag. In consumentenbevragingen wordt vaak gebruik gemaakt van Likert items, een itemformaat waarin respondenten aanduiden in hoeverre ze akkoord gaan met bepaalde uitspraken. Antwoorden op zulke items kunnen echter vertekend zijn door responsstijlen, gedefinieerd als de tendens van bepaalde respondenten om onevenredig gebruik te maken van bepaalde responsopties. Een bekend voorbeeld is de instemmingstendens (d.i. de tendens om onevenredig gebruik te maken van de opties die instemming uitdrukken) maar respondenten kunnen eveneens onevenredig veel kiezen voor extreme opties, de middelpunt optie of de opties die staan voor niet-akkoord.

Ondanks herhaalde waarschuwingen voor de vertekenende effecten van responsstijlen, wordt in het meeste vragenlijstonderzoek niet gecontroleerd of gecorrigeerd voor hun impact. Mogelijke redenen hiervoor zijn de onvolmaakte theorievorming rond responsstijlen en hun antecedenten, en de moeilijkheden bij het meten van responsstijlen.

Het onderzoeksprogramma dat wordt gerapporteerd in deze dissertatie wil bijdragen aan een beter begrip van responsstijlen in consumentenonderzoek door de conceptualisering van responsstijlen verder vaste vorm te geven, door het optimaliseren van de meting van responsstijlen en door het verklaren van de processen die ten grondslag liggen aan responsstijlen. Hiertoe werden vijf empirische studies uitgevoerd.

Een eerste studie onderzocht de manier waarop respondenten omgekeerde items in een vragenlijst begrijpen. Omgekeerde items zijn gerelateerd aan hetzelfde construct

als hun niet-omgekeerde tegenhangers, maar in de tegengestelde richting (bv. ‘Ik hou ervan om nieuwe producten aan te kopen’ is een omkering van ‘ik koop niet graag innovaties’). Deze studie toonde aan dat antwoorden op items beïnvloed worden door de aanwezigheid van andere items die hetzelfde construct meten. De juiste functionele vorm van deze invloed verschilt tussen omgekeerde items en niet-omgekeerde items, hetgeen wijst op een verschil in de wijze waarop respondenten beide soorten items verwerken. Aangezien dit onderzoek de validiteit van omkeringen voor het meten van responsstijlen in vraag stelt, werd in de volgende studies een meetmethode voorgesteld van responsstijlen gebaseerd op antwoordpatronen die zich voordoen over toevalssteekproeven van items.

In de tweede studie werd aangetoond dat responsstijlen tendensen zijn die stabiel zijn over de loop van een enkele vragenlijstsessie. Studie 3 stelde vast dat responsstijlen grotendeels stabiele tendensen zijn over verschillende vragenlijsten heen die werden afgenomen met een jaar tussentijd en gebruik makend van verschillende itemreeksen. Een vierde onderzoek vergeleek responsstijlen tussen verschillende methodes van data-collectie, met name papieren vragenlijsten, telefoninterviews en online vragenlijsten. De studie toonde aan dat er tussen deze methodes verschillen kunnen optreden in responsstijlen die niet gedetecteerd kunnen worden met de traditionele toetsen voor meetinvariantie.

De laatste studie vond twee grote groepen van respondenten terug die verschillen in hun manier van *satisficing*, d.i. het besparen op tijd en energie die geïnvesteerd wordt in het beantwoorden van vragenlijsten. De ene groep heeft de neiging onevenredig veel gebruik te maken van de middelpunt responsoptie. De andere groep maakt daarentegen niet alleen onevenredig veel gebruik van de middelpunt responsoptie, maar ook van beide extremen.

Hoewel er nog vele vragen onbeantwoord blijven, draagt deze dissertatie bij tot een beter inzicht in responsstijlen. In het bijzonder werd de theorievorming verbeterd door (1) een verdere afbakening van de conceptualisering van responsstijlen, hetgeen werd vertaald in een voorgestelde meetmethode, (2) bewijs ter ondersteuning van de stabiliteit van responsstijlen, (3) de vaststelling dat responsstijlen een potentiële vertekenende factor zijn in vergelijkingen van verschillende methodes van data-collectie, en (4) een model dat de relatie tussen responsstijlen en *satisficing* expliciteert.

SUMMARY

In researchers' efforts to better understand consumers, questionnaires are an indispensable source of data. In consumer surveys the Likert item format, where respondents rate their agreement with specific statements, is very popular. However, responses to such items may be biased by response styles, defined as respondents' tendencies to disproportionately select specific response options. A well-known example is the acquiescence response style, i.e. the tendency to disproportionately use the response options expressing agreement, but respondents may also make disproportionate use of the extreme options, the midpoint option, or the options expressing disagreement.

Despite repeated warnings regarding the biasing effect of response styles, most survey research does not control or correct for their impact. A reason for this may be the incomplete understanding of response styles and their antecedents, as well as the difficulties encountered in measuring response styles.

The research programme reported in this dissertation aimed to contribute to the understanding of response styles in consumer research by further crystalizing the conceptualization of response styles, by optimizing measurement of response styles, and by explaining the processes that underly response styles. To this end, five empirical studies were carried out.

A first study investigated respondents' understanding of reversed items in questionnaires. Reversed items relate to the same construct as their non-reversed counterparts, but in the opposite direction (e.g. 'I love to buy new products' is a reversal of 'I dislike the purchase of innovations'). This study indicated that responses to items are influenced by the presence of other items that measure the same

construct. The exact functional form of this influence is different for reversals and non-reversals, indicating a difference in the way respondents process both types of items. Since this study questioned the validity of reversals for measuring response styles, in the subsequent studies a measurement method for response styles was proposed that captures response tendencies across random samples of items.

In a second study, it was shown that response styles are tendencies which are largely stable over the course of a single questionnaire administration. Study 3 established response styles as largely stable tendencies across different questionnaire administrations with a one year time gap in between and using different sets of questions.

A fourth study compared response styles across different modes of data collection (self-administered paper and pencil questionnaires, telephone interviews and self-administered online questionnaires). This study showed that there may be differences in response styles across modes of data collection that cannot be detected by the traditional measurement invariance tests.

A fifth and final study found two major segments of respondents that differ in the way they satisfice, i.e. economize on the time and effort invested in responding to questionnaire items. One group tends to disproportionately use the midpoint when satisficing. A second group, when satisficing, disproportionately uses the midpoint as well as the negative and positive extremes of the response scale.

In sum, though many questions remain unresolved, this dissertation contributes to a better understanding of response styles. More specifically, theory is enhanced by (1) a further delineation of the concept of response styles, which is translated in a proposed operationalization of response styles, (2) evidence in support of the stabililty of response styles, (3) the establishment of response styles as a potential biasing factor in

cross-mode comparisons, and (4) a model that captures the relation of response styles to satisficing.

CHAPTER 1: INTRODUCTION

CHAPTER OUTLINE

The topic of the current dissertation, response styles in consumer research, is introduced. It is demonstrated how self-report measures are indispensable for consumer research, but also that the validity of such measures is threatened by response styles. An outline of the dissertation is given.

IMPORTANCE OF CONSUMER SURVEY RESEARCH

In consumer research as in many other behavioral sciences, questionnaire data are an indispensable source of information. While it might be possible to make direct observations of what, when and how much consumers buy, one usually needs self-reports to understand why they do so and what they might prefer to do in the future. If large numbers of consumers need to be questioned regarding their beliefs and/or evaluations, closed-ended questions provide the most efficient solution (Converse 1984). Casual inspection of the major marketing journals provides ready evidence of the widespread use of closed ended self-report measures, most often based on Likert items, where respondents are asked to rate their level of agreement with a statement (Likert 1932)[1].

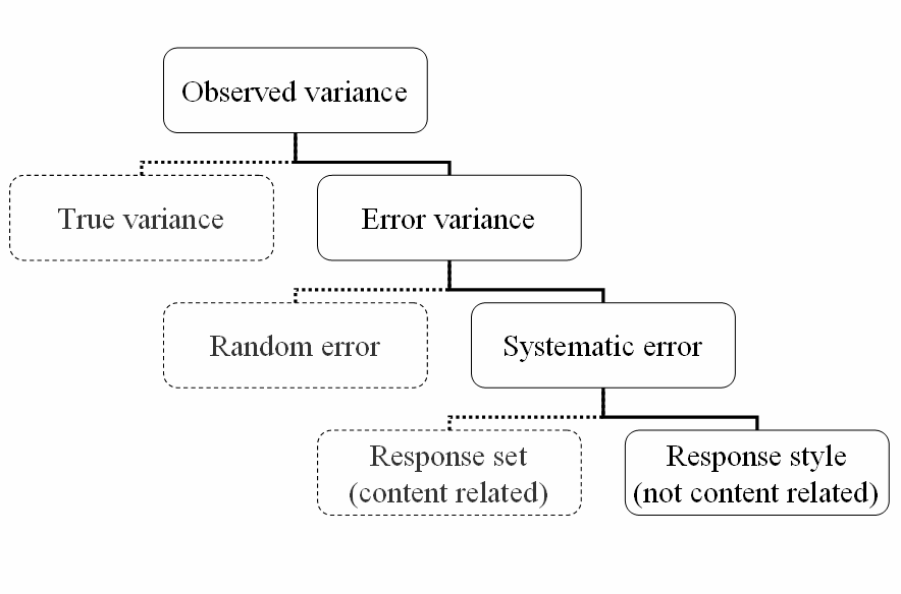
Within the field of consumer research a host of domains make ample use of Likert item measurement. These domains include customer satisfaction and loyalty (Mittal and Kamakura 2001), service evaluation (Parasuraman, Zeithaml and Malhotra 2005), attitudes (Ajzen 2001), personal values (Steenkamp, Ter Hofstede and Wedel 1999), affect and mood (Shiv and Fedorikhin 1999), consumer innovativeness (Steenkamp and Gielens 2003), other individual difference variables like technology readiness (Parasuraman 2000), and numerous others. In many of these domains, like service quality and consumer innovativeness, to name just two, it may even appear that most part of the research efforts reported in the literature are directed towards the development, validation and optimization of multi-item scales to measure constructs of interest.

[1] Some key terms related to questionnaire research are briefly defined and discussed in Appendix A-2.

COMPONENTS OF ERROR

Several threats to the procedure's validity have been identified though. A useful way to think of this is in terms of true and error variance, a central concept in classical test theory (Traub 1994)[2]. The components of an item's observed variance are graphically shown in Figure 1-1.

Figure 1-1
Decomposition of observed variance



Items are being designed to measure true variance (Traub 1994). Unfortunately, this aim is not fully met due to the presence of error variance. Error variance has two components, a random and a systematic component. The effect of random error has been generally accepted and is accounted for by using multi-item scales (Churchill 1979) and correcting for measurement error during data-analysis (Fornell and Larcker

[2] A more detailed conceptual and operational definition of response styles will be given later in this text, after a further elaboration of the conceptual framework. The current discussion mainly aims to offer a first intuitive frame of reference.

1981). The effect of systematic error, on the other hand, poses more serious problems to the validity of survey research because it provides an alternative explanation for the observed relationships between measures of different constructs (Podsakoff et al. 2003). While the methodological necessity of controlling/correcting for systematic error may often be acknowledged, it is commonly honored in the breach (as shown by Baumgartner and Steenkamp 2001; Podsakoff et al. 2003). Following the classic article by Rorer (1965), the systematic error component can be split up into content related systematic error due to response sets and non-content related systematic error due to response styles. A response set is related to content, and more specifically refers to the extent to which the respondents want to create an impression of themselves with regard to the item content. Social desirability is a well-known example of this (Leite and Beretvas 2005). A response style, on the contrary, is a tendency to answer items in a certain way regardless of content (Rorer 1965). The best known example of a response style probably is acquiescence response style, i.e. the tendency to agree with statements regardless of their content (Billiet and McClendon 2000). Contrary to social desirability, this response style is cognitively rather than socially based (Knowles and Nathan 1997; Ayidiya and McClendon 1990). Moreover, by definition response styles are not limited to specific content domains, such as socially sensitive variables or so-called ‘dark side variables’ (Mick 1996) and can therefore be expected to be omnipresent in survey research. The essence of the response style problem is that the same response can have different meanings for different respondents (Rossi, Gilula and Allenby 2001). In particular, individuals differ in their tendency to use certain types of responses: extreme, neutral, agree, or disagree (Stening and Everett 1984). Hence, to know the meaning of the

responses recorded in questionnaire based data, a correct understanding of response styles is indispensable. The current dissertation aims to add to this understanding.

GENERAL OBJECTIVE

The current dissertation wants to contribute to the knowledge of response styles by optimizing the conceptualization, operationalization and explanation of response styles. The insights related to these issues may be helpful in improving diagnosis and correction of response style bias.

In terms of conceptualization, in the current dissertation the response style phenomenon is integrated in a broader theoretical framework drawing from theories of survey response and contemporary measurement models. The operationalization of response styles is advanced through testing and evaluating alternative measurement methods of response styles (including the use of reversed items). A measurement method is proposed in a means and covariance structure context. Closely related to this, the explanation of response styles starts from an assessment of the short term stability (within a single questionnaire) and the long term stability (across two data collections separated by a time lag) of response styles. Additionally, the relation of response styles to demographics is confirmed and two major types of antecedents are established: mode of data collection (online, paper and pencil, telephone interview; see study 4), and satisficing (i.e. minimizing the investment of time and effort in the response process from the part of the respondent; Krosnick 1991; see study 5).

OUTLINE OF THE DISSERTATION

In the current dissertation, first a conceptual background is drawn. Building on this conceptual framework, five empirical studies are reported:

(1) "*Respondents' understanding of reversed items in questionnaires: The interaction between item content and item location*". This study investigates how respondents may change their interpretation of items depending on the item's proximity to other items that have the same meaning, the opposite meaning (i.e. reversed items), or no related meaning. Among others, it is found that reversed items correlate more strongly (negatively) the further they are apart. These results indicate that inconsistent responses to reversed items may be due to interpretational reasons rather than response style bias; a finding that clearly has repercussions on the question of how to measure response styles.

(2) "*The short term stability of response styles*" shows that the effect of response styles on items in a single questionnaire have a substantial stable component.

(3) "*The long term stability of individual response styles*" assesses the extent to which response styles of individuals are stable across two independent questionnaire administrations separated by a one-year time lag. Substantial stability is found.

(4) In "*Assessing response styles across modes of data collection*", a comparison is made between online, telephone and paper and pencil surveys in terms of the level of response style bias. The telephone mode is found to be rather different than the other two modes.

(5) "*Response styles as satisficing strategies*" investigates which response styles may be satisficing strategies, i.e. strategies used by the respondents to save time and cognitive effort. Two major segments of respondents are found, each using different satisficing strategies in terms of response styles.

Finally, the last chapter of the dissertation provides some concluding remarks, summing up the main findings and integrating them with one another and the

conceptual framework. The main limitations of the studies are discussed and related to opportunities for future research.

NOTE: HOW TO READ THIS DISSERTATION

All chapters can be read in isolation. Thus, the reader who is interested in a specific topic can directly go to the chapter in question without missing any information necessary for a correct understanding of the chapter. Obviously, this implies that some information will be repeated. The overlap is kept to the necessary minimum. Still, the most logical order of reading the chapters is in the order they are presented.
On a practical note, the abbreviations used in the text are always given in unabbreviated form at least once. As a backing option, all abbreviations and their referents are also listed in Appendix A-1. Additionally, to ensure a shared understanding of the words used in the text, some key concepts are defined in Appendix A-2. The bibliographic references of all chapters are grouped at the back of the current volume.

NOTE: WHAT THIS DISSERTATION IS NOT ABOUT

The research reported in this dissertation focuses on response styles in consumer self-reports using Likert-type agreement rating items related to non-factual unbounded information; unbounded refers to variables that have no absolute scale or zero-point. By definition, unbounded non-factual data have no directly observable counterpart. This is the case for attitudes and beliefs, for example, but not necessarily for probabilities, percentages, etc. The latter are therefore not studied in the current dissertation. Other topics not considered in detail in this dissertation include response sets, random error, sampling error, item non-response and unit non-response.

CHAPTER 2: CONCEPTUAL BACKGROUND

"*[...] survey responses, as we are so often reminded, are not merely self-reports of preexisting states and behaviors; they are behaviors themselves.*" (Schuman 1992, p. 20)

CHAPTER OUTLINE

Before describing the empirical studies that were conducted, a conceptual framework is set up based on the following elements: a general framework of how response styles may relate to latent and observed variables; a sketch of the process of survey response; theories on how respondents map beliefs and evaluations onto response scales; a model on how response styles may influence this mapping process; and a review of potential effects this may have on univariate and multivariate data.

A PROPOSED GENERAL FRAMEWORK

To understand response styles, it is good to start from a conceptualization of the survey response process. The following quote traces this process back to the essence: "When we talk about attitudes we are talking about constructs of the mind as they are expressed in response to our questions. But usually all we really know are the questions we ask and the answers we get." (Burleigh Gardner, 1978, as cited in Churchill 1979). Gardner brings to mind here the reality that pervades most of the behavioral sciences: what researchers observe are stimuli ("our questions") and the responses to these stimuli ("the answers we get"), while usually in measurement models the reponse is conceptualized as a direct effect of the construct of interest. This idea is represented graphically in Figure 2-1 a and b. Figure 2-1a depicts a measurement model as it is very commonly used in the context of confirmatory factor analysis: an individual i's response R_i is shown as the consequence of i's level of latent construct ξ_i.

Figure 2-1

Graphical representation of the applied model versus the presumed true model

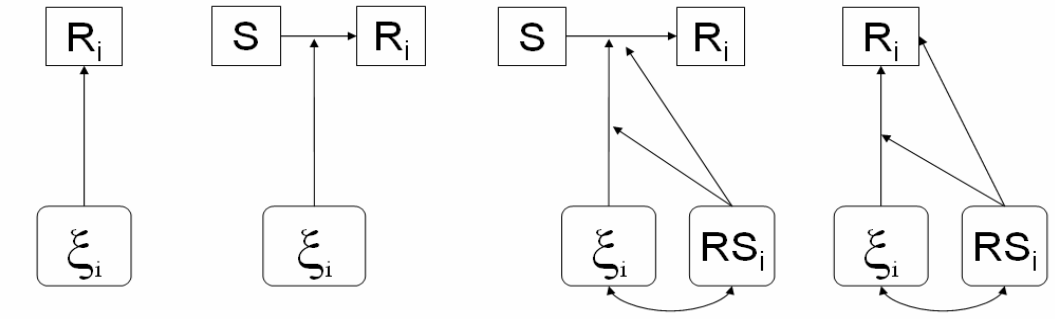


Figure 2-1a Figure 2-1b Figure 2-1c Figure 2-1d

Figure 2-1b shows the underlying causal model: stimulus S, which is constant across repondents, leads to response R_i (the subscript i indicates that R varies over individuals). The S-R relation is moderated by ξ_i. This means that for different levels of ξ, the relation between S and R is different. On the other hand, the model also implies that for given levels of ξ, the relation between R and S is identical across respondents. The latter assumption is challenged by the concept of response styles (RS), defined as tendencies to disproportionately select a particular subset of response options (Rorer 1965; O'Neill 1967), where disproportionate in the current text is interpreted as disproportionate for given levels of ξ. Formally,

$$E(R_i|\xi_0; RS_i) = E(R_j|\xi_0; RS_j) <=> RS_i = RS_j \qquad (1)$$

where ξ_0 corresponds to a given level of latent construct ξ, RS_i and RS_j are the response style levels of individual i and j, and R_i and R_j are responses to a valid indicator of ξ by the same respondents. When a response style (RS) is added to the causal model underlying a response, as shown in Figure 2-1c, it may have the following effects. First, like ξ, RS may moderate the S-R relation. Second, the moderating effect of ξ on the S-R link itself may be moderated by RS. As an aside, RS may or may not be related to ξ. Since S is kept constant, this model reduces to measurement model d in Figure 2-1, where RS is modeled to have a direct effect on R, as well as a moderating effect on the relation between ξ and R. An important question relates to the status of the latent construct ξ that is being measured, which may be a pre-existing state (as in Schuman's quote above), as well as a judgment that is constructed on the spot by the respondent. While this question has not received a definite answer in the literature and will not get one here either, it is relevant to consider the plausible possibility discussed below.

THE PSYCHOLOGY OF SURVEY RESPONSE

While several models have been proposed that capture the psychological process leading to a response, the model recently proposed by Tourangeau, Rips and Rasinski (2000) integrates much of the previous work and seems to be well accepted in the literature (Podsakoff et al. 2003). Tourangeau et al. (2000) state that the response process consists of four major components. It is not a necessity to go through all processes sequentially, and some respondents will skip particular processes, or will go back and forth between some of the processes. Also, respondents may choose to put more or less effort in each of them. The components are (1) comprehension, which requires respondents to attend to the questions and instructions, interpret the relevant terms in a question and decide on what information is being sought; (2) retrieval, referring to the process of 'looking up' (activating and bringing to mind) relevant information in memory; (3) judgment, where the information that was retrieved is evaluated and integrated into an overall judgment; and (4) response, consisting of an editing and a mapping process. Editing refers to respondents' evaluating their judgment before actually disclosing it, and adapting it if deemed desirable. Mapping refers to translating the judgment into the format required by the questionnaire context, for example a rating scale. The latter two processes are especially relevant in light of the current issue. It seems meaningful to conceptualize response styles as operating at the level of response mapping, while response sets operate at the level of response editing. Although Tourangeau et al. (2000) discuss editing as the last process in the most common sequence (remember that this sequence is optional though), it seems plausible that editing commonly occurs before mapping, especially in the context of Likert item measurement, for it is the revealed judgment that has the potential of being socially undesirable, not the selected response category as such.

An important point that should be made based on the above is that the response to a questionnaire item does not usually correspond to a pre-existing chunk of information that is reported. Rather, several chunks of information are retrieved, integrated, edited and mapped. If this process model is linked to response styles and how they were conceptualized above, it is not immediately clear what the latent construct ξ actually corresponds to. It appears that positing the existence of such latent construct may be somewhat of a simplification of the response process. On the other hand, given only a stimulus and a response it is impossible to determine all the processes that occur in between. Hence, the reduction of the source of an actively generated response to a one-dimensional construct is necessary for the specification of measurement models that are uniquely identified (i.e. that have unique parameter estimates). The response process can be slightly rephrased to more clearly identify what ξ may refer to as follows. External stimulus S, the question, leads to an internal representation of the same via the process of comprehension. The internal stimulus activates beliefs via the process of retrieval. These beliefs result in a private judgment (via the process of integration/judgment). The private judgment leads to an edited judgment by editing the former. And, finally, the edited judgment is mapped onto a response option and reported. In this framework, it is proposed that response sets such as social desirability response set operate at the level of editing. The edited judgment than corresponds to (the level of) the latent construct that will be measured, i.e. ξ. Response styles determine how this edited judgment will be mapped onto a specific response option.

MEASUREMENT MODELS

Since in the current conceptualization it is proposed that response styles operate at the level of the construct-response link, two contemporarily dominant measurement models are briefly introduced: Confirmatory factor Analysis (CFA) and Item

Response Theory (IRT). CFA models will also be used in specifying a measurement model for response style indicators (in study 2, 3, 4 and 5, corresponding to Chapter 5, 6, 7 and 8). The brief introduction to the IRT model provides useful background for a correct understanding of study 5. Below, the discussion of the CFA and IRT models draws from Meade and Lautenschlager (2004), Raju, Lafitte and Byrne (2002), and Reise, Widaman and Pugh (1993). The interested reader is referred to these texts for more details; the current discussion will be limited to specifying the model implied item-construct relations and investigating how this relation may be affected by response styles.

CONFIRMATORY FACTOR ANALYSIS (CFA)

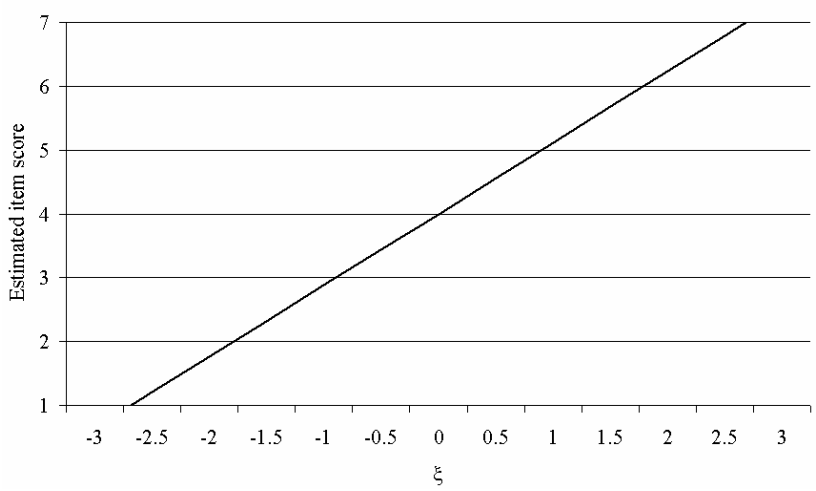
For a given individual and a given item, the CFA model can be mathematically described as follows:

$$x = \tau + \lambda\xi + \delta \qquad (2)$$

where x is the observed response to a specific item, τ is the intercept for the item, λ is the factor loading, ξ is the latent construct and δ the residual term. As is apparent from equation (2), the CFA model assumes linearity of the regression function of the response on the construct. An example of a regression plot of an observed item/indicator on its latent construct is given in Figure 2-2. The graph indicates that a respondent with a ξ level of zero (the sample mean), has an expected item score of 4, the midpoint.

Figure 2-2

Example of CFA item regression plot between construct and item



ITEM RESPONSE THEORY (IRT)

While the CFA model specifies a linear relation between the construct and the item, Item Response Theory (IRT) models specify the probability of the response categories of an item conditional on the construct level.

For the analysis of Likert items, the graded response model (Samejima 1969) has been shown to be most appropriate in general (Maydeu-Olivares 2005). The fundamental equation of this model is

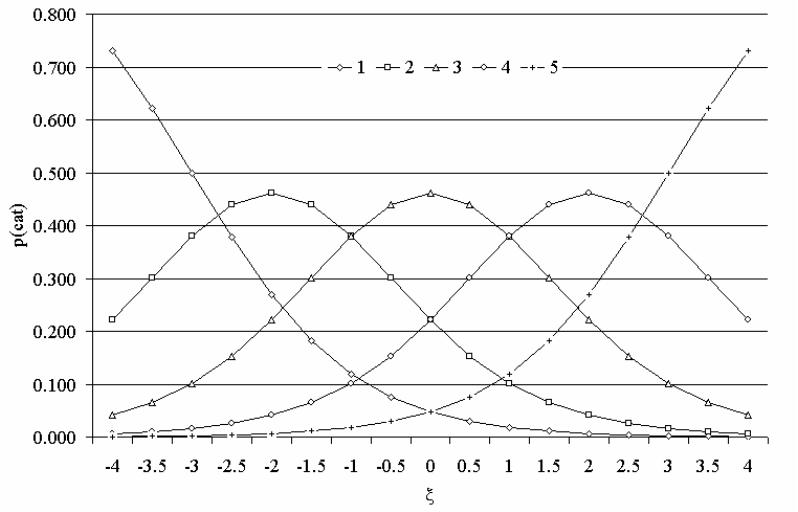
$$P(x=k|\xi) = 1/[1+\exp(-a(\xi - b_{j-1}))] - 1/[1+\exp(-a(\xi - b_j))]$$
$$= P^*(j-1) - P^*(j) \quad (3)$$

where P(x=k) refers to the probability of an individual responding in category k for variable x; this probability is modeled conditional on the level of latent construct ξ. The response categories are assumed to be separated by thresholds on the underlying ξ dimension corresponding to the b parameters. For each response category, an Item Characteristic Curve (ICC) is estimated which captures the probability of a specific category response as a function of ξ. a is the item discrimination parameter, and its

value is proportional to the slope of the Item Response Functions. An example of an ICC for a five point scale is given in Figure 2-3. Note that the ICC concerns a different item than the CFA example. To illustrate the interpretation, the ICC graph shows that individuals who have a ξ level between approximately -1 and 1 will most probably select response category 3, the midresponse.

Figure 2-3:

Example of IRT Item Characteristic Curve



Given these construct-item relations, it is now possible to more clearly delineate the potential effects of response styles on observed item responses. First, however, an overview is given of the response styles treated in this dissertation.

RESPONSE STYLES

As stated above, response styles relate to the probability that a respondent selects a specific subset of response categories (for a given level of the latent construct). Such subset may consist of the categories expressing agreement, disagreement, extreme positions at either side of the agreement scale, or neutrality (Stening and Everett

1984). The current dissertation focuses on the four corresponding response styles, summarized in Table 2-1.

TABLE 2-1

OVERVIEW OF RESPONSE STYLES TREATED IN THIS DISSERTATION

ARS	Acquiescence Response Style	Tendency to make disproportionate use of response categories at the favorable/agreement side of the agreement rating scale
DRS	Disacquiescence Response Style	Tendency to make disproportionate use of response categories at the unfavorable/disagreement side of the agreement rating scale
ERS	Extreme Response Style	Tendency to make disproportionate use of response categories at the extreme sides of the agreement rating scale
MRS	Midpoint Response Style	Tendency to make disproportionate use of the middle response category

The biasing effect of response styles operates at two levels (Baumgartner and Steenkamp 2001; Podsakoff et al. 2003). First, the univariate distributions of observed item scores are affected. Second, the multivariate relations between measures of constructs are affected. Each is discussed in turn. The univariate distribution bias is linked to the CFA and IRT models.

Indicator bias due to response styles

Response styles affect the item-construct relation. Cheung and Rensvold (2000) discuss the effect in a CFA context. In this linear model, two parameters, representing the slope and the intercept, are needed to capture the expected relation between item and construct. Hence, the impact of ARS and DRS reduces to the effect of NARS (Net Acquiescence Response Style; Baumgartner and Steenkamp 2001) on the intercept. In

particular, respondents with high (low) NARS have a higher (lower) intercept (Cheung and Rensvold 2000), as illustrated in Figure 2-4a. As a consequence, for equal levels of a latent construct, high (low) ARS respondents have higher (lower) observed scores.

The effect of ERS is rather subtle. Essentially, it can be conceived as an amplification factor in the mapping function of internal states/latent variables to reported responses (Van der Kloot, Kroonenberg and Bakker 1985). For measures of which the mean is not equal to the scale's midpoint, this may result in directional bias of observed scores (Baumgartner and Steenkamp 2001). In the more general case, ERS will lead to differences in the relation between latent variables and observed variables (Cheung & Rensvold 2000). Respondents with high (low) ERS levels will show a steeper (shallower) slope of the item-construct function line. This is illustrated in Figure 2-4b. Hence, for latent scores above (below) the intercept, ERS has an inflating (deflating) effect on observed scores. Note that this interaction effect is implicit in the model proposed by Greenleaf (1992a). Baumgartner and Steenkamp (2001) capture this effect in a parsimonious way (without directly estimating the interaction of latent score and ERS) by studying the interaction between ERS and the average deviation from the midpoint of a given scale.

Figure 2-4a: NARS effects on the CFA measurement model (*)

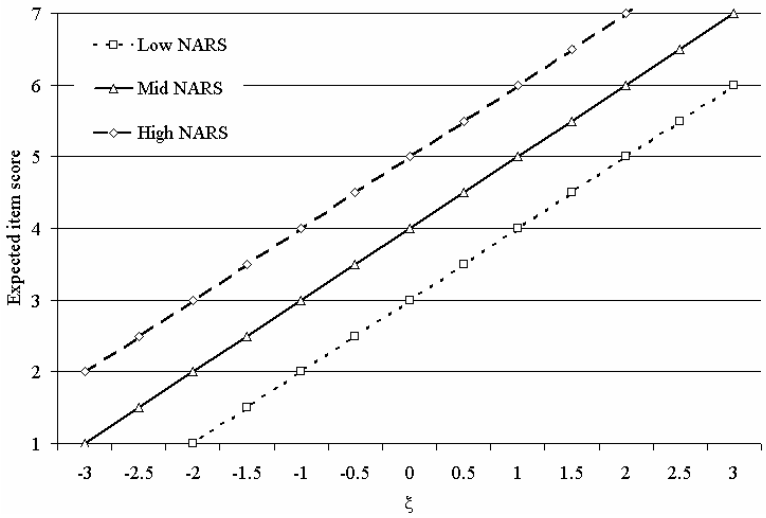
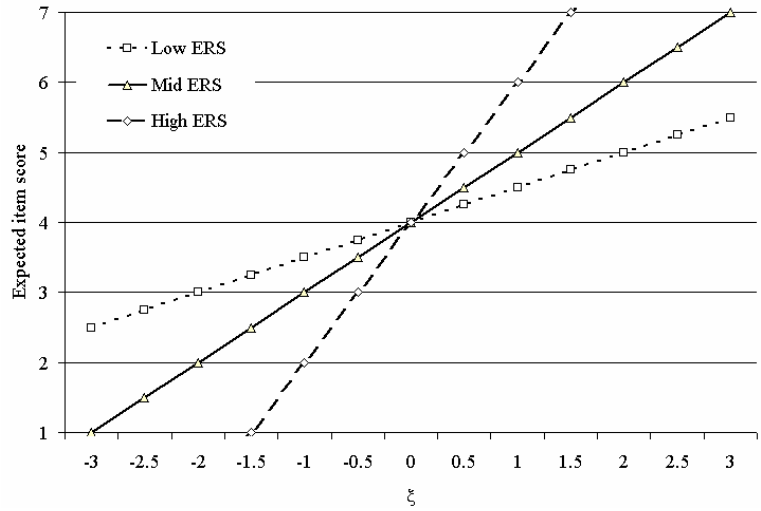


Figure 2-4b: ERS effects on the CFA measurement model (*)



(*) based on Cheung and Rensvold (2000)

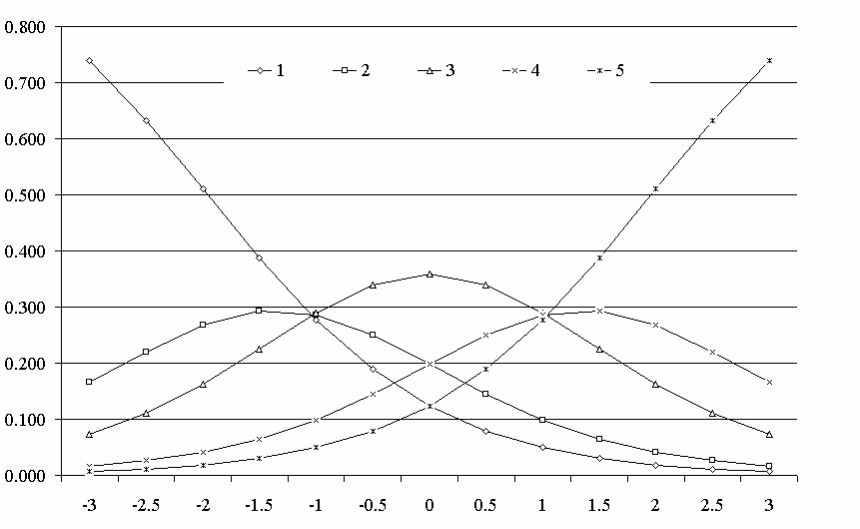
In the current dissertation the link between response styles and the CFA model is further elaborated in study 4 (Chapter 7), where measurement invariance and response style differences across modes of data collection are studied.

The relationship between IRT measurement models and response styles is discussed in study 5 (Chapter 8). There, results are provided that suggest that two major

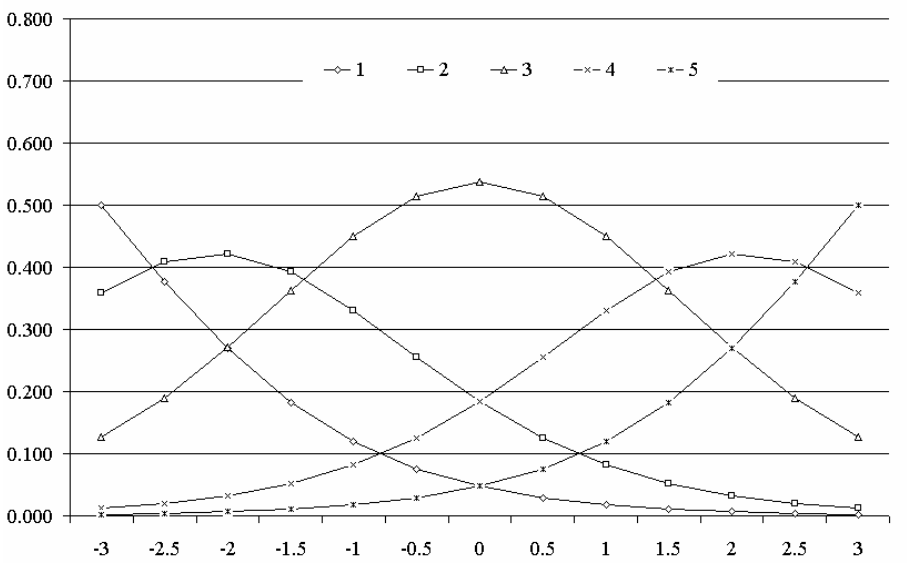
segments of respondents exist, each of which may have different threshold values in the linking function between constructs and items. If it is assumed that all respondents, irrespective of their response style levels, are drawn from the same underlying normal distribution of ξ, the operation of response styles can be captured by the threshold parameters (b in equation 3). For example, respondents with higher ARS levels may have lower threshold parameters. Respondents with higher ERS levels may have a higher threshold for the lowest category and a lower threshold for the higher category. This is illustrated with a simulated example in Figure 2-5 a. Respondents with higher MRS levels may have a lower left hand threshold for the midpoint combined with a higher right hand threshold for the same category. This effect is illustrated with a simulated example in Figure 2-5b.

Figure 2-5

Examples of Item Characteristic Curves



2-5 a. Example of ICC for high ERS respondents



2-5b. Example of ICC for high MRS respondents

Biasing effect of response styles on multivariate relations

Since response styles are a source of variance that is common across several measures, they lead to common variance that is not due to content. This phenomenon affects relations between measures of the same construct as well as relations between measures of different constructs. The biasing effects of response styles are graphically illustrated in Figure 2-6: for each model, the left hand pane shows what might be observed if an unspecified response style is not taken into account (labeled the apparent model) while it is present in reality; the right hand panel shows what would be observed if the response style would have been taken into account (labeled the true model). The misestimated relations are shown in dashed lines in the right hand panel. Residuals are not shown.

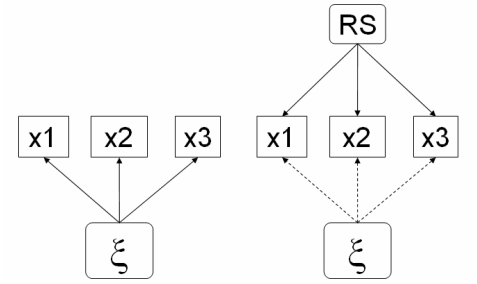
Response style bias of within-construct multivariate relations

Response style variance shared by indicators of a same construct may inflate the observed internal consistency of measures of this construct. As Mirowsky and Ross

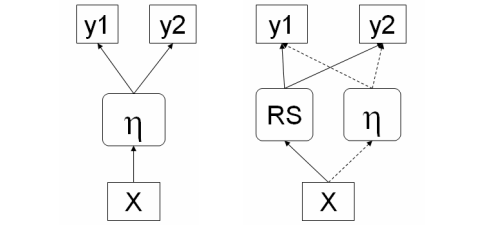
(1991) stated, "Other things being equal, the reliability of an unbiased measure is lower than that of a measure containing reproducible bias." In extreme cases, a content factor might be observed where none is present, as illustrated in Figure 2-6a. On the other hand, indicators that are scored in reversed directions may have artificially weak or even wrongly signed correlations (Bentler 1969).

Figure 2-6

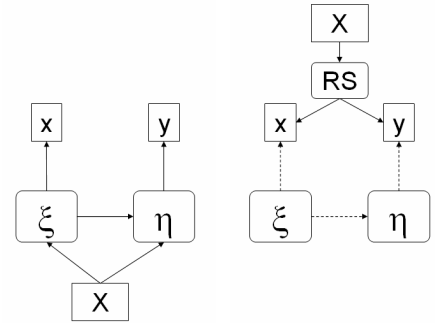
Potential biases due to response styles: apparent (left) versus true (right) models



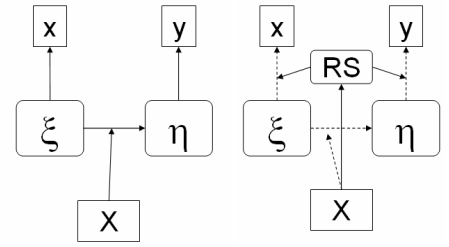
2-6a



2-6b



2-6c



2-6d

In Figure 2-6, RS stands for response style; x and y refer to the observed indicators of an independent and a dependent latent construct ξ and η respectively; X indicates an observed independent variable.

These observations are important for marketing research in light of the current focus of many marketing researchers on internal consistency. Domain sampling and classical test theory were set as the norm for marketing research by Churchill (1979).

The ‘Paradigm for developing better measures of marketing constructs’ he proposed, has undoubtedly improved the quality of measurement in marketing. In brief, domain sampling theory proposes that items used to measure a particular construct are mutually interchangeable and are sampled from a large population of items constituting the content domain. The domain sampling model makes the assumption that all items, if they belong to the same content domain, have an equal amount of common core (Churchill 1979, p. 68) and that the relations among the items are due to this common core. As the author puts it, “Interestingly, all of the errors that occur within a test can be easily encompassed by the domain sampling model. All of the sources of error occurring within a measurement will tend to lower the average correlation among the items within the test.” This is true only because error is defined as uncorrelated across items. Researchers may tend to forget this specific definition of error, often leading to overreliance on the coefficient of internal consistency alpha, while neglecting the deleterious effects of correlated error due to common sources of bias (Green and Hershberger 2000), of which response styles are a major component (Mirowsky and Ross 1991). Unfortunately, to the detriment of validity, it appears that many researchers have developed a single-minded focus on multi-item scales with high internal consistency, even if this consistency is achieved by selecting items that have a high chance of sharing common sources of bias (Rossiter 2002; Drolet and Morrison 2001; Green and Hershberger 2000; Mirowsky and Ross 1991).

Response style bias of between-construct multivariate relations

Just as common variance due to response styles may provide an alternative explanation for shared variance between indicators intended to measure the same construct, response style variance may also inflate or deflate relations between measures of different constructs (Podsakoff et al. 2003; Baumgartner and Steenkamp

2001). Some examples of such situations are given to make clear how diverse and widespread the influence of response styles potentially is. First, relationships between latent constructs and background variables (like demographics) may often be partly due to response style bias, as depicted in Figure 2-6b. For example, if no controls for acquiescence response style would be in place, measures of specific attitudes, e.g., distrust towards immigrants (η), might be artificially inflated among respondents with lower levels of education (X) due to higher ARS (Billiet and McClendon 2000). Study 4 of the current dissertation demonstrates how samples questioned by different modes of data collection (X) may also show artificial differences in their levels of trust in employees (η). Figure 2-6c shows another common scenario, where a latent construct is modeled as the consequence of another latent construct, each measured by observed indicators (in the figure only one indicator is shown for illustrative purposes) and controlled for a covariate. An example of this would be the situation where respondents are asked to rate several dimensions of service quality, which are then used as antecedents of an overall service evaluation. All indicators might be sharing substantial amounts of response style bias, which would lead to apparently good levels of explained variance. More subtle is the scenario in Figure 2-6d, where an apparent moderating effect of a background variable is actually due to a moderating effect of this background variable on the measurement relations, mediated by a response style. For example, since age generally is positively associated with ERS (Hamilton 1968), any relation between latent constructs that is found to be stronger among older respondents, should be interpreted with caution. A similar scenario is discussed in Study 5 (Chapter 8) of the current dissertation.

In sum, response styles have the potential to affect observed measures and the relations between them in many different ways. Moreover, it is not yet known with

great certainty when and where response styles operate (Baumgartner and Steenkamp 2001). Consequently, measurement of response styles is of key importance for the validity of survey based research. Measurement of response styles has two major goals: diagnosis and correction of bias. Obviously, the measures of response styles need to be valid themselves if used for diagnosis, because otherwise wrong research decisions may be taken. The validity requirements may even be greater if response style measures are used for corrective purposes, because correcting observed scores by means of invalid response style measures would increase the level of error rather than reduce it. The following section discusses different methods of measuring response styles that have been proposed in the literature and evaluates their merits and disadvantages.

CHAPTER 3: MEASURES OF RESPONSE STYLES

"Suppose no one asked a question, what would be the answer?"

(Gertrude Stein 1928)

CHAPTER OUTLINE

In this section, an overview is given of how response styles have been measured in the literature. Three aspects of the operationalization issue are discussed: the focus on stimulus or respondent; the basic formulas of response style measures; and the way content in indicators of response styles has been treated.

INTRODUCTION

When gathering information on consumers' evaluations and beliefs, researchers often have little alternative but to directly ask respondents what their evaluations and beliefs are. It is then hoped that respondents are willing to go through the process of understanding the question, retrieving the right information to subsequently form a judgment, and finally translate the judgment into the format specified by the researcher (Tourangeau, Rips and Rasinski 2000). In the case of non-factual information, the veracity of the obtained response can impossibly be ascertained through comparison with an objectively observable criterion. Or as Bohrnstedt (1983) put it, in the case of subjective phenomena, the concept of a Platonic true score does not apply, and the researcher is left with only the responses themselves to assess both the content and possible errors and biases in the same. What makes response styles problematic is that they provide an alternative explanation of why a respondent endorses a particular response option for a particular item. On the reverse side of the issue, the measurement of response styles is problematic because content provides an alternative explanation for the same observed behavior, namely endorsing a particular response option to a particular question (Hamilton 1968). Not very surprisingly, operationalization issues have traditionally be the Achilles' heel of response style research, as illustrated by Ray's statement (1979, p.639): "It is in fact a little odd that although we normally require reliability evidence for any scale score we use, acquiescence scores have been used in the past without such evidence." An evaluation of different methods of operationalizing response styles seems necessary.

A distinction can be made between different aspects of the operationalization issue. First, the way response styles are measured is influenced by the focus of the research design, which may be on the stimulus (questions, task design, interviewer effects, etc.)

or the respondent (personality and stable background variables, transient factors like fatigue, etc.). This distinction will be made first, indicating the inclination of the current research program to the respondent oriented individual differences approach. Second, the basic formulas used to distill response style measures from observed scores are briefly reviewed. Third, a typology is proposed of how researchers have treated the problem that items contain both content and style information (Jackson and Messick 1958) and how they have solved the question of which items to use as the basis for response style measures.

TWO SOURCES OF RESPONSE STYLES

Referring to the general framework set out above, two major sources of response styles are conceivable: first, the stimulus (mainly the question, but also other task related factors) to which a response is given; second, the respondent. Admittedly simplifying matters a bit, two traditions could be distinguished, each of which focuses somewhat more on either the stimulus side or the respondent side of the problem, respectively public opinion research and psychological research.

The so-called public opinion tradition, exemplified by the work of Schuman and Presser (1981) and research published in the Public Opinion Quarterly, can be said to be focused somewhat more on the stimulus-side. Of course, the moderating effect of respondent characteristics is studied in this tradition as well, albeit mostly from the perspective of demographic groups (Narayan and Krosnick 1996; Knauper 1999, Bachman and O'Malley 1984). Usually the main question relates to how to optimally design questions/task definitions, and the methodology typically involves split-ballot experiments, which essentially entail a focus on the between-stimuli aspect of the research design (e.g., Schuman and Presser 1981; Bishop 1987; Kalton, Roberts and Holt 1980; Hippler and Schwarz 1986; Shaeffer, Krosnick, Langer and Merkle 2005;

etc.). Questions have a somewhat different status in this tradition than they have in psychology, in that in public opinion survey research, the answer to the question is often important as such, considered in isolation. For example, it may be of interest whether the population is for or against a given topic and this position may be captured by a very small set of questions, possibly just one.

As discussed above, the other tradition, led by personality and social psychology (and followed by marketing), is strongly influenced by psychometrics and the domain sampling model (Churchill 1979, see above). In this tradition the latent construct is central, while the items are mutually interchangeable stimuli used to tap this construct. The respondent is the focus of attention here, and the typical methodology is to measure variables and correlate (and/or factor analyze) them across respondents (e.g. Bentler, Jackson and Messick 1971; Forehand 1962; Hamilton 1968). It is not surprising that the first major wave of studies on response styles in this field stressed personality correlates of response styles (Couch and Keniston 1960; Frederiksen and Messick 1959) or even nearly equated style to personality as a general orientation towards outside stimuli (McGee 1967; Gage, Leavitt and Stone 1957).

Needless to say, these two traditions are prototypical rather than being mutually exclusive and exhaustive, and much research is done on the interface between both, for example investigating the interactions or simultaneous effects of respondent and question characteristics in fields like sociology (Alwin and Krosnick 1991), statistics (McClendon 1991b) and education (Elliot 1961), where this was the main problem to begin with (Cronbach 1946; 1950).

Nevertheless, realizing the existence of both major perspectives may be helpful in better understanding the heterogeneity of the methods by which the response style measurement issue has been tackled. Much of the early response style literature is

reminiscent of the quote at the beginning of this chapter, wondering what the answer would be in the absence of a question. From this perspective the question is seen as a confounding factor that needs to be controlled for, leading response style researchers to make statements that might sound surprising if read out of context, like "*It seems almost impossible to escape the possibility that questionnaire items influence the responses given by respondents*" (Moxey and Sanford 1992, p. 295).

Also, insight in the major perspectives makes it easier to situate the view taken in the current dissertation. In particular, the current research program, with the exception of Study 1 (which compares responses to different stimuli in different conditions while making abstraction of respondents), relates most closely to the psychological paradigm, with a focus on between-subjects / individual difference variables. However, a special effort is made to study such differences as they apply across relevant sets of stimuli. This perspective is further clarified below. First, an overview is given of operationalizations of response styles in the tradition of response styles as individual difference variables. Whether these individual differences are stable or transient in nature is an empirical question that will be addressed by the appropriate means later in this text.

BASIC FORMULAS FOR RESPONSE STYLE MEASURES

From the perspective of response styles as individual difference variables, response styles are variables that need to be computed for each respondent. To this end, different formulas or computational methods have been proposed to extract the stylistic part from questionnaire responses. The items that are used as the basis for the methods will be discussed below. The general idea behind most of these techniques is largely similar however, and it has been noted that the particular formula used to compute response style measures may be rather inconsequential. Bachman and

O'Malley (1984) for example remark that different operationalizations of ARS and ERS led to similar conclusions. Similarly, Baumgartner and Steenkamp (2001) find convergent validity for different measures of ARS, DRS and ERS. Specifically for the latter style, these authors indicate that response range, though a theoretically distinct construct (Greenleaf 1992a, b), is empirically sufficiently similar to be used interchangeably with ERS.

The most common measures of ARS (DRS) use the frequency/proportion of (dis)agreements (e.g. Bachman and O'Malley 1984; Couch and Keniston 1960; Gage, Leavitt and Stone 1957; Peabody 1966), a weighted count of (dis)agreements, in which the strength of agreement is taken into account (Baumgartner and Steenkamp 2001; Jordan, Marcus and Reeder 1980), a count of double agreements to reversed items (Johnson et al. 2005; Baumgartner and Steenkamp 2001) or a factor on which all items load positively (negatively) (Bentler, Jackson and Messick 1971; Billiet and McClendon 2000). Note that these options apply a different weighting scheme to the same information and will correlate highly by design[3]. For this reason, correlating an ARS factor (on which reversed and non-reversed items load positively) with a count of agreements (Billiet and McClendon 2000) confirms the theoretically expected convergent validity of both measures without necessarily supporting criterion validity. As pointed out by Baumgartner and Steenkamp (2001) the difference between ARS and DRS may be used as an indicator of Net Acquiescence Response Style (NARS; e.g. Greenleaf 1992a), but it is theoretically relevant to treat ARS and DRS distinctly (Couch and Keniston 1960; Baumgartner and Steenkamp 2001).

[3] It was analytically demonstrated in a general context by Peabody (1962) that the weighting of the extremeness of Likert responses did affect overall scores only slightly. Peabody's argument directly applies to response style measures just as well.

As a measure of ERS it is common to use the frequency/proportion of extreme responses, for example one and five in a five point rating scale (e.g. Arthur and Freemantle 1966; Bachman and O'Malley 1984; Baumgartner and Steenkamp 2001; Greenleaf 1992b; Hui and Triandis 1985), though other methods have been used as well (see Hamilton 1968). As noted above, findings by Baumgartner and Steenkamp (2001) indicate the empirical convergence of ERS and response range.

In some studies, ERS and MRS have been treated as opposites of a same dimension (e.g. Jordan, Marcus and Reeder 1980). However, while ERS and MRS are negatively correlated in general, this does not always need to be the case (Osgood 1941; Stening and Everett 1984).

Self-evidently, Midpoint Response Style (MRS) is only relevant in case odd numbers of response categories are offered: MRS is usually measured as the frequency/proportion of midpoint responses (e.g. Kraut, Wolfson and Rothenberg 1975; Baumgartner and Steenkamp 2001; Stening and Everett 1984).

A TYPOLOGY OF RESPONSE STYLE OPERATIONALIZATIONS

A review of the literature suggests that operationalizations of response styles could be organized along two dimensions. First, the status of the items on which the response style measures are based (A) can be multifunctional, meaning that the items are used in both a substantive model and as response style measures or (B) they can be specific to the response style measure and hence not substantively relevant. A second dimension relates to the treatment of content in the items used for response style measurement. This dimension has four levels: (1) no specific controls are put in place ex ante (i.e. the content of the items is not deliberately manipulated or selected before data collection), and the items that happen to be available are used as the basis for computing response style measures post hoc; (2) content can be eliminated with the

aim of measuring style in the absence of content; (3) content can be manipulated to take on specific known levels (e.g. opposite meanings) that can be used to cancel out the influence of content by means of specific computations or modeling techniques; (4) content can be randomized, such that it has no systematic influence on responses.

A method that has been used but is not considered here uses an external criterion variable to assess the true value of the response to a questionnaire item (Greenleaf 1992a). Since the focus of the current research is on non-factual measures that have no observable true counterpart (see above), such methods lie beyond the scope of the current research. The use of behavioral measures as criterion variable for attitude measures may not be valid (Welkenhuysen-Gybels, Billiet and Cambré 2003), especially since the attitude-behavior relation has many moderators other than response styles (De Cannière 2006). These variables might include variables that may correlate with response styles and/or their antecedents.

Table 3-1 summarizes the levels of both dimensions of the proposed typology. Each of the cells in this matrix is discussed in turn.

TABLE 3-1

OVERVIEW OF POSSIBLE OPERATIONALIZATIONS OF RESPONSE STYLES

		I. Function of item set used for response style measures	
		A. Multi-functional items	B. Response Style measure specific items
II. Treatment of content	1. No ex ante control for content	A1	B1
	2. Elimination of content	(A2)	B2
	3. Experimental control	A3	B3
	4. Randomization	(A4)	B4

Cell labels between brackets indicate a combination that is not theoretically meaningful.

A1. NO SPECIFIC ITEMS, NO EX ANTE CONTROL FOR CONTENT

In some instances, researchers compute or model a response style measure based on items that are simultaneously used in a substantive model of interest, consisting of related constructs. In one such scenario, researchers might simultaneously use responses to a series of items on the one hand as content indicators (e.g. of personality or customer satisfaction) and as the basis for response style measures on the other hand (e.g. Couch and Keniston 1960; Rossi, Gilula and Allenby 2001). Such approach may lead to confounding of style and content (Arce-Ferrer 2006), and for this reason has been forcefully condemned (Rorer 1965). A somewhat related practice has been used in structural equation models where a common method factor has been created by loading the same items on both substantive and a method factor. This approach has also been severely criticized (Lindell and Whitney 2001). The main advantage of the approach is that no additional response style measures have to be included in the

questionnaire. Consequently, the procedure could also be used to carry out secondary analyses of data that were collected without taking into account response styles. The problems with this method clearly outweigh this advantage. First, such method factor is very general and does not distinguish between the effects of different response styles. Associated with this is the problem that the conceptual meaning of such factor may be vague. Consequently, it cannot be identified as a specific response style. Additionally, if two items are correlated and load on both a substantive and a method factor, the estimates may become somewhat more unstable, in that the estimation algorithm has more possibilities of accounting for given covariances with the same amount of data (resulting in less degrees of freedom and less power). The common method factor might therefore 'absorb' common variance that is not due to method bias (Podsakoff et al. 2003). For this reason, it has been argued that partialling out a general method factor that has no own indicators may produce virtually meaningless results (Lindell and Whitney 2001).

In sum, the basic problem of these methods is that it is hard (if not impossible) to correctly assign portions of covariance to method/response style factors and substance/content factors. Therefore, it seems recommendable to avoid this approach.

A special case where no response style measure specific items are included while content is related, occurs when using measurement invariance tests to assess response styles (as proposed by Cheung and Rensvold 2000 and criticized by Little 2000). Some of the above problems apply to this procedure (see Study 4; Chapter 7). Further, invariance is not a guarantee for the absence of response styles (Little 2000). Study 4 makes a more thorough evaluation of the relation between response styles and measurement invariance tests.

A2. NO SPECIFIC ITEMS, NO CONTENT

This combination is not possible since content-free items cannot be used in substantive models.

A3. NO SPECIFIC ITEMS, EXPERIMENTAL CONTROL FOR CONTENT

When using the same indicators to measure both content and style, a method to disentangle these two dimensions is by manipulating question form independently of content. This is the essential idea behind two methods. The first method, the Multi-Trait Multi-Method approach, manipulates form while keeping constant the measured content. The second method, using balanced items method factor, uses items with opposite meanings. Both methods are discussed in turn.

First, in the Multi-Trait Multi-Method (MTMM) approach the same construct (trait) is measured repeatedly by means of different measures/methods (Campbell and Fiske 1959). The idea is that observed variance can be decomposed in variance due to the trait that is being measured and variance due to the method used to measure it. To disentangle both sources, the classic MTMM design uses three measures of three traits, resulting in a total of nine measures. In the initial MTMM approach, the resulting correlations are put in a matrix. Nowadays, it is common to use Structural Equation Modeling in analyzing MTMM data (Coenders and Saris 2000; Saris, Satorra and Coenders 2004). In such model, each set of three measures measuring the same trait is then modeled to load on the same trait factor, while each set of three measures using the same method is modeled to load on the same method factor. Such model allows a researcher to assess the relative impact of content (validity) versus method (bias).

As Podsakoff et al. (2003) point out in their review of method biases and related remedies, MTMM models may encounter serious problems of identification and

specification. The identification problem has been countered by estimating MTMM data with models that have correlated uniqueness terms (Saris and Aalberts 2003). Such solution does not allow for the estimation of a method factor though.

Another limitation of MTMM designs in general is the requirement that the same respondent answers the same question repeatedly in a different form (method). This might lead to consistency bias, memory effects and/or fatigue effects (Saris, Satorra and Coenders 2004). While these disadvantages are limiting the applicability of the MTMM approach, a continuous stream of research is providing solutions to most of them, though it remains hard to counter all potential problems simultaneously in a single design (Coenders and Saris 2000; Saris, Satorra and Coenders 2004). From the perspective of response style research, the main limitation is that MTMM only has one method factor, which essentially captures directional bias (NARS) specific to each method, while the influence of MRS and ERS is not accounted for. In other words, no complete set of response style measures can be estimated in the MTMM design.

A second method that capitalizes on the manipulation of content independent of form is the balanced scale method (Billiet and McClendon 2000; Mirowsky and Ross 1991). This method can be used to model the factor structure and construct relations of scales that are balanced (i.e. made up of equal proportions of reversed and non-reversed items). Reversed and non-reversed items respectively have negative and positive loadings on the content factor they relate to, and all have positive loadings on an ARS factor (Billiet and McClendon 2000). This procedure is elegant in its efficiency, since no specific RS measures are needed. However, its use is limited to the operationalization of ARS in balanced scales. Other response styles cannot be accounted for using this method, and while balancing scales has been recommended,

not all commonly used scales in the literature have reversed items (Baumgartner and Steenkamp 2001). Part of the reason may be that the formulation of reversals is very difficult (Ray 1983; Billiet and McClendon 2000). The fundamental issue in this regard is that it may be impossible to independently manipulate content and form. Moreover, in many cases it makes sense for respondents to agree to both an item and its proposed reversal (Rorer 1965). Another common observation is that items and their reversals are too extreme, thus 'creating a middle ground' that allows respondents to disagree with both (Schuman and Presser 1981; McClendon 1991a). Moreover, differences in the way reversals are responded to have been shown to be due to interpretational factors rather than due to ARS (Wong, Rindfleisch and Burroughs 2003). The latter issue is studied in-depth in Study 1.

A4. NO SPECIFIC ITEMS, RANDOMIZATION OF CONTENT

Since items used to operationalize a substantive model are selected for their specific and related content, it is impossible to have random content across such items.

Next to methods that compute response style measures based on the items that are also used in a substantive model of interest, in some studies specific items have been used only to compute/model response styles. These methods are discussed now.

B1. SPECIFIC ITEMS, NO EX ANTE CONTROL FOR CONTENT

A first method that uses specific items is to measure response styles as they operate within a set of contentwise related items, i.e. without ex ante controls for content. The reason why this approach is so prevalent probably is that it can be used to analyze secondary data. Often, researchers decide post hoc to study the presence and extent of bias due to response styles in data that they obtained for other purposes (e.g. Bachman

and O'Malley 1984; Jordan, Marcus and Reeder 1980; Kiesler and Sproul 1986; Shulman 1973; Van Herk, Poortinga and Verhallen 2004). In other cases, researchers have used existing scales and computed response style indicators based on the items in these scales (e.g. Bentler, Jackson and Messick 1971; Gage, Leavitt and Stone 1957). In such context, it may be impossible to construct sets of items that are not contentwise related (Rorer 1965). The main advantage of this approach is its general applicability and the chance it offers to assess the extent to which response styles operate in specific studies in hindsight. The main disadvantage is that internal validity is low. More specifically, it is nearly impossible to disentangle variation in responses due to content and variation in responses due to response styles. Even if this would be achieved, the observed response styles might be content specific (Rorer 1965).

B2. SPECIFIC ITEMS, ELIMINATION OF CONTENT

Some researchers have tried to create content free items to try and eliminate or minimize content, such that responses could be attributed purely or mainly to response styles. Husek (1961) created a content free measure of ARS. The ESP acquiescence test "involves giving agree-disagree answer alternatives to a set of subjects and asking them to read the experimenter's mind and answer questions he is purportedly thinking of. However, the experimenter is not thinking of items, but merely counting from 1 to 10 over and over again" (Husek 1961). Similarly, Forehand (1962) used a phony language exam composed of items "whose content appears to be meaningful but is not."

A first problem with this approach is that together with content it eliminates external validity. That is, it is doubtful that response styles to content-free stimuli generalize to common measures of attitudes and other psychological variables. Second, although the aim of the content-free approach is to optimize internal validity, there is reason to

question its success in doing so. More specifically, it is plausible that the absence of content results in a qualitative shift in the process under study: freeing stimuli/questions of content also frees the response options of meaning. Hence, responses to such stimuli are merely gambles/guesses or random number generation tasks. Consequently, it could be argued that what is studied in such case are guessing/gambling styles, not response styles. Using items with so-called low content saturation (Hamilton 1968) suffers from the same limitations as the content-free stimuli. As Block (1971) phrased it rather eloquently: "This design decision astonishes me for it suggests that in order to 'find' acquiescence, one must look for it under artificially constrained and irrelevant circumstances rather than in typical inventory domains where acquiescence was first sighted. I am reminded of the drunk who, having lost his wallet in a dark alley, proceeded to look for it under a convenient street light rather than in the place where the wallet should be found."

B3. SPECIFIC ITEMS, CONTROL FOR CONTENT

To eliminate the effect of content on response style measures, some researchers have used specific sets of items containing particular items and their reversals (Baumgartner and Steenkamp 2001 for their ARS2 and DARS2 measures; Watson 1992). The same remarks apply as those listed for method A3. As mentioned there, the issue of reverse coded items is discussed in Study 1, which is dedicated specifically to this issue.

B4. SPECIFIC ITEMS, RANDOMIZATION OF CONTENT

A final method to measure response styles makes use of a set of items that is maximally heterogeneous in content. Such approach is advocated by Greenleaf (1992a, b). It could be said that the basic idea behind this approach is to reduce the

effect of content in the set of items to random noise: if all the items represent different constructs that are (on average) unrelated, it can be expected that there is no consistency in the responses other than that induced by response styles. Greenleaf (1992 a, b) - for all measures - and Baumgartner and Steenkamp (2001) - for most measures - use a convenience sample of items that are quite representative of items used in consumer research, since they are taken from a typical consumer survey. One step further, the use of a random sample of items taken from a relevant sampling frame (e.g., an inventory of multi-item scales) would even further optimize both internal validity and external validity: internal validity because the relation or similarity of an individual's responses to widely heterogeneous items is mainly due to response styles, not content; external validity because operation of the response style can be expected to generalize to the population of items from which the random sample was drawn. In the studies presented in this volume, such samples of items are used.

Finally, an issue that merits some further attention relates to the number of indicators used to measure response styles. In principle, a single indicator based on one set of items can be used for each response style, a method applied in two relatively recent and influential response style studies (Greenleaf 1992a, b; Watson 1992). The use of multiple indicators has the potential benefit that measurement error in the response style measures can be accounted for by modeling the response styles in a Structural Equation Model. Though not capitalizing on this possibility, Baumgartner and Steenkamp (2001) use indicators based on different measurement methods. Unfortunately, the use of different measurement methods is not possible for all response styles, MRS in particular. As mentioned above, the method by Billiet and McClendon (2000) models ARS as a latent variable, but is limited to ARS. In the

current dissertation, ARS, DRS, ERS and MRS are all modeled as latent variables with multiple indicators based on random subsets of the marker items. Splitting the total item set into 'testlets' is preferable to grouping all information in one indicator, because it allows for measurement error in the response style indicators (Podsakoff et al. 2003). As detailed in Study 2, 3 and 4 (Chapter 5, 6 and 7), this is particularly important since different response style indicators are based on the same sets of items, leading to correlations that are indicator specific rather than structural. As discussed in Study 4 (Chapter 7), an additional advantage of this approach is that the response style factors can be subjected to measurement invariance tests.

IMPLICATIONS FOR THE CURRENT RESEARCH PROGRAM

Based on the literature review the following decisions were made regarding the current research program. Two recent response style studies have been particularly important with regard to the question of how to measure response styles (Billiet and McClendon 2000; Baumgartner and Steenkamp 2001). The method by Billiet and McClendon (2000) focuses on the relation between ARS and the response to reversed items. Since recent research has suggested that this relation may be more complicated than initially hoped (Wong, Rindfleisch and Burroughs 2003), Study 1 (Chapter 4) of the current dissertation investigates this relationship further. The other empirical studies focus not alone on ARS, but also on DRS, ERS and MRS. For this reason, an operationalization is used that enables the study of all these response styles. Extending the approach advocated by Baumgartner and Steenkamp (2001) and Greenleaf (1992a, b), response style measures in these studies are based on representative samples of consumer research items (listed in Appendix B). Moreover, in line with recommendations by Podsakoff et al. (2003), multiple subsets of items will be created to take into account measurement error in the response style measurement models.

CHAPTER 4: RESPONDENTS' UNDERSTANDING OF REVERSED ITEMS IN QUESTIONNAIRES: THE INTERACTION BETWEEN ITEM CONTENT AND ITEM LOCATION (EMPIRICAL STUDY 1)

CHAPTER OUTLINE

As is apparent from the contradictory recommendations by measurement experts, the issue of whether or not to use reversed items is far from resolved, mostly because too little is known about how consumers respond to reversed items. This study investigated the response to reversed items as a function of their distance to their non-reversed counterparts. Over three thousand respondents filled out an online questionnaire containing a heterogeneous sample of seventy-six items. Regression analyses on the observed correlations between contentwise unrelated, positively related and negatively related items revealed that the correlation between two nearby positively related items decreased with increasing inter-item distance, while the absolute correlations between negatively related items increased with increasing inter-item distance. The latter finding lends support to the Unipolar rather than the Bipolar Response Model.

INTRODUCTION

In consumer research, measurement depends heavily on the use of self-report scales of different forms, often Likert scales. At the time he introduced his popular scale, Likert (1932, p. 46) already recommended the use of reversals. A reversed item i' is assumed to relate to the same latent variable as its non-reversed counterpart i, but in a negative instead of a positive way. To illustrate with two items taken from the Mavenism scale by Steenkamp and Gielens (2003), i could be "I don't talk to friends about the products that I buy", and i' could be "I like introducing new brands and products to my friends", but the order of the items and the labels could as well be inverted[4].

An advantage of balancing a scale, i.e. mixing equal amounts of reversed and non-reversed items, is that is may correct summed or averaged scale scores for the influence of Acquiescence Response Style (ARS), i.e. yeah-saying (Paulhus 1991). Several researchers have reported that ARS biased results (Baumgartner and Steenkamp 2001; Bentler 1969; Billiet and McClendon 2000; Paulhus 1991). ARS is assumed to lead respondents to agree to items regardless of content, even if one item is the reversal of the other (Ray 1983). The presumed mechanism behind the

[4] Since the scaling of a latent construct is essentially arbitrary, it seems most appropriate to consider the attribute of being reversed as a characteristic of an item-item pair rather than an item-construct pair (McPherson and Mohr 2005). As McPherson and Mohr (2005) put it: "*[...] the keying direction of an item is entirely relative to the definition of the construct of interest: For example, positively keyed items from a depression scale may resemble negatively keyed items from a happiness scale.*" Hence, the processes that we will investigate in the current study cannot be attributed to characteristics of negatively worded items (in isolation or because they are part of a dominantly positively keyed scale), such as negations or other semantic attributes (as is the case in studies by Cordery and Sevastos 1993; Schmitt and Stults 1985; and Schriesheim, Eisenbach and Hill 1991).

acquiescence response style correction by using reversals can be described as follows. Assume that the observed score X_i on item i can be decomposed $X_i = T + ARS + R_i$, where T is the so-called true score, ARS refers to systematic error due to acquiescence response style, and R_i refers to random error, which has an expected mean of zero and is orthogonal to random error components of other items as well as T (Churchill 1979). The reversal of the item, labeled i', then has as an observed score $X_{i'} = -T + ARS + R_{i'}$ (Mirowsky and Ross 1991). The expected weighted sum or difference of R_i and $R_{i'}$ is zero (Andrews 1984). The expected weighted sum of X_i and $X_{i'}$ will yield ½ $(X_i - X_{i'}) = ½(T + T + ARS – ARS + R_i - R_{i'}) = T$. For the reversal to have the desired effect (i.e. correct for ARS), some conditions have to be met. First, the effect of acquiescence response style should be constant for i and i' (Billiet and McClendon 2000). Second, the shared variance between the items should result only from the latent causes they have in common, i.e. T and ARS. Included in this condition is that measurement of i should not influence measurement of i' directly, that is, the items should not interact (Tuerlinckx and De Boeck 2001). If the latter condition, which is labeled non-reactivity, is not met, it would be incorrect to attribute the covariance between i and i' to content and acquiescence response style alone. Such faulty attribution would result in biased estimates of the relationship between the items and their underlying construct (Tuerlinckx and De Boeck 2001).

The objective of the current paper was twofold. A first objective was to investigate how and to what extent ARS influences inter-item correlations. A second objective was to test the assumption of non-reactivity. Specifically, it was investigated whether responses to i' were biased as a function of the presence and proximity of i. To this end, inter-item correlations of unrelated items, same-direction items (i.e. items that are related to the same construct in the same direction) and reversals were studied.

CONCEPTUAL BACKGROUND AND HYPOTHESIS DEVELOPMENT

REVERSED ITEMS AND THE ITEM-FACTOR RELATION

Researchers who studied reversals have repeatedly pointed out the near impossibility of formulating such items. Schuman and Presser, for example, dismissed 12 out of 14 items for analyses based on their presumed non-validity as reversals (Schuman and Presser 1981; Appendix D, p. 345-348). Likewise, Ray (1983, p. 83) listed some scales that are deemed to be nearly 'irreversible'. McClendon (1991a, p. 69) discussed such concerns in detail and stated there is a consensus on two criteria for valid reversals: "*First, and most obviously, the reversal must change the direction of the content, that is, it must be a logical reversal. And second, the reversal should not be too extreme, that is, it should not be a polar opposite."* While these criteria are valid conditions for defining 'perfect' reversals, in a measurement context the objective usually is not to have a perfect logical and symmetrical reversal, but to have items that have approximately equally strong relations (usually factor loadings) to the same construct ξ, albeit in the opposite direction. Reverse items often are not logical opposites, but neither are most same-direction items logical equivalents. If they were, they would be considered essentially identical and hence redundant (Churchill 1979; Rossiter 2002). Research that studies reversed items in real measurement scales consequently uses imperfect reversals, i.e. items that are negatively correlated but not strict logical opposites (e.g. Billiet and McClendon 2000; Motl and DiStefano 2002; Wong, Rindfleisch and Burroughs 2003).

In addition to being hard to design, reversed items are also hard to analyze. It is well-known to researchers who have used reversals that these items tend to load on a different factor than the non-reversals or an additional method factor (Bentler 1969; Herche and Engelland 1996; Marsh 1996; Motl and DiStefano 2002; Quilty, Oakman

and Risko 2006), and/or that the use of reversals leads to data-model fit problems in confirmatory factor analyses (Cordery and Sevastos 1993). These problems are so pervasive and bothersome, that some oppose the use of reversals (Barnette 2000; Marsh 1996; Schmitt and Stults 1985; Schriesheim and Hill 1981; Schriesheim, Eisenbach and Hill 1991). Based on the finding that balancing scales reduces at least part of the ARS bias, others remain in favor (Baumgartner and Steenkamp 2001; Paulhus 1991). In order to sort out this debate, it is necessary to understand what causes the artificial factors (or 'artifactors' as Marsh calls them). Tuerlinckx and De Boeck (2001) give two possible causes of relations between items that are not explained by the common factor they are both intended to relate to. The first refers to the presence of more than one underlying dimension, resulting in a residual correlation after accounting for the common factor. A latent variable that is believed to have pervasive effects of this nature is ARS (Mirowsky and Ross 1991). A second possible cause of residual correlation between items refers to item interaction. In the case of item interaction, a individual's response to item i affects her/his response to item i'.

To sum up, this leaves three sources of shared variance between any two items: (1) the intended common factor; (2) ARS, which operates independently of content (Rorer 1965); and (3) the interaction between items measuring the same construct (in the same or reversed direction). The latter effect is dependent on accessibility of item i when responding to item i', as will be discussed later. Now the effect of ARS and item interactions will be focused upon, after which hypotheses will be generated.

BIAS DUE TO ACQUIESCENCE RESPONSE STYLE

Researchers often find low absolute correlations between items presumably measuring opposite poles of one bipolar dimension. This has led to intense debates on whether or

not given constructs, like valence of affect, self-esteem and others, are best conceptualized as one bipolar dimension or two unipolar dimensions (Bentler 1969; Russell and Carroll 1999; Carroll, Yik, Russell and Barrett 1999; Marsh 1996; Motl and DiStefano 2002; Warr, Barter and Brownbridge 1983; Watson 1988). Several researchers have identified Acquiescence Response Style (ARS) as the main culprit for the confusion (McClendon 1991a; Bentler 1969; Russell and Carroll 1999; Carroll, Yik, Russell and Barrett 1999; Motl and DiStefano 2002; Warr, Barter and Brownbridge 1983; Mirowsky and Ross 1991). In particular, ARS variance in measures of affect is assumed to lead to a spurious increase in observed correlations, inflating positive correlations and biasing negative correlations upwards towards zero (Green, Goldman and Salovey 1993; Tomas and Oliver 1999). This effect has been acknowledged to be present in other content domains as well (Paulhus 1991; Podsakoff et al. 2003). The net result of this effect is that the baseline correlation between two unrelated items is expected to take on a positive value, rather than zero. Hence the following hypothesis is advanced:

H1: After controlling for content, the expected correlation between two items is positive.

The effect of ARS has been shown to generalize at least over the items within one questionnaire (Greenleaf 1992a). However, using an ad hoc set of items, Hui and Triandis (1985) find that nearby items may share more common response style bias than do items that are further apart. This shows in the correlations between scores of neighbouring parts of a questionnaire: the closer two subsets of items in a questionnaire, the higher their correlation. The theoretical base for this phenomenon would be that ARS has at least a component that is unstable over time (within the

period of filling out a questionnaire) and hence shows its effects in a local rather than a general manner.

In line with this, a second hypothesis is proposed:

> H2: The positive correlation between unrelated items will decrease as a function of inter-item distance.

While the above-mentioned hypotheses apply to any item-pair, regardless of content, the next discussion will focus on how the placement of items might further affect inter-item correlations for contentwise related items in particular.

ITEM INTERACTIONS: THE EFFECT OF ITEM LOCATION

In marketing research, there are two common methods of positioning contentwise related items within a questionnaire (Ostrom, Betz and Skowronski 1992). In the first, the researcher positions items that measure the same construct together in blocks. Other researchers use the second method, dispersing same-construct items over the questionnaire, mixing them with other-construct items. The idea of the latter method is that the content and meaning of an item should be clear in and of itself and that grouping same-construct items might lead to an artificially high internal consistency (Budd 1987; McFarland, Ryan and Ellis 2002). It is not clear, however, how these practices affect the interpretation of and responses to the items, and how this in turn might affect the validity of reversals.

Budd (1987) shows that respondents' degree of consistency across related items increases when the relationships between these items are obvious. To the respondent, topical organization of the items often is a clear indication of conceptual organization. As Ostrom, Betz and Skowronski (1992, p. 297) see it, "*People do not just passively respond to survey questions as if they were looking up answers in a dictionary, but they actively form cognitive representations of the survey and its items. These*

representations, in turn, guide the respondent's answers." Studies by Knowles (1988) and Knowles et al. (1992), Ostrom et al. (1992) and Budd (1987) have shown that responses to items are not merely a function of the item itself, but are also affected by the presence and proximity of other items measuring the same construct. Specifically, Budd (1987) has found that grouping items that measure the same construct lead to higher inter-item consistency in components of the Theory of Reasoned Action. Similar findings were obtained by McFarland, Ryan and Ellis (2002) in a personality assessment context. Ostrom et al. (1992) suggest that respondents construct a cognitive representation of what the questionnaire is actually measuring. This representation, which can be continuously updated, then guides responses to subsequent items.

Items that are near one another are more readily interpreted as tapping the same construct (Budd 1987; Ostrom, Betz and Skowronski 1992). Moreover, carryover effects have been shown to be rather local, fading out with increasing inter-item distance (Feldman and Lynch 1988; Tourangeau, Rasinski, Bradburn and D'Andrade 1989; Tourangeau, Singer and Presser 2003). The foregoing leads to the prediction that same-direction items will correlate more strongly the nearer they are to one another:

> H3: After controlling for ARS, the correlation between a pair of items measuring the same construct in the same direction will decrease with increasing inter-item distance.

It is less clear, however, how inter-item distance will affect the correlation between reverse-direction items. At least two outcomes are plausible, each of which is in line with current theorizing. The negative correlation between an item pair i and i' can

either decrease or increase in strength with increasing inter-item distance (after ARS has been controlled for).

Hypothetically two basic models of the way respondents process reversals can be proposed: “Unipolar Responding” (UR) versus “Bipolar Responding” (BR). In the BR model, respondents react to item i and item i’ the way the researcher intended. This means that the respondents interpret the items as opposite in meaning, retrieves all information relevant to this construct, and base their answers to both items on this information, making sure to reverse the response to item i’ when mapping the overall judgment to the response scale.

In the UR model, the respondents interpret item i as related to construct ξ, retrieve information relevant to this item and formulate a response (for a discussion of the response process as a whole, see Tourangeau, Rips and Rasinski 2000). When confronted with item i’, they interpret this item as linked to another construct ξ’, retrieve information they deem relevant to this construct and answer to item i’ based on this information. The major issue here is that the respondents interpret item i’ as relating to a different dimension than i. Whether this is due to a conscious act requiring interpretation, hypothesis generation about the construct, and continuous updating of this hypothesis (Ostrom, Betz and Skowronski 1992) or an effect based on the retrieval of a different set of beliefs (Tourangeau 1992) is of secondary importance for the current study. It seems most plausible that both processes are closely related, in that interpretation of the question guides retrieval of relevant information (Tourangeau, Rips and Rasinski 2000).

Research has indicated that inter-item distance dissipates the effect exerted by preceding items on target items (Feldman and Lynch 1988; Tourangeau et al. 1989; Tourangeau et al. 2003). Under the UR model, when presumed opposite-direction

items are placed next to one another, respondents seem to be focusing on the construct that was activated by the first item, and may regard the second item as irrelevant to this construct. This interpretational process would lead nearby reversals to have a relation that is orthogonal rather than opposite (Ostrom et al. 1992). At least, this is what is expected if respondents fail to acknowledge the bipolarity of the construct and the opposite relations that items i and i' have towards it. Under the BR model, that is if respondents would respect the intended bipolarity, proximity of i and i' would result in a highly negative correlation between the two.

To sum up, current theory leads to two competing hypotheses concerning the outcomes of reversed items. Therefore, both are proposed as mutually exclusive hypotheses for empirical testing:

H4a: After controlling for ARS, the correlation between a pair of reversed items will become less strongly negative (closer to zero) the closer both items are located to one another in the questionnaire. This is called the Unipolar Response model.

H4b: After controlling for ARS, the correlation between a pair of reversed items will become more strongly negative (diverging from zero) the closer both items are located to one another in the questionnaire. This is called the Bipolar Response model.

To clarify, the inter-item correlations as expected under both models are depicted in Figure 4-1a and b.

Figure 4-1a

Hypothetical graph of r as a function of inter-item distance for the Bipolar Response model

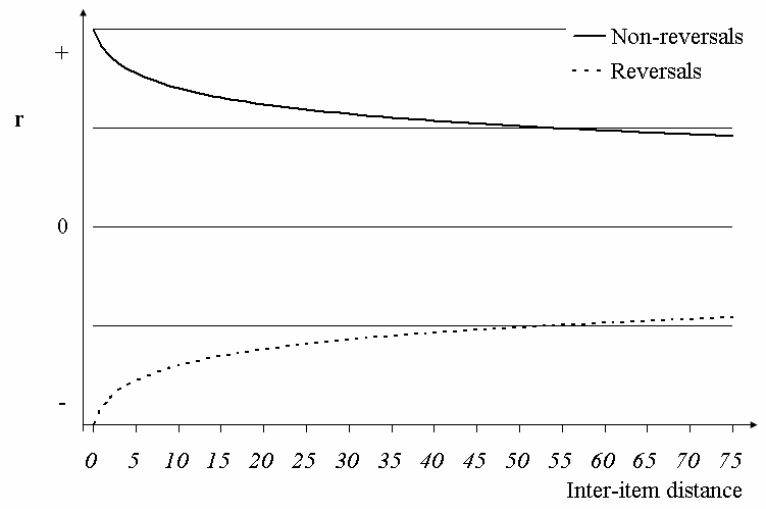
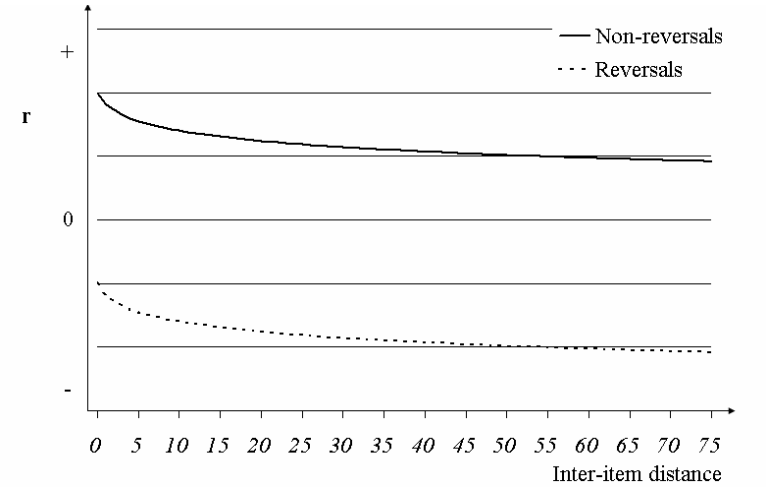


Figure 4-1b

Hypothetical graph of r as a function of inter-item distance for the Unipolar Response model



METHODOLOGY

RESPONDENTS

A sample was taken from the general online population by recruiting respondents on multiple major portal websites for the Dutch speaking part of Belgium. Data were collected by means of an online questionnaire which did not allow respondents to scroll back to previous pages. Respondents were told that the online survey was part of an academic study mapping the opinions of the population with regard to a wide variety of issues. 3114 valid responses were obtained. In this sample, 1607 respondents (51.6%) were male, 1179 (37.9%) had a higher education (i.e. formal education after secondary school), and the average age was 39.4 years (s=13.9).

ITEM SELECTION

The above hypotheses were tested using a data set based on a questionnaire that contained a wide variety of items, 76 in total, consisting of the following sets. (1) 10 pairs of reverse items (totaling 20 items) were randomly chosen from the scales compiled by Bruner, James and Hensel (2001). Each of the items was positioned randomly throughout the questionnaire, resulting in different distances between the respective pairs of reversed items. (2) Further, the items of two balanced multi-item scales were dispersed throughout the questionnaire (Dispositional Innovativeness, consisting of 3 positive and 5 negatively scored items; and Market Mavenism, consisting of 2 positively and 2 negatively scored items; Steenkamp and Gielens 2003). (3) The items of one unbalanced multi-item scale were also dispersed throughout the questionnaire (Susceptibility to Normative Influence, consisting of 8 positively scored items; Steenkamp and Gielens 2003). (4) Also included in the questionnaire was one unbalanced scale, the items of which were placed together as a

block of items (Trust and Loyalty in a clothing retail context, consisting of 4 positive trust and 4 positive loyalty items; Sirdeshmukh, Singh and Sabol 2002[5]). (4) Finally, 28 filler items were randomly selected from the scales compiled by Bruner, James and Hensel (2001). More specifically, a two step sampling procedure was used: first, scales were randomly sampled, after which one item was randomly sampled from each scale. If two scales related to the same content domain (e.g. price sensitivity), one was excluded from the sample. Consequently, these items were not contentwise related neither to the other items nor to one another. In addition, they were randomly dispersed throughout the questionnaire, in particular by having research assistants who were not informed about the purpose of the study, randomly assign the items to positions in the questionnaire.

DEPENDENT VARIABLE

As the data points in the analyses observed inter-item Pearson correlations were used. Therefore, Pearson correlations between all 76 items were computed. To account for missing data (all item pairs had at least 3000 valid observations), the correlation matrix was estimated using the EM (Expectation Maximization) algorithm in NORM (Schafer 1999). The EM algorithm is a method for obtaining maximum-likelihood estimates of parameters from incomplete data. The demographic variables age, sex and education level were used as covariates in estimating the correlation matrix (in line with the missing at random assumption; Schafer and Graham 2002).
Of a total of 2850 correlations, 29 were based on reverse coded item pairs and 71 were based on same direction item pairs. The other correlations were based on items

[5] These scales were coded as one construct because they were very closely related.

that had no contentwise relation to one another. All these correlations made up the dependent variable in a multiple linear regression model.

The observed correlations were regressed on independent variables that reflected questionnaire design and content factors that varied across the item-pairs under study. Studying correlations as the dependent variable was relevant because studies in the domain of reversals have focused on inter-item correlations (e.g. Wong et al. 2003), or methods based on correlations (e.g. Billiet and McClendon 2000), since inter-item correlations capture the variance shared by the items and indicate both the strength and direction of their association. The aim of the current study was to add to the understanding of how items correlate as a function of their shared content, response style bias and inter-item distance. The current approach required a shift in the data set from respondents to item pair correlations. In other words, the unit of analysis was not the respondent, but the inter-item correlation (computed across respondents). For a statistical discussion of the Pearson correlation, see Appendix 4-1. Similarly restructured data sets were used before to study response styles (Knowles 1988; Baumgartner and Steenkamp 2001, p. 153). Baumgartner and Steenkamp (2001) used correlations between scales as the dependent variable in a multi-level regression model. Likewise, Knowles (1988) used item-total correlations as the dependent variable in a regression model. In Knowles' regression model, serial position of the item was the main independent variable and all items measured the same construct. In the current study, the items tapped a wide diversity of constructs. Therefore, variables were included that capture this aspect of the correlation. More specifically, dummies were created that indicated whether a correlation was based on two items measuring

the same construct or not. Further, in stead of serial position[6], the inter-item distance was used as an independent variable of interest.

REGRESSION MODEL

The following regression equation was tested: $r_{ij} = \beta_0 + \beta_1 * LN\_DIST_{ij} + \beta_2 * SAME\_\xi_{ij} + \beta_3 * REVERSE\_\xi_{ij} + \beta_4 * DIST\_SAME_{ij} + \beta_5 * DIST\_REVERSE_{ij} + \varepsilon_{ij}$, where r_{ij} is the correlation between item i and j.

The intercept β_0 corresponds to the expected inter-item correlation for two subsequent items, controlling for contentwise relations. Hence, this intercept indicates the baseline correlation that is due to ARS variance shared by the items (Hypothesis 1).

Also, a variable was created indicating the distance between the two items in each correlation, expressed as the number of intervening items (i.e. the number of items positioned in between the two focal items). Because the effect of distance was expected to show a decreasing effect, the natural logarithm is taken of (distance + 1) resulting in the independent variable LN_DIST. This transformation compresses the distance scale as it takes on higher values, which is in line with theoretical expectations (Feldman and Lynch 1988). The main effect of LN_DIST on r corresponds to the notion that nearby items may share more common response style bias than do items that are further apart (Hypothesis 2).

Further, two dummy variables were created: the first dummy marks item pairs assumed to tap a same latent construct in the same direction (SAME_ξ). A second dummy variable flags item pairs that tap a same latent construct in the reverse direction (REVERSE_ξ). The variable DIST_SAME is equal to LN_DIST for

[6] which was relevant given the presence of only one construct in Knowles' study, such that serial position corresponds to the cumulative exposure to measures of the same construct.

SAME_ξ pairs, zero otherwise, and DIST_REVERSE is equal to LN_DIST for REVERSE_ξ pairs, zero otherwise. In other words, these terms represent the interactions between distance on the one hand, SAME_ξ (Hypothesis 3) and REVERSE_ξ (Hypothesis 4a and Hypothesis 4b) respectively on the other hand. Finally, the disturbance term (ε_{ij}) captures the variance in inter-item correlations that has not been accounted for by the above variables, including correlations due to specifics in content and/or form, not captured by the dummy indicating their measuring the same construct.

RESULTS

With an R^2 of .454 the regression model explained a sizable proportion of variance in the observed correlations ($p<.001$; adjusted $R^2=.453$). The multiple linear regression analysis assumptions were met. First, all condition indexes were below 7, indicating there was no problem of multicollinearity. The standardized residuals showed approximately normal distributions (as revealed on a normal P-P plot of the regression standardized residuals). Additionally, the regression coefficient estimates were robust, since they varied only mildly when estimating the model on different subsamples of correlations and using different model specifications (see below).

Table 4-1 lists the results of the regression analysis. The observed correlations between same- and reverse-direction items as a function of LN_DIST are shown in Figure 4-2a. Figure 4-2b depicts the regression implied predicted values of inter-item correlations over inter-item distance (untransformed).

Figure 4-2a:

Observed inter-item correlations (y-axis) and linear trend of same and reverse direction item pairs only as a function of LN_DIST (x-axis)

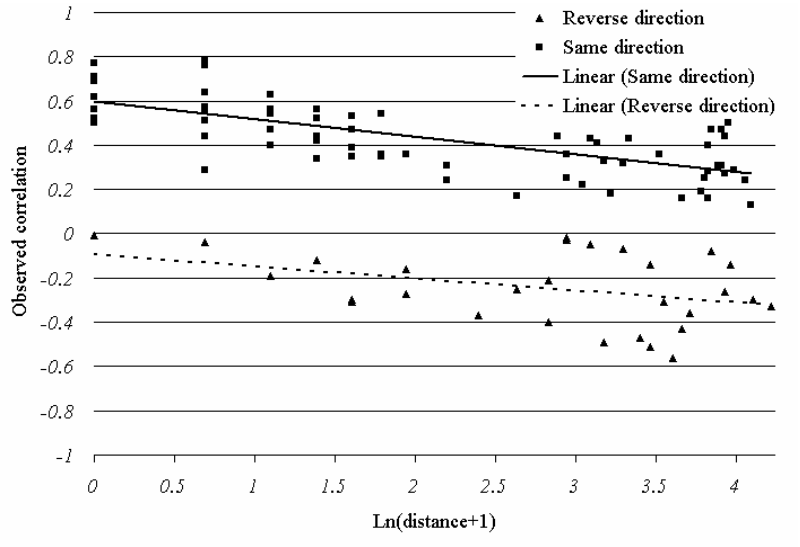


Figure 4-2b

Predicted inter-item correlation (y-axis) as a function of non-transformed inter-item distance (x-axis), based on regression estimates

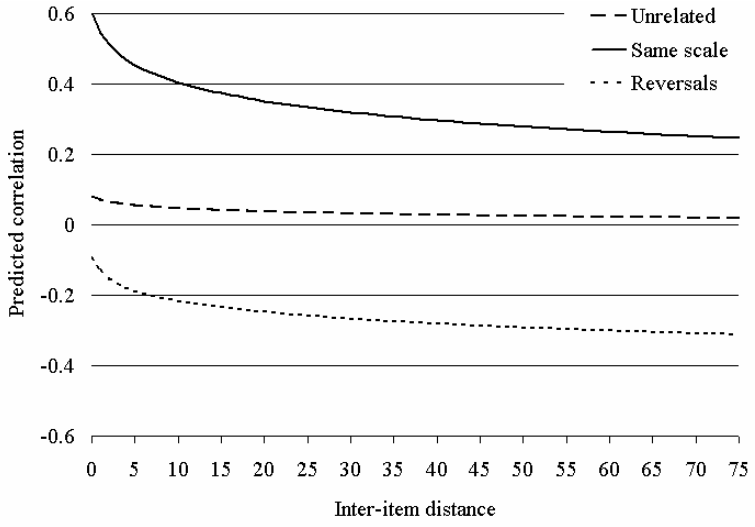


TABLE 4-1

RESULTS OF REGRESSION ANALYSIS ON OBSERVED CORRELATIONS

$R^2 = 0.45$	Unstandardized Coefficients[a]		95% Confidence Interval for B			
	B	s.e.	Lower Bound	Upper Bound	t	Sig.
Intercept	0.082	0.005	0.072	0.091	16.85	< 0.001
LN_DIST	-0.014	0.002	-0.017	-0.011	-9.13	< 0.001
SAME_ξ	0.521	0.017	0.487	0.554	30.41	< 0.001
REVERSE_ξ	-0.179	0.041	-0.260	-0.098	-4.34	< 0.001
SAME_DIST	-0.068	0.007	-0.081	-0.054	-9.87	< 0.001
REVERSE_DIST	-0.035	0.014	-0.062	-0.008	-2.56	0.010

[a] Only unstandardized coefficients are reported since both the independent variables and the dependent variable are expressed in a metric that is readily interpretable.

The intercept of the regression equation, 0.082, was positive and significantly different from zero ($p<.001$). This indicates that the average correlation between two items that are situated next to each other in a questionnaire (i.e. distance is zero) is positive, even after controlling for contentwise relatedness. This is consistent with the notion that ARS inflates correlations, as posited in Hypothesis 1. Further, as stated in Hypothesis 2, the main effect of inter-item distance was statistically significant and negative, but rather small ($B = -.014$). Linear extrapolation of this result beyond the range of the data - to obtain a mere indication - suggested it would take an inter-item distance of over 200 items to obtain a zero correlation (after rounding to two decimals) between two contentwise unrelated items.

The main effect of SAME_ξ was highly significant, positive and substantial in size. Specifically, the expected correlation of two items probing the same construct was 0.521 after controlling for the baseline correlation (i.e. the intercept, corresponding to

ARS) and the effect of distance (LN_DIST). For a pair of reversed items, the expected correlation at inter-item distance zero was -0.179.

The interaction effect between distance and respectively SAME_ξ and REVERSE_ξ were both significant and in the direction that is consistent with the Unipolar Response Model. Specifically, as inter-item distance increased, both the correlations between same-scale items and between reverse-scale items decreased. This means that the contentwise consistency for same-direction items goes down with distance, while going up with distance for reverse-direction items. Hence, Hypothesis 4a and the Unipolar Responding (UR) model were supported, while Hypothesis 4b and the Bipolar Response (BR) model were refuted by the results. It is important to note that the discrepancy between the expected correlation for reversed items and same-direction items is dependent on the inter-item distance at which the correlations are considered. Using the parameter estimates in Table 4-1, it is estimated that the absolute expected correlation between a pair of same-direction items (SAME_ξ) is equal to the absolute expected correlation between a pair of reversed items (REVERSE_ξ) if both pairs have inter-item distances around 45. In other words, if the distance measure would be centered on 45, the expected absolute correlation between reversals and non-reversals (considered at the intercept) would be equal in size. The reported results should therefore not be interpreted as indicating that reversals lead to lower absolute correlations as such. Rather, reversals that are positioned right next to their non-reversed counterparts lead to lower absolute correlations.

The same analyses were carried out taking polychoric correlations instead of Pearson correlations as the dependent variable. The results are described in Appendix 4-2 and led to the same substantive conclusions as reported above.

DISCUSSION

In this study, two findings are key: (1) the presence of a non-zero baseline correlation for nearby items which decreases as a function of inter-item distance; and (2) the reactivity of measurement, leading to an upward bias in both same-direction and opposite-direction correlations the nearer the items are to one another. Next, each of these is discussed in more detail.

POSITIVE BUT DECREASING BASELINE CORRELATION

Consistent with hypothesis 1, a positive correlation between items after controlling for content (SAME_ξ; REVERSE_ξ) was found. Note that this correlation did not emerge among an ad hoc set of related items, but among a very heterogeneous set of items, sampled from the scales compilation by Bruner, James and Hensel (2001). This result therefore adds considerable weight to previous findings and clearly corroborates the proposition that even unrelated items from validated scales are significantly correlated as the result of acquiescence response style (Baumgartner and Steenkamp 2001; Billiet and McClendon 2000). While the size of the baseline correlation is not huge, a correlation of 0.082 is definitely worrisome in light of the range of effect sizes of correlations and regressions commonly reported in social sciences (Green 1991). Response style bias can be expected to lead to overestimation of internal consistency of scales (Green and Hershberger 2000), and relations between scale variables (Baumgartner and Steenkamp 2001). The current results once again highlight that this problem should not be neglected and that researchers should take into account this bias in their analyses (see, e.g. Watson 1992).

In line with hypothesis 2, a decline in the positive inter-item correlation as a function of inter-item distance was observed. This finding lends some support to Hui and Triandis' (1985) finding that nearby items in a questionnaire share more common

response style variance than do items that are further apart. The decline is very shallow, however. Further research on the (in)stability of response styles seems warranted.

DECREASING UNIPOLARITY OVER INTER-ITEM DISTANCE

Since the attribute of being reversed or non-reversed applies to an item pair and not to a single item, there is no reason to expect that reversals have any specific characteristic that non-reversals do not have, since the two are interchangeable by definition. In this research, the first item was considered the non-reversed item i, and the one that follows this item as the reversal i'. Since the items were randomly assigned to serial positions, it was impossible to consider reversals and non-reversals as two separate classes of items to which different response processes apply due to the item considered in isolation. Keeping this in mind, the response to a reversed item seemed to be biased by the presence of its non-reversed counterpart. The net effect of this is that the expected absolute correlation between two nearby reversed items is much weaker than the expected absolute correlation between two same-direction items. While this finding seems to confirm the problematic status of reversed items as discussed by Marsh (1996) and Wong, Rindfleisch and Burroughs (2003), the current results also offer an important qualification. In particular, inter-item distance moderates the discrepancy between same- and reversed-direction items: negatively related items will have larger absolute correlations the further they are apart in the questionnaire, while for positively related items the opposite occurs. As Figure 4-2 illustrates, in the current data set the estimated absolute correlation for non-reversed and reversed pairs of items became similar once the two items were approximately 45 items apart in the questionnaire. This finding makes it plausible that in the absence of contamination by their reversals, inverse scored items may relate equally strongly to

the latent construct they operationalize as do their same-direction counterparts (after correcting for ARS; Schuman and Presser 1981; McClendon 1991). This makes perfect sense in light of the observation that the scaling - and hence the direction - of latent constructs is essentially arbitrary.

Moreover, the findings support a unipolar responding (UR) model: on average, respondents seem to interpret a reversal i' as orthogonal to its non-reversed counterpart i' if the two are positioned next to one another in the questionnaire. This effect dissipates over increasing inter-item distance. This decreasing reactivity of item with increasing inter-item distance is in line with previous research, including Feldman and Lynch (1988). However, paraphrasing Feldman and Lynch (1988), it could be argued that in the case of reversals, grouping items that measure the same construct might lead to 'self-generated non-validity' of the measurement model (rather than self-generated validity). The observed reversed item effect will lead to a factor structure in which reversed items show a loading near zero in stead of the expected negative loading. How strong this effect is, can be directly read from the data presented here, in that estimated factor loadings for a factor measured by two items i and i' will be equal to the square root of their absolute correlation, adding a negative sign for one of the items. So, for example, for two items that are next to one another in a questionnaire, one would expect loadings of approximately .28 ($=0.082^{1/2}$) if both are used to operationalize ARS, loadings of 0.72 if the two items are measuring the same construct in the same direction, and 0.42 if they are measuring the same construct in the opposite direction.

IMPLICATIONS FOR QUESTIONNAIRE RESEARCH

Obviously, the reported findings also bear upon the literature concerning the psychology of survey response. While balanced scales are used to partially correct for

bias due to response styles, the current results clearly show that the process of responding to reversals is somewhat more complicated than a straightforward acquiescence response style related account (as construed in the introduction) may imply. Specifically, it is found that reversals are not always responded to as such by respondents, and that this error is systematically related to the presence of a non-reversed item and its proximity to the reversal. Therefore, in addition to ARS, balanced scales may also be affected by other sources of error which are clearly content-related. These sources of error by definition do not classify as response styles as delineated by Rorer (1965). Rorer dismissed most of the response style literature based on the observation that it could not disentangle content from style. In addition, the response to reversals seems clearly distinct from the so-called baseline correlation that was observed between a large heterogeneous set of items. One important implication for response style research is that it may be most valid to measure response styles (conceptualized as pure behavioral tendencies not related to content; O'Neill 1967; Rorer 1965) by measuring consistent patterns of response selections over a heterogeneous set of items (Greenleaf 1992a, b) rather than as the number of double agreements, i.e. agreements to an item and its reversal (Watson 1992; Billiet and McClendon 2000). The latter method might be measuring a mix of response styles, interpretational differences and content. In this regard, it is significant that measures of double agreements to non-reversed and reversed items have also been used to measure attitude ambivalence (Wegener et al. 1995, p. 457). Double agreements with reversals may be partly due to non-content related response styles, but clearly are also a function of content-related context, mediated by top-down processing. Wong, Rindfleisch and Burroughs (2003) also point out that double agreements with reversed and non-reversed items are not merely the result of ARS,

but of interpretational problems due to the presence (and proximity, although this is not stated as such) of non-reversed items. In their study, Northern American respondents seem to be less context sensitive in this regard than are Eastern Asian respondents. Therefore, it would be interesting to investigate how the current results, obtained from a European sample, would generalize to other cultures.

IMPLICATIONS FOR THE USE OF BALANCED SCALES

This leaves the researcher with the question of whether or not to use reversed items and balanced scales. Based on the current as well as previous findings, the following recommendations can be proposed.

First, given the current state of knowledge, reversals should not be used to create measures of ARS. The process leading to double agreement to both an item and its reversal is more complex than a constant additive ARS model would imply. Incidentally, such measures have quite often shown low reliability (e.g. Watson 1992; Johnson, Kulesa, Cho and Shavitt 2005). It is safer to measure response styles as a general tendency to select particular responses (expressing agreement in the case of ARS) over a broad set of unrelated items (Greenleaf 1992b).

Second, when using balanced scales (e.g. because they are the only validated alternative available), it may be recommendable not to group the items. For example, a scale consisting of two same-direction items (i and j) and one reversal (k'), could be positioned in the beginning (i), the middle (k') and the end (j) of the questionnaire. This would reduce artificial inflation of the correlation between i and j, as well as artificial bias towards zero of the correlations between k' and i and between k' and j. Ideally, both recommendations have to be applied simultaneously in research designs.

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

The number of correlations based on unrelated items (2750 so-called baseline correlations) might seem disproportionate relative to the number of correlations based on reversed items (29 correlations) and non-reversed related items (71 correlations). This is a consequence of the fact that any filler item could be correlated with any other filler item, while the other types of correlations were much more selective by design. The apparent imbalance of contentwise unrelated items to contentwise related items is not problematic. By using the dummy specification that reflected the different types of correlations, and by creating interaction terms of these dummies with each of the effects, separate effects were estimated for the different categories of items, and all estimates had their own appropriate standard errors. At the same time, ARS was being controlled for in a highly reliable way (based on the many baseline correlations), such that the main effect and the effect moderated by distance of ARS could be assessed independently of the item-interaction effects. Further, the correlations were based on a large number of respondents (over 3000) which enhanced their stability and reliability (Zimmerman, Zumbo and Williams 2003), and the items were randomly assigned to positions in the questionnaire. These factors made it possible not to include extraordinarily large numbers of reversals in the questionnaire, which might have led respondents to become acutely aware of the set-up, possibly even leading them to see the task as a ‘reversal examination’ rather than an ordinary questionnaire.

The specific curve of reversed item correlations as a function of inter-item distance was attributed to a unipolar response model. The varied contents of the questionnaire in the current study renders implausible an otherwise appealing alternative explanation of this phenomenon. Specifically, if respondents fill out a series of

positively related items and are then confronted with a reversed item, careless reading might lead some respondents to misinterpret the reversed item as a same direction item (Schmitt and Stults 1985). However, for this effect to occur, it seems that many similar items should occur in an uninterrupted series (cf. Drolet and Morrison 2001). Though this was the case in studies in which a negative item method effect has been observed (e.g. Marsh 1996; Motl and Distefano 2002), it was not in the current study. Since the length of the questionnaire used for this study was limited to 76 items, most inter-item distances were quite small. The median inter-item distance in the data is 23. It would be useful to further study the current phenomena using longer questionnaires. Possibly, the effect of distance on r fades out completely after a given distance. The current data are too limited in scope to find out.

For now, good fit was obtained using the natural logarithm of (distance + 1). Though the natural logarithm is an often-used transformation (Greene 2003, p. 11-13; Tabachnick and Fidell 1996, p.80-82), other specifications are also possible, and some of these possibilities are shortly reviewed below. Note that the substantive findings were found to be robust over different specifications.

As an exploratory exercise several specifications of the regression model were tested: (1) a strictly linear model; (2) a model with quadratic effects of distance (and its interaction terms); (3) a spline regression, where the effect of distance (and its interaction terms) was allowed to be different in the inter-item distance range of 0-10 versus 11-76. However, the different specifications resulted in the same substantive conclusions, where (1) there is a significantly positive base correlation (the intercept) in the range of .05 to .08, which is slowly declining towards zero over distance, (2) a negative correlation between reverse-direction items which grows in strength (becomes more negative) over increasing inter-item distances, and (3) a stronger

correlation between same-direction items which also more pronouncedly declines over inter-item distance.

In addition to further quantitative research, it would be most interesting to further validate the current findings by means of cognitive interviews (DeMaio and Rothgeb 1996; Jobe and Mingay 1989). Specifically, it would be enlightening to study respondents' processing of unrelated items, same direction items and reversed items in a controlled setting. Using questionnaires similar to the one used in the current study, respondents could be asked to think aloud as they process the meaning of items and retrieve information. It would be especially relevant to observe the extent to which respondents refer to previous items and how respondents use the intended scoring direction of the items (non-reversed or reversed) and inter-item distance as input for the comprehension process. Another interesting probing technique would be to ask respondents to paraphrase reversed items, i.e. to word these items in the respondents' own words. This would be indicative of whether or not respondents refer to related concepts when processing reversed items.

Finally, a study is planned that approaches the issues investigated here from a different perspective. The current study used a between-item design with a one-time random assignment of items to locations. In a follow-up research, a between-subject design will be used. In this study, item content will be kept constant by investigating a pair of reversed items and a pair of non-reversed items. Item location of item i and i' will be randomized over respondents. The following regression model will be tested: $x_{i'} = \alpha + \beta_1 ARS + \beta_2 x_i + \beta_3 (x_i * LN\_DIST_{ii'}) + \varepsilon$, where $x_{i'}$ and x_i are the observed scores on item i and i', ARS is a measure of acquiescence measured over a set of heterogeneous filler items, $LN\_DIST_{ii'}$ is the natural logarithm of the distance +1 between item i and i', and α, β_1, β_2 and β_3 are the regression intercept and weights. α

corresponds to the mean of $x_{i'}$, β_1 to the effect of ARS, β_2 is expected to be negative and corresponds to the extent to which the extremity of a respondent's position on the construct underlying both items is identical in size (but opposite in direction) for i and i', and β_3 captures the effect of distance on this relation.

APPENDIX 4-1: STATISTICAL DISCUSSION OF THE PEARSON CORRELATION

Observed sample Pearson correlations are not without their limitations. A combination of factors leads observed inter-item correlations in general to be imperfect, and this from two perspectives: (1) the absolute population correlation $|\rho|$ between two items tapping the same construct is almost never equal to 1, and (2) the observed sample correlation r is not equal to the population correlation ρ. The main reason why inter-item population correlations will not be exactly 1 or -1 is that such items would be considered to be identical and hence redundant. The reasons why observed sample correlations are smaller in absolute size than ρ include coarseness of measurement scales (Green and Rao 1970), violations of distributional assumptions (Kraemer 1980), a slight structural bias towards zero (Zimmerman, Zumbo and Williams 2003), range restriction (Sackett and Yang 2000; Chan and Chan 2004), and random error in measures (Charles 2005). On the other hand, for rating scales having at least five response options, the use of Pearson correlations is defendable and quite commonly accepted (Bollen and Barb 1981; Srinivasan and Basu 1989). Moreover, the Pearson correlation remains a popular statistic in the social sciences, and most researchers readily understand the meaning of the size and direction of correlations. Therefore it is relevant to use Pearson correlations as the variable of interest in this study. To ensure that this choice does not influence the results in some way, Appendix 4-2 also presents the results of the same analysis using the polychoric correlation coefficients as the dependent variable.

In the analyses, untransformed correlations are used rather than a Fisher z-transformation for several reasons. First, raw correlations are more meaningful and easier to interpret (e.g. the meaning of a .05 change in r is readily interpretable to most researchers). Second, the correlations in the current study have a mean value of .044

(SD=.11), with a minimum of -.56 and a maximum of .78. Consequently, most observed values are removed far enough from (-) 1 not to worry about the instability of the variance of r near (-) 1. In addition, the estimates in the current empirical study will be based on a sufficiently large sample of respondents to reasonably assume stable and nearly unbiased estimates (Zimmerman, Zumbo and Williams 2003). Finally, the z transformation in the first place applies to r estimates sampled from the same population of real correlations, while in this study, each observed r is an estimate of a different true correlation.

APPENDIX 4-2: REPLICATION USING POLYCHORIC CORRELATION COEFFICIENTS

The polychoric correlation coefficient is a measure of association that serves as an alternative to the Pearson r in situations in which the variables of interest are continuous but the measurement instruments yield ordinal data (Pearson and Pearson 1922). Procedures for estimating the polychoric developed by Olsson (1979) are based on the assumption that the unseen underlying variables are continuous and have a bivariate normal distribution. The polychoric correlation coefficient, calculated from ordinal transformations of bivariate normal variables, results in an unbiased estimate of the correlation between the original bivariate normal variables (Olsson 1979). Babakus and Ferguson (1988) recommend its use when data are ordinal.

The polychoric correlation matrix of the 76 items was estimated in Mplus 4.0. Application of the regression model discussed in the main text to these data gave the estimates in Table 4-2-1.

TABLE 4-2-1

REGRESSION ESTIMATES FOR POLYCHORIC CORRELATIONS

$R^2 = 0.41$	Unstandardized Coefficients[a]		95% Confidence Interval for B			
	B	s.e.	Lower Bound	Upper Bound	t	Sig.
Intercept	0.092	0.006	0.080	0.104	15.24	<0.001
LN_DIST	-0.017	0.002	-0.020	-0.013	-8.43	<0.001
SAME_ξ	0.578	0.021	0.536	0.620	27.00	<0.001
REVERSE_ξ	-0.192	0.052	-0.293	-0.091	-3.73	<0.001
SAME_DIST	-0.064	0.009	-0.081	-0.047	-7.45	<0.001
REVERSE_DIST	-0.044	0.017	-0.077	-0.010	-2.55	0.011

The results led to the same substantive conclusions, but some remarks are in place. First, the intercept was even higher than in the analysis using r as the dependent variable. This indicates that the effect of acquiescence response style may be underestimated if the coarseness of the scale is not taken into account. In line with this, the correlations between same-construct and reverse-construct items were slightly stronger (i.e. respectively more positive and more negative) in the current analysis. The distance effects were similar to those obtained when using the Pearson correlation, with the main effect and the REVERSE_DIST effects somewhat stronger, the SAME_DIST effect a little weaker when using polychoric correlations. In sum, the findings reported above are not limited to Pearson correlations, but also generalize to polychoric correlations.

CHAPTER 5: THE SHORT TERM STABILITY OF RESPONSE STYLES (EMPIRICAL STUDY 2)

CHAPTER OUTLINE

Based on a literature review, nine models are proposed that specify the extent of (in)stability over a single questionnaire administration of four response styles: acquiescence, disacquiescence, midpoint and extreme response style. Using secondary data (Hui and Triandis 1985) and primary data, a comparison of these nine models is made based on model fit and model estimates. It is concluded that response styles have a major stable component that might need to be complemented by an autoregressive component in specific cases. Implications of these results are discussed.

INTRODUCTION

Much of the research in the social sciences heavily depends on respondents' self-reports. A good deal of these self-reports use scales consisting of closed-ended agree-disagree items. Unfortunately, such measures are often biased by response styles, defined as behavioral tendencies to disproportionately select a subset of the available response options (Rorer 1965; O'Neill 1967). The following such response styles have been defined and studied in the behavioral sciences: acquiescence response style (ARS), disacquiescence response style (DRS), extreme response style (ERS), and midpoint responding (MRS), which respectively refer to disproportionate use of the alternatives at the positive end, the negative end, the extreme ends, and the middle of the rating scale (e.g. Baumgartner and Steenkamp 2001; Greenleaf 1992b; O'Neill 1967; Rorer 1965; Van Herk, Poortinga and Verhallen 2004; Johnson et al.2005). The extent to which these response styles should be expected to systematically affect agreement-disagreement scores, and the relations between such scores, revolves around the issue of their stability. In the best case scenario, the effect of a response style does not generalize across any two items and reduces to random error. Since behavior that does not generalize across different stimuli or time stops being a tendency, in that case response styles are but a myth, as Rorer (1965) has stated. At the other extreme of the range of possibilities, response styles may be highly stable personal characteristics that cause bias with a high within-subject consistency (Jackson and Messick 1958; Hamilton 1968). The worst case scenario is the situation in between, where individuals' response styles show both a generalizable and an idiosyncratic component. In this case, item responses will be biased by response styles, but the bias is hard to correct for. The reason is that correction for response styles depends on the ability to construct reliable and valid measures of response

styles (Greenleaf 1992a, b), something that is impossible if they fluctuate substantially (Hui and Triandis 1985, p. 259). While this matter is far from trivial, previous research has had to take position on this issue without a thorough empirical comparison of the alternative models that may apply. The current study makes a systematic assessment of the (in)stability of response styles over the items within a single questionnaire by comparing alternative models that have been proposed implicitly or explicitly in the literature. To this end, alternative models of response styles are fitted to data that were collected with the specific aim of studying response styles. Before that, a secondary analysis is conducted of data presented by Hui and Triandis (1985) in support of the instability of response styles.

First, the literature on response styles is reviewed and from it alternative conceptual models on the styles' stability are distilled. Next, these conceptual models are translated in operational models, more specifically common factor and auto-regressive models as well as hybrids of the same. These models are then subjected to a methodical comparison in a structural equation modeling framework. The results of these model comparisons answer the question of how stable response styles are over the course of a questionnaire.

CONCEPTUAL FRAMEWORK

THEORIES ON THE STABILITY OF RESPONSE STYLES

The following discussion focuses on the situation where individuals would respond to a questionnaire consisting of subsequent sets of contentwise unrelated items. Response style indicators could be computed for each set of items. Indeed, since the items do not share content variance, the variance they share is to be attributed to response styles (Greenleaf 1992a, b). Assume there are k such indicators based on k

subsequent parts of the questionnaire. The question now is what different response style researchers would predict in terms of the relations between the response styles present in these k subsequent sets of items for the case[7] where k = 5. The k = 5 response style scores for a respondent will be represented by a 5*1 vector [y_1, y_2, y_3, y_4, y_5]' and are assumed to be mean centered. Consequently, no intercept term will be included in the equations.

Non-existence of response styles

Rorer (1965) dismissed the complete response style literature up till 1965 by pointing out it did not prove any generalizable effect of response styles. Basically, Rorer stated that response style researchers seemed to have forgotten the possibility that their respondents actually might have responded to content. Based on his extensive literature review, he reached the conclusion that response styles do not exist, and that one should not expect sets of items that are contentwise unrelated to show shared variance merely due to respondents' tendencies of systematically selecting certain response options rather than others. Operationally, this would imply that response style indicators based on subsequent contentwise independent sets of items do not correlate. This is labeled the independence model, in which

$$[y_1, y_2, y_3, y_4, y_5]' = [\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5]' \qquad (1a),$$

where the ε's represent the individual deviation score and are uncorrelated. Hence,

$$\Sigma_{RS} = \text{Diag}(\Sigma_{RS}) \qquad (1b)$$

[7] This number is arbitrary in the current context, but will be the number of indicators used in the empirical part of this study, since it is the number of available indicators in the data reported by Hui and Triandis (1985), and because it is the minimal number of indicators for which all models are identified. This will become clear later on in the text.

This model is especially relevant because it is the implicitly assumed model when studying relations between self-report measures in the same format without taking into account response style bias, a common practice in many studies (that is criticized by Ray 1979; Paulhus 1991, and defended by Schimmack, Böckenholt and Reisenzein 2002).

Instability of response styles

A more moderate approach was taken by Hui and Triandis (1985), who posited that response styles are not stable, but that they gradually evolve over the course of a questionnaire. In other words, response styles in a set of items can be predicted best by the response styles in the preceding set. The authors based this conclusion on the observation that the correlation matrix of subsequent response style indicators shows a simplex pattern, i.e. the size of the correlations declines the further one moves away from the main diagonal. This indicates that response style indicators based on subsequent parts of the questionnaire correlate more highly than response style indicators based on remote parts of the questionnaire.

Conceptually, Hui and Triandis suggested that the response style level in a part of the questionnaire relates directly only to the response style levels in the preceding part of the questionnaire, rather than being stable throughout. The authors stressed this apparent instability of response styles (hence the title of their article) and questioned the validity of measures of response styles that generalize across a whole questionnaire (p. 259). Operationally, the direct effect from a response style indicator to the subsequent one only (and indirect effects to the following indicators mediated by this effect) translates into an autoregressive model (Marsh 1993; Green and Hershberger 2000). Formally, this means a response style indicator can be decomposed in the effect from the preceding indicator and a random component.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \beta_{21} & 0 & 0 & 0 \\ 0 & \beta_{32} & 0 & 0 \\ 0 & 0 & \beta_{43} & 0 \\ 0 & 0 & 0 & \beta_{54} \end{bmatrix} * \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix} \quad (2a),$$

where y is a k*1 vector of subsequent response style indicators, β is a k*k lower diagonal matrix with autoregressive weights and ε is a k*1 vector of unique components. Two alternative versions of this model are conceivable. In (2a), the autoregressive coefficient is time variant. It can also be time invariant, such that

$$\beta_{21} = \beta_{32} = \beta_{43} = \beta_{54} = \beta \qquad (2b).$$

The data presented by Hui and Triandis in support of their instability hypothesis do not seem to definitely rule out the presence of a stable component of response styles, in that even response styles in remote parts of the same questionnaire were substantially correlated. To further probe this issue, in the empirical part of the current study the relative weight of the local and generalizable components of response styles in Hui and Triandis' data will be assessed.

Stability of response styles

Paulhus (1991) - and based on his work also Baumgartner and Steenkamp (2001) - took the view that response styles are due to an interaction of person and content. In other words, for a given respondent, the level of response style bias in a given set of items is decomposable into the influence of a common response style factor and a unique factor characteristic of the set of items. The influence of the common and the unique factor varies across sets of items without there being an order effect present (the relative position in the questionnaire is not considered as being of major relevance).

Operationally, the latter model is a congeneric factor model (Anderson and Gerbing 1988), in which the response styles in all sets are related to a single underlying factor, where factor loadings and unique variances can freely vary across sets of items.

Assuming $E(\xi\delta')=0$,

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix} * \begin{bmatrix} \xi \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix} \quad (3a)$$

Greenleaf (1992b) specified conditions under which different response style indicators show tau-equivalence, which means the impact of the common response style factor would be the same for all indicators[8]. Other researchers have imposed tau-equivalence in models of response styles where this constraint could not be tested for reasons of identifiability (Billiet and McClendon 2000; Mirowsky and Ross 1991).

Tau equivalence translates into the additional condition that

$$\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 \quad (3b).$$

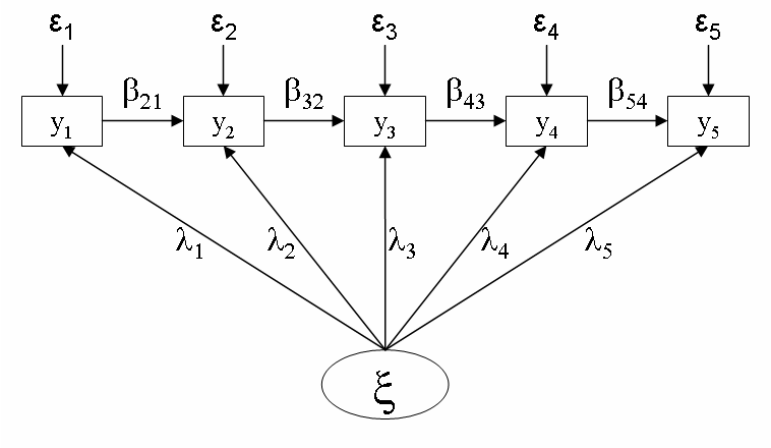
A comparison of competing models

The different models presented above have divergent consequences for research using agree-disagree items, both in terms of bias (is it general and stable) and in terms of the potential to solve for such bias (can it be reliably measured). For this reason, it is important to formally compare these alternative models of response style stability. This was the purpose of the current study.

[8] Since the current study focuses on covariance structures not including mean structures, for reasons of readability the term tau-equivalence is used to refer to essential tau-equivalence (and no constraints are formulated for the intercepts). These concepts are used in their traditional meaning, see Traub (1994, p. 56-57).

To give more structure to the model comparison, all models are organized along two dimensions. The first dimension relates to the autoregressive coefficient, which can be zero, time-invariant, or time-variant. The second dimension relates to the common factor, the loadings on which can be zero, equal across sets, or set-specific. Figure 5-1 depicts the model in which both a common factor (with loadings labeled λ) and autoregressive effects (labeled ß) are present.

Figure 5-1
Hybrid model of response styles



Using the notation presented in equations 1 through 3, this general model can be expressed as follows.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \beta_{21} & 0 & 0 & 0 \\ 0 & \beta_{32} & 0 & 0 \\ 0 & 0 & \beta_{43} & 0 \\ 0 & 0 & 0 & \beta_{54} \end{bmatrix} * \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} + \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix} * \begin{bmatrix} \xi \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix} \quad (4)$$

Table 5-1 provides an overview of the nine alternative models that can be specified based on this general model, by restricting parameters along the two dimensions discussed above (common factor constraints, autoregressive coefficient constraints).

TABLE 5-1: OVERVIEW OF THE MODELS OF RESPONSE STYLE STABILITY

	A. Congeneric	B. Tau-equivalent	C. No common factor
1. Time-variant autoregressive	q = 3k-1 $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ $\beta_{21}, \beta_{32}, \beta_{43}, \beta_{54}$	q = 2k $\lambda, \beta_{21}, \beta_{32}, \beta_{43}, \beta_{54}$	q = 2k-1 $\beta_{21}, \beta_{32}, \beta_{43}, \beta_{54}$
2. Time invariant autoregressive	q = 2k+1 $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \beta$	q = k+2 λ, β	q = k+1 β
3. Non-autoregressive	q = 2k $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$	q = k+1 λ	q = k

Shown are nine models with their respective number of freely estimated parameters, and (in italics) the labels of the freely estimated autoregressive coefficients and factor loadings. k = number of indicators; λ = factor loading; β = autoregressive coefficients; q = number of parameters that have to be estimated; Note that for each model, each indicator has a (residual) variance to be estimated, adding k parameters to each model.

METHODOLOGY

To assess the stability of response styles, nine structural equation models were specified. Data on subsequent response style indicators were used. First, the correlation matrices reported by Hui and Triandis (1985) were analyzed, because they provided some of the little information on the stability of response styles available in the literature. Second, primary data based on a random set of items measured on seven-point scales were analyzed.

SECONDARY DATA (HUI AND TRIANDIS 1985)

Hui and Triandis (1985) reported three studies. Since only the correlation matrices of the first two studies were provided in the article, the current discussion focuses on these data. Attention is also limited to data of net acquiescence response style (NARS; equivalent to ARS – DRS) and extreme response style (ERS), since these are the tendencies that fall under the strict definition of response styles used here, in line with

Rorer (1965). For the data themselves, the reader is referred to the original article. The first study (henceforth called H&T1) was based on object ratings on 10-point semantic differentials (N=219). The second study (henceforth referred to as H&T2) was based on evaluations of self-concept related statements on 5-point Likert rating scales (N=145).

PRIMARY DATA

Also, primary data were collected with the specific aim of measuring response styles. The questionnaire consisted of a randomly selected set of items. This made it particularly well-suited for measuring response styles.

Respondents

Respondents were recruited from the panel of an online market research company. The sample was selected to represent a cross-section of the Belgian population in terms of age, gender and education levels. From the 1372 panel members who were contacted by e-mail for participation, 604 provided valid responses (response rate = 44%). 490 of these were one hundred per cent complete.

Items

Items were sampled from the Marketing Scales Handbook by Bruner, James and Hensel (2001) and Measures of Personality and Social Psychological Attitudes by Robinson, Shaver and Wrightsman (1991). From these books, 112 items from different scales were randomly selected. The items were put together in an uninterrupted random list making up the complete questionnaire.

Response style indicator calculation

The items were divided into five sets, corresponding to five subsequent parts of the questionnaire. Each set consisted of 22 or 23 items. Five sets were used because this

resulted in the minimum number of indicators that allowed estimating all nine proposed models (Table 5-1). Also, this meant that each set consisted of a sufficient number of items to reasonably assume their validity as response style indicators (Greenleaf 1992a). The five sets were used to compute five indicators for every response style (ARS, DRS, ERS and MRS). For ARS, the number of agreements per set of items was summed after weighting a seven as three points, a six as two points, and a five as one point. A similar method was applied to obtain DRS measures (Baumgartner and Steenkamp 2001). ARS and DRS indicators reflect the expected deviation from the midpoint due to ARS or DRS respectively if means would be computed based on the item responses. ERS indicators reflect the proportion of extreme responses (1 or 7). Similarly, MRS indicators reflect the proportion of midpoint responses (4).

DATA-ANALYSIS

The independence model (C3 in Table 5-1) corresponds to the position that response styles do not generalize across different sets of items. The other two models in column C of Table 5-1 correspond to the position that response styles are unstable (no common factor) and only have a local effect (the autoregressive coefficient), which can be time variant (C1) or time invariant (C2). Model B3 corresponds to the stance that all sets of items are affected only by a common response style factor and this with equal strength for all sets. This model is assumed when constraining response style factor loadings to one for different (sets of) items (Billiet and McClendon 2000; Mirowsky and Ross 1991). The other B models hold the latter assumption too, but allow for an additional autoregressive component of response styles. Model A3 assumes a single underlying response style that may have a different impact on different sets of items (Greenleaf 1992a, 1992b; Watson 1992; Baumgartner and

Steenkamp 2001). The other models in column A again present hybrid extensions of this model that may be important because the autoregressive model and the common factor model are not mutually exclusive, but seem to have been treated as such in the literature nonetheless.

Models that are in the same row or column are nested within one another, that is, the set of freely estimated parameters of each model is a subset of those estimated in the model(s) preceding it in the same row as well as the model(s) preceding it in the same column. Note that A1 is not nested in any other model. This model is overly liberal, in that for small numbers of sets (like in the current study, where k=4 or k=5), the degrees of freedom are limited. This model will mainly serve as a reference model.

Each model is estimated for each response style and evaluated in three major ways. As pointed out by Marsh, Hau and Wen (2004), meeting common goodness-of-fit cutoff criteria is not a sufficient criterion for having a valid model. Goodness-of-fit criteria usually perform better in comparing alternative models based on the same data (Marsh, Hau and Wen 2004). Therefore the different models are also evaluated with respect to one another. Additionally, the theoretical viability, statistical significance and substantial size of the parameter estimates are assessed.

To sum up, first, model fit of the stand-alone models will be evaluated. Second, model fit will be evaluated relative to the other models (taking into account nesting). Third, the substantive meanings of the model estimates are appraised. Each of the three steps is now discussed in more detail.

Absolute model fit

The chi square statistic allows for a formal test of model fit. However, since some sample sizes are large enough to expect some oversensitivity of the chi square test statistic (Marsh, Balla and McDonald 1988), alternative fit indices are also taken into

account (Hu and Bentler 1999). The RMSEA (Root Mean Square Error of Approximation, Steiger 1990; Browne and Cudeck 1993) takes into account model complexity by dividing the minimum discrepancy by the number of degrees of freedom for testing the model. This is important since the number of parameters relative to the number of distinct sample moments varies widely over the models and parsimony is considered a plus. Additionally, the confidence intervals around the RMSEA estimates are helpful in comparing models. The CFI (Comparative Fit Index; Bentler 1990) is particularly relevant in this context since it evaluates the decrease in misfit (captured by the noncentrality parameter) relative to the independence model, i.e. model C3. This means that the CFI of model C3 will be zero by definition, while a saturated model will have a CFI of 1. The range and meaning of the CFI precludes its use in assessing model C3, but if the latter model is rejected based on other criteria, the CFI becomes useful in assessing how well the other models account for the covariances between the indicators that are constrained to zero in model C3. Values close to 1 indicate very good fit, .95 is commonly used as a cut-off value (Hu and Bentler 1999). The CFI and RMSEA are two alternative fit indices often referred to by experts (e.g. Flora and Curran 2004).

Relative model fit

Since models in the same column or row are nested, nested chi square difference tests are performed. Here again, chi square may be oversensitive due to the sample size (in the primary data). Therefore, a decrease in CFI equal to or higher than .01 is evaluated as indicative of a relevant deterioration in fit (Grouzet, Otis and Pelletier 2005), a decrease of .05 or more as a substantial non-acceptable deterioration in fit (Little 1997; note however, that this recommendation was based on multi-group invariance tests; generalization to the current setting is therefore somewhat tentative). Another

marker of a substantial deterioration of fit is the extent of separation/overlap between RMSEA confidence intervals.

Estimates

In addition to the above evaluations of model fit, model estimates were evaluated by checking whether the relevant estimates were significantly different from zero and were signed in the expected direction. In particular, in the congeneric and tau-equivalent models (all models A and B), factor loadings were expected to be significantly positive. If the loading of a specific response style indicator was not significantly different from zero, this would imply that the indicator in question is not significantly related to a common response style factor. If its loading is negative, this would indicate that higher levels of response styles in other sets of items are predictive of lower levels of response styles in the set in question. In the autoregressive models (all models 1 and 2), the autoregressive weights were expected to be significantly positive. A similar reasoning applied here. If the autoregressive coefficient of a specific response style indicator was not significantly different from zero, this would imply that the indicator in question was not significantly related to the previous indicator. If its coefficient is negative, this would indicate that higher levels of response styles in the previous item set is predictive of lower levels of response styles in the set in question. In addition to the evaluation of significance, size and direction of the loadings and autoregressive coefficients separately, the relative size of the estimates related to autoregression were compared with those related to a common factor.

RESULTS

The correlation matrices provided by Hui and Triandis (1985) were analyzed using a ML estimator (MPlus version 4; Muthén and Muthén 2006). The primary data were

analyzed using a FIML estimator which takes into account missing values (Amos 5.0.1; Arbuckle 1994-2003). All proposed models were fit to four correlation matrices (NARS and ERS in H&T1; NARS and ERS in H&T2) and four covariance matrices (ARS, DRS, ERS and MRS for the primary data). It was chosen to estimate a separate model for each response style to get results that could be directly compared to the results obtained from the H&T data and because this allowed being very specific about what causes misfit in the models. Also, the scenario where data on different response styles fit different models is considered a possibility. Note that model A1 (the time variant autoregressive congeneric model) cannot be estimated with four indicators because this would result in negative degrees of freedom. Hence, model A1 was not estimated for NARS and ERS in H&T1. All other models were identified and the estimations converged without any problems. There were no instances of inadmissible solutions.

A CAUTIONARY NOTE ON MODEL A1

Before discussing the other models, it is worth focusing the discussion shortly on model A1 alone. As expected, an investigation of the estimates shows that the value of model A1 is questionable. While it fits the data good for all response styles and all data sets, this seems to be due to the absence of constraints rather than good validity. This allows the algorithm to approach the observed correlation/covariance matrices with estimates that are not necessarily meaningful but that are admissible within the set of constraints. As discussed above, the factor loadings and the autoregressive coefficients would be expected to be positive and significantly different from zero.

Several autoregressive coefficients did not meet these requirements[9]. Close inspection of the estimates leads to the following conclusions. First, free estimation of an autoregressive coefficient for each pair of response style indicators and a factor loading for each individual indicator has questionable validity and leads to over fitting. The superior fit of this model should be treated as confirmation of its status as a nearly-saturated reference model without much value as a stand-alone model. Second, the freely estimated autoregressive coefficients are quite unstable and small relative to the factor loadings on the common factor.

MODEL FIT EVALUATION

The model fit indices based on the H&T data are listed in Table 5-2. Table 5-3 lists the fit indices based on the primary data. Figure 5-2 shows the 90% confidence intervals for the RMSEA's of all models (based on the H&T1, H&T2 and primary data respectively). Although the sheer amount of information may be overwhelming at first, some clear and remarkable trends are apparent that seem to generalize across the response styles and the data sets. When reviewing the results, it will become apparent that H&T1 is exceptional in several regards, so the reader is cautioned not to focus exclusively on this first data set. In Figure 5-3 the results of the nested model comparisons are presented. To read this figure, one should start from model A1. From there, it was tested whether the imposition of additional constraints led to a significant

[9] In particular, in the H&T data, 3 out of 4 AR coefficients were non-significant at the .05 level (i.e. t-values under 1.96) for the ARS model, as were 2 out of 4 in the ERS model. In the latter model, one coefficient was (non-significantly) negative. All factor loadings were significantly positive, with one exception (which had a t-value of 1.93). In the primary data, all factor loadings were highly significant, while 9 out of 16 autoregressive coefficients were non-significant at the .05 level, of which 4 were negative.

chi square difference test (significant difference at the .01-level depicted in light grey) and to a substantial increase in CFI (difference test larger than .05 depicted in dark grey). Note that these results are clearly in line with the RMSEA plots (Figure 5-2).

TABLE 5-2: MODEL FIT INDICES FOR H&T DATA

	NARS H&T1				p(χ^2 diff)		ERS H&T1				p(χ^2 diff)	
Model	df	χ^2	P	CFI	RMSEA	Within CF [a]	Within AR [b]	df	χ^2	P	CFI	RMSEA	Within CF [a]	Within AR [b]
A1	1	0.70	0.403	1.000	0.000			1	0.32	0.572	1.000	0.000		
A2	4	2.87	0.580	1.000	0.000	0.538		4	7.25	0.123	0.997	0.061	0.074	
A3	5	43.67	0.000	0.959	0.188	0.000		5	45.49	0.000	0.966	0.192	0.000	
B1	5	4.76	0.446	1.000	0.000		0.398	5	11.42	0.044	0.995	0.077		0.025
B2	8	6.20	0.625	1.000	0.000	0.696	0.504	8	12.13	0.145	0.996	0.049	0.871	0.300
B3	9	55.48	0.000	0.951	0.154	0.000	0.019	9	55.43	0.000	0.961	0.153	0.000	0.041
C1	6	49.42	0.000	0.954	0.182		0.000	6	75.07	0.000	0.941	0.229		0.000
C2	9	52.63	0.000	0.954	0.149	0.360	0.000	9	78.10	0.000	0.941	0.187	0.387	0.000
C3	10	959.29	0.000	0.000	0.658	0.000	0.000	10	1190	0.000	0.000	0.734	0.000	0.000

	NARS H&T2				p(χ^2 diff)		ERS H&T2				p(χ^2 diff)	
Model [c]	df	χ^2	P	CFI	RMSEA	Within CF [a]	Within AR [b]	df	χ^2	P	CFI	RMSEA	Within CF [a]	Within AR [b]
A2	1	1.5	0.221	0.999	0.059			1	10.3	0.001	0.981	0.253		
A3	2	3.9	0.144	0.985	0.080	0.124		2	11.3	0.003	0.981	0.179	0.310	
B1	2	1.5	0.475	1.000	0.000			2	10.5	0.005	0.983	0.171		
B2	4	2.8	0.592	1.000	0.000	0.519	0.729	4	10.8	0.029	0.986	0.108	0.852	0.912
B3	5	7.3	0.199	0.994	0.056	0.034	0.330	5	12.7	0.026	0.984	0.103	0.166	0.701
C1	3	31.9	0.000	0.920	0.258		0.000	3	53.6	0.000	0.899	0.341		0.000
C2	5	32.3	0.000	0.925	0.194	0.839	0.000	5	53.8	0.000	0.883	0.259	0.905	0.000
C3	6	368.8	0.000	0.000	0.646	0.000	0.000	6	505.4	0.000	0.000	0.758	0.000	0.000

[a] Within CF refers to model comparisons for which the common factor specification remains identical; these models share the same letter, but are denoted with different numbers. [b] Within AR refers to model comparisons for which the autoregressive specification remains identical; these models share the same number, but have a different letter. [c] For H&T2, model A1 is not identified (df=-1).

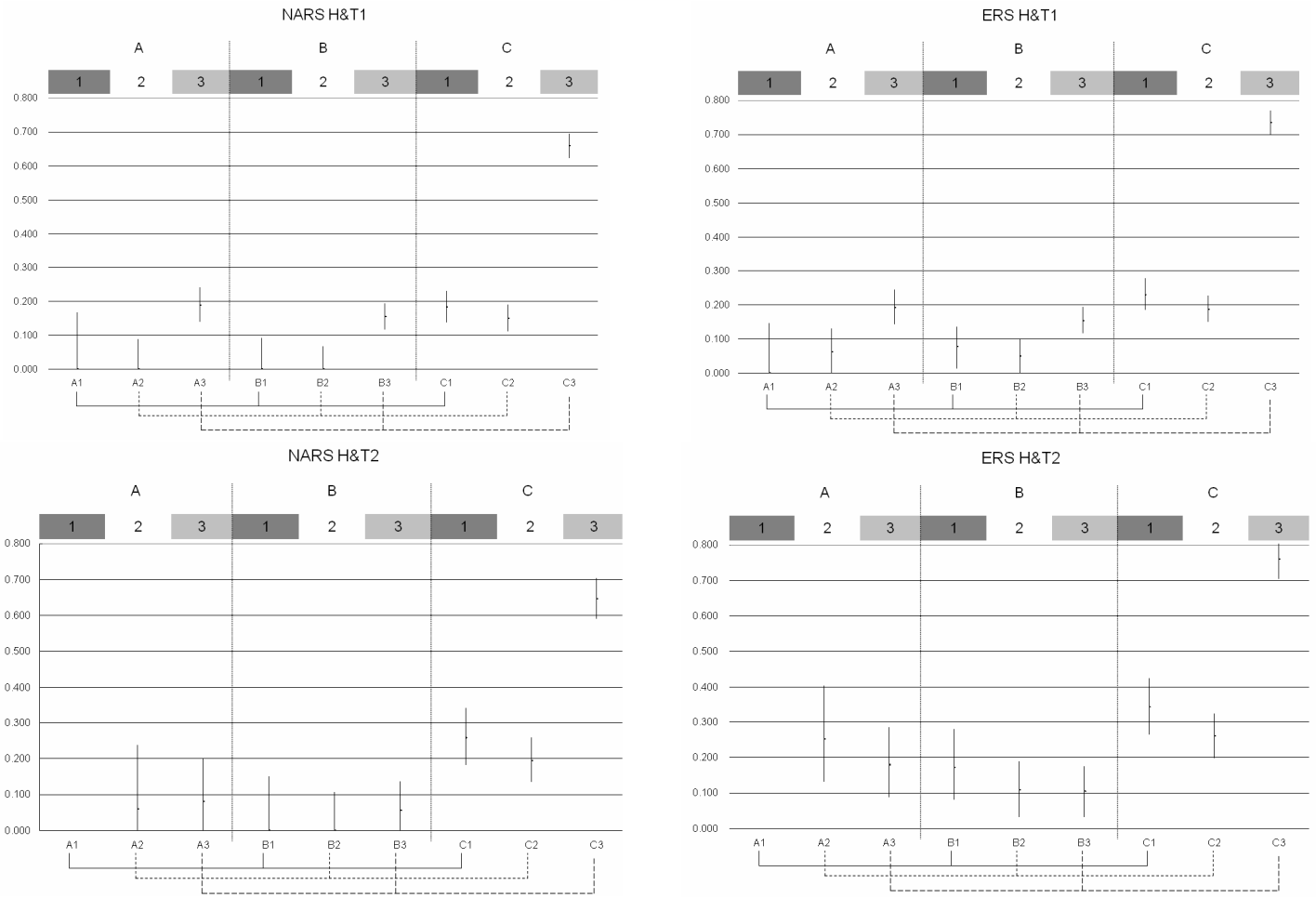
TABLE 5-3 MODEL FIT INDICES FOR PRIMARY DATA

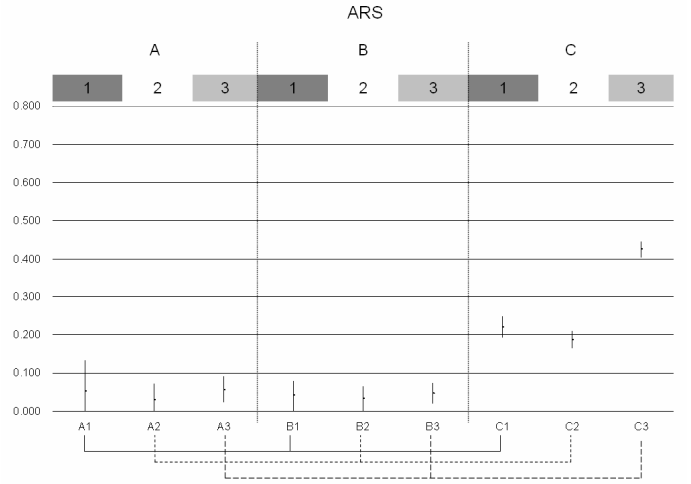
	ARS					p(χ^2 diff)		DRS					p(χ^2 diff)	
Model	df	χ^2	P	CFI	RMSEA	Within CF [a]	Within AR [b]	df	χ^2	P	CFI	RMSEA	Within CF [a]	Within AR [b]
A1	1	2.65	0.103	0.998	0.052			1	0.509	0.476	1.000	0.000		
A2	4	5.88	0.209	0.998	0.028	0.358		4	7.12	0.130	0.997	0.036	0.085	
A3	5	14.52	0.013	0.991	0.056	0.003		5	8.108	0.150	0.997	0.032	0.320	
B1	5	10.21	0.069	0.995	0.042		0.109	5	7.882	0.163	0.997	0.031		0.117
B2	8	13.31	0.102	0.995	0.033	0.377	0.115	8	38.69	0.000	0.968	0.080	0.000	0.000
B3	9	20.59	0.015	0.989	0.046	0.007	0.194	9	51.7	0.000	0.956	0.089	0.000	0.000
C1	6	180.70	0.000	0.838	0.220		0.000	6	203.5	0.000	0.795	0.234		0.000
C2	9	198.22	0.000	0.824	0.187	0.001	0.000	9	216.8	0.000	0.784	0.196	0.004	0.000
C3	10	1091.60	0.000	0.000	0.424	0.000	0.000	10	976.5	0.000	0.000	0.400	0.000	0.000

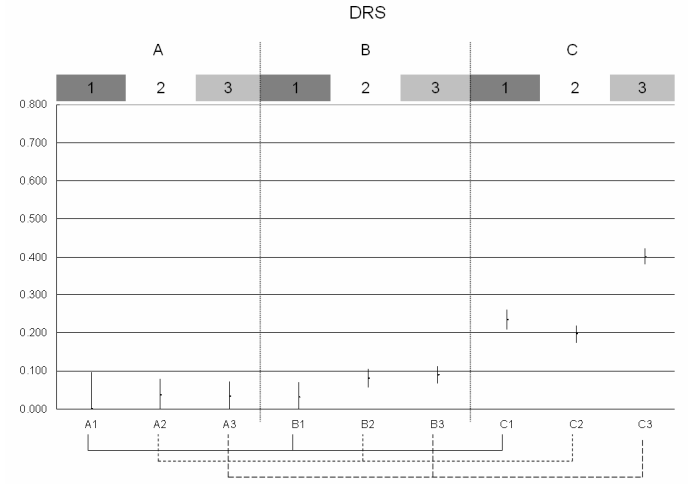
	ERS					p(χ^2 diff)		MRS					p(χ^2 diff)	
Model	df	χ^2	P	CFI	RMSEA	Within CF [a]	Within AR [b]	df	χ^2	P	CFI	RMSEA	Within CF [a]	Within AR [b]
A1	1	0.053	0.817	1.000	0.000			1	0.661	0.416	1.000	0.000		
A2	4	2.621	0.623	1.000	0.000	0.463		4	3.443	0.487	1.000	0.000	0.426	
A3	5	25.15	0.000	0.991	0.082	0.000		5	36.03	0.000	0.984	0.101	0.000	
B1	5	3.778	0.582	1.000	0.000		0.445	5	10.42	0.064	0.997	0.042		0.045
B2	8	18.52	0.018	0.995	0.047	0.002	0.003	8	44.18	0.000	0.981	0.087	0.000	0.000
B3	9	39.99	0.000	0.986	0.076	0.000	0.005	9	154.2	0.000	0.925	0.164	0.000	0.000
C1	6	266.6	0.000	0.885	0.268		0.000	6	199.7	0.000	0.900	0.231		0.000
C2	9	273.8	0.000	0.884	0.221	0.066	0.000	9	214.4	0.000	0.894	0.195	0.002	0.000
C3	10	2288	0.000	0.000	0.615	0.000	0.000	10	1954	0.000	0.000	0.568	0.000	0.000

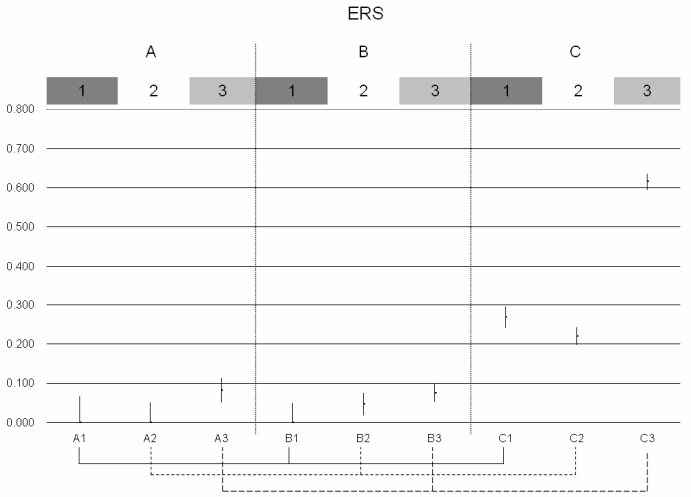
[a] Within CF refers to model comparisons for which the common factor specification remains identical; these models share the same letter, but different numbers. [b] Within AR refers to model comparisons for which the autoregressive specification remains identical; these models share the same number, but have a different letter.

Figure 5-2: RMSEA confidence intervals




ARS
A
B
C
1 2 3 1 2 3 1 2 3
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
A1 A2 A3 B1 B2 B3 C1 C2 C3


DRS
A
B
C
1 2 3 1 2 3 1 2 3
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
A1 A2 A3 B1 B2 B3 C1 C2 C3


ERS
A
B
C
1 2 3 1 2 3 1 2 3
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
A1 A2 A3 B1 B2 B3 C1 C2 C3

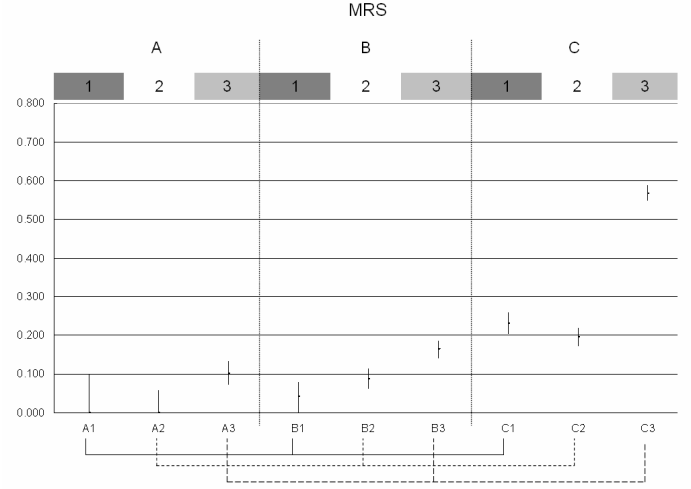
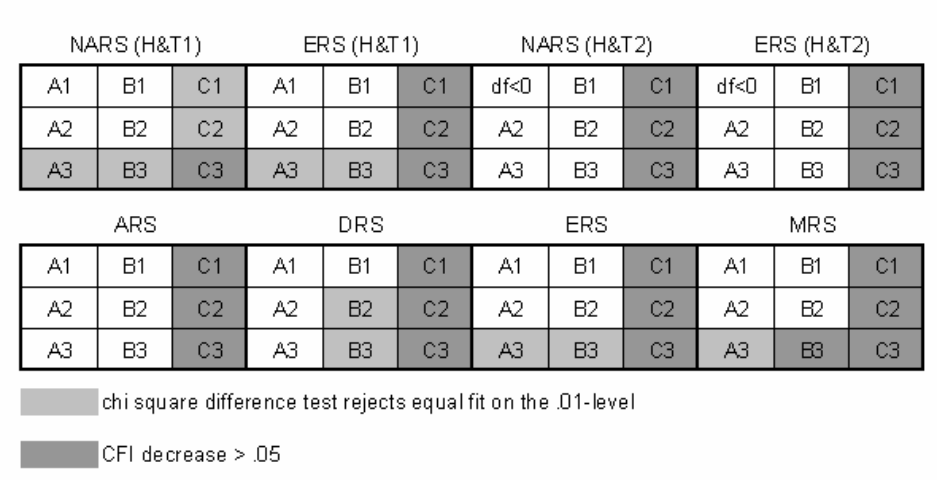

MRS
A
B
C
1 2 3 1 2 3 1 2 3
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
A1 A2 A3 B1 B2 B3 C1 C2 C3

Figure 5-3:

Graphical summary of model fit evaluation based on chi square and CFI



First and foremost, all C models, i.e. models that assume no common factor, fit the data rather poorly, both in the H&T and the primary data. From the perspective of absolute fit, this is evidenced by the chi square tests that were consistently significant at the .001 level, the RMSEA's that were consistently above .100 and the CFI's that were almost consistently below .95 (NARS H&T1 model C1 and C2 were the sole exception to the latter rule). Additionally, from a relative fit perspective, moving from any model B to its C counterpart, which corresponds to constraining the common factor loadings to zero, resulted in a significant and substantial deterioration of fit. All chi square difference tests between any B model and its C counterpart were significant at the .01 level (see the three bottom right cells of each sub table in Table 5-2 and Table 5-3). The decrease in CFI was at least .05 (with the exception of a .046 decrease for NARS model C1 and C2 in H&T1). Finally, the RMSEA confidence intervals clearly show a disparity between C and B models, with C models having substantially larger misfit relative to their degrees of freedom. It is reasonable to conclude from these findings that response styles in different sets of items in the same questionnaire

share a common factor. When measuring response styles, neglect of this factor will lead to serious model-data misfit. Thus, the current findings convincingly show the presence of a stable component to response styles.

Evidence in support of the autoregressive component of response styles is less unequivocal. Some of the A3 and B3 models (in which the autoregressive coefficient is constrained to zero) showed acceptable levels of fit: while only a few chi square tests were non-significant, most CFI's were above .95, and several RMSEA's were below .08 (some below .05; see Table 5-2 and Table 5-3). From a nested model comparison perspective, only the data from H&T1 provided strong evidence of a significant and substantial decrease in fit when the autoregressive coefficient was constrained to zero, as apparent from the significant chi square difference when moving from A2 to A3 or from B2 to B3 (see the 'within CF' column in Table 5-2 and Table 5-3), as well as the CFI decrease of over 5 percentage points when imposing the same constraints. The MRS and ERS models based on the primary data show a similar but less pronounced pattern. Here, the chi square difference tests were significant and the RMSEA increased notably, but the decrease in CFI was smaller than .05 (with the sole exception of the move from B2 to B3 for MRS). This seems to indicate that the common response style factor in these cases can be complemented with an autoregressive component. In the remainder of the data sets (ARS and ERS in H&T2; ARS and DRS in primary data) the autoregressive coefficient did not seem to add to the validity of the model. Where present, constraining the autoregressive coefficient to be constant across time seems granted, based on an evaluation of absolute and relative model fit (Table 5-2 and Table 5-3), and a comparison of the coefficients (Table 5-4 and Table 5-5, discussed below). For now, it seems safest to conclude that a time-invariant autoregressive effect may be present in some response

styles in some data sets, while usually a common factor suffices to account for the shared variance between response style indicators.

EVALUATION OF MODEL ESTIMATES

In addition to an evaluation of overall model fit, the relative value of autoregressive versus common factor specifications is evaluated by investigating the parameter estimates. Since model A2 and B1 showed acceptable fit for all data sets, the estimates of these models were used to evaluate the relative contribution of the common factor and autoregressive components to understanding response styles. The estimates are summarized in Table 5-4 and Table 5-5.

The major trend that emerges from these estimates is in accord with the findings based on model fit: the loadings on the common factor were larger in size and more consistently significant than were the autoregressive coefficients. Only in dataset H&T1 were all autoregressive coefficients significant at the .05-level when estimated freely (i.e. in model B1). For model B1 in dataset H&T2, only one out of six autoregressive coefficients was significant at the 0.05 level. In the primary data, 11 out of 16 of these coefficients were significant. This is most consistently the case for MRS. Taken over all analyses, the average standardized factor loading was 0.71; the average standardized autoregressive coefficient was 0.15.

TABLE 5-4: PARAMETER ESTIMATES FOR MODEL A1 AND B1 (H&T DATA)

Model		NARS H&T1			ERS H&T1			NARS H&T2			ERS H&T2		
		Est.[a]	s.e.	t-value	Est.[a]	s.e.	t-value	Est.[a]	s.e.	t-value	Est.[a]	s.e.	t-value
A2	λ1	0.72	0.06	11.59	0.78	0.06	13.40	0.76	0.07	10.27	0.84	0.07	12.40
	λ2	0.60	0.07	8.80	0.63	0.07	9.81	0.78	0.10	7.61	0.81	0.11	7.52
	λ3	0.60	0.07	8.12	0.67	0.07	10.27	0.74	0.11	7.04	0.83	0.11	7.72
	λ4	0.62	0.07	9.07	0.64	0.07	9.57	0.70	0.10	7.26	0.79	0.10	0.79
	λ5	0.59	0.07	8.90	0.67	0.06	10.99						
	β	0.31	0.05	5.64	0.28	0.05	5.58	0.12	0.08	1.44	0.09	0.09	0.97
B1	β21	0.28	0.05	5.68	0.24	0.05	5.26	0.07	0.07	1.03	0.08	0.06	1.48
	β32	0.23	0.05	4.68	0.21	0.04	4.97	0.11	0.07	1.53	0.05	0.06	0.97
	β43	0.27	0.05	5.62	0.21	0.04	4.74	0.15	0.07	2.29	0.06	0.06	1.02
	β54	0.24	0.05	4.75	0.22	0.04	5.34						
	λ	0.66	0.05	13.54	0.72	0.05	15.16	0.75	0.07	11.12	0.83	0.06	12.95

[a] Since the model was based on a correlation matrix, the estimates are standardized. Est. = Estimated parameter value; s.e. = standard error; λ = factor loading; ß = autoregressive coefficient.

TABLE 5-5: PARAMETER ESTIMATES FOR MODEL A1 AND B1 (PRIMARY DATA)

Model		ARS				DRS				ERS				MRS			
		Est.	s.e.	t-value	Stand. est.	Est.	s.e.	t-value	Stand. est.	Est.	s.e.	t-value	Stand. est.	Est.	s.e.	t-value	Stand. est.
A2	λ1	0.25	0.02	16.38	0.69	0.19	0.01	15.20	0.65	0.16		23.08	0.84	0.11	0.01	19.41	0.76
	λ2	0.21	0.02	12.31	0.66	0.26	0.02	16.26	0.76	0.15		15.86	0.74	0.10	0.01	14.14	0.67
	λ3	0.24	0.02	15.01	0.74	0.26	0.02	15.61	0.76	0.15		15.79	0.75	0.13	0.01	9.36	0.73
	λ4	0.24	0.02	13.78	0.69	0.20	0.02	12.81	0.71	0.13		14.44	0.72	0.12	0.01	16.09	0.72
	λ5	0.21	0.02	12.74	0.62	0.19	0.01	13.97	0.67	0.15		17.16	0.75	0.12	0.01	15.37	0.73
	β	0.10	0.03	2.90	0.10	0.03	0.03	0.99	0.03	0.16		4.44	0.16	0.19	0.04	5.26	0.18
B1	β21	0.05	0.04	1.39	0.05	0.13	0.04	3.13	0.12	0.14		4.50	0.13	0.12	0.04	3.40	0.11
	β32	0.12	0.04	3.07	0.12	0.21	0.04	5.64	0.20	0.13		4.74	0.13	0.22	0.03	6.26	0.20
	β43	0.09	0.04	2.15	0.09	0.02	0.03	0.55	0.02	0.03		1.22	0.04	0.27	0.03	8.13	0.25
	β54	0.06	0.04	1.54	0.06	-0.01	0.04	-0.28	-0.01	0.09		2.95	0.09	0.34	0.03	10.69	0.31
	λ	0.23	0.01	20.59	0.69	0.21	0.01	20.44	0.67	0.16		24.95	0.81	0.11	0.01	22.53	0.69

Est. = Estimated parameter value; s.e. = standard error; Stand. est. = Standardized estimates. λ = factor loading; ß = autoregressive coefficient

Finally it is noted that the amount of explained variance in the response style indicators remained constant when the autoregressive coefficients were set to zero (when moving from model A2 to A3). Specifically, the average[10] indicator R squared remained at 0.71. On the other hand, when constraining the common factor loadings to zero (i.e. moving from model B1 to model C1), the average indicator R squared dropped from 0.71 to 0.55. Note that these results should not be considered a decomposition of variance components, but a comparison of the ability of different types of models to explain a certain portion of variance in the observed variables while optimizing model-data fit.

TAU EQUIVALENCE

While the main focus of the current study is on the presence versus absence of a common factor and an autoregressive component in response styles, the results can be read in a similar way to assess the validity of the tau-equivalence hypothesis. This is especially relevant given the major significance of a common response style factor; the question now becomes how constant its effect is. Without going into details, it is concluded that an assessment of absolute and relative model fit as well as the loading estimates (Table 5-2, Table 5-3, Table 5-4, and Table 5-5) indicates that tau-equivalence may be a reasonable assumption in most of the data, with the exception of MRS and DRS in the primary data.

[10] The first response style indicator was not included in the evaluations of R squared (both in the A2-A3 and the B1-C1 comparison) because in the autoregressive models its explained variance is zero by design.

DISCUSSION

THE STABILITY OF RESPONSE STYLES

The current research provides convincing support for the notion that response styles share a common factor which is stable across sets of items in the same questionnaire, even when these items are not related to one another in terms of content. Remarkably, it is found that not only primary data, but also the data brought forward by Hui and Triandis (1985) indicated the presence of a stable common factor that showed good model fit as well as significant and high factor loadings and that explained a good deal of the variance in response style indicators. The autoregressive component was less significant and substantial, especially in light of the observation that the remarkable pattern in H&T's data set 1 (H&T1) might have been the direct reason for postulating the instability hypothesis and could hence hardly be considered a fair test of the same. Also note that H&T1 concerned object ratings on 10-point scales, which set it apart from the other data and which may invalidate generalization from these data to response styles in more common data, like five and seven point Likert items. In particular, H&T1 concerned stimulus-centered rather than respondent-centered scales and used a suboptimally high number of response alternatives (Cox 1980). Nevertheless, it is clear that even if an autoregressive component is present in the response style data, it operates in addition to a common underlying factor, rather than alone. Moreover, the autoregressive component of response styles compares rather faintly to the effect of a common factor, both in terms of model fit and effect size.

IMPLICATIONS FOR RESEARCH

The current findings indicate the presence of systematic response style bias in self-reports using closed-ended questions. More specifically, it was found that random sets

of contentwise unrelated items share stable response style variance, in that respondents show systematic differences in their preference for positive (ARS), negative (DRS), extreme (ERS) or middle (MRS) response options. On the positive side, the observed stability of response styles implies the possibility of constructing reliable and valid measures of the same. It is therefore recommended to researchers to include such measures in research designs when using questionnaire data. The current study offers guidelines to construct measures of ARS, DRS, MRS and ERS in a structural equation modeling framework, where random sets of items from heterogeneous item domains are used as the basis for response style indicators. Such procedure has not been commonly implemented yet to measure response styles, though it would offer important benefits (Podsakoff et al. 2003). First, it allows for methodical model comparisons, addressing the question of stability, or in particular the presence of a common factor and/or an autoregressive component as well as their respective tau-equivalence and time invariance. This issue cannot be addressed by coefficients of internal consistency or split-half correlations. Second, it allows for further evaluation of measurement models in terms of discriminant and convergent validity (Fornell and Larcker 1981) as well as the assessment of measurement invariance across different groups of respondents like different modes of data collection, cultural groups, etc. (Little 1997; Cheung and Rensvold 2002). In the methods used to measure response styles in the literature, such measurement issues seem to have been taken for granted, while there is little reason to treat response style measures differently than any substantive measure in this regard.

Based on the current findings, it is suggested that response styles are best modeled as a congeneric or tau-equivalent common factor with or without a time invariant autoregressive effect (i.e. model A2, A3, B2 and B3). These models quite consistently

showed good model fit (in absolute and relative terms) combined with theoretically sound estimates for the factor loadings and the autoregressive coefficient. The choice between congeneric and tau-equivalent models as well as the choice between autoregressive and non-autoregressive models can ideally be based on model comparisons as those presented here. For stand-alone models of response styles as used in the current study, it is recommended to use at least 4 indicators of response styles, such that models A2 and A3 are identified and can be compared. In more extended models, it may be desirable to use 3 indicators, since this number of parcels allows for stable yet efficient estimation of the factor variance and loadings (Little, Cunningham and Shahar 2002).

THE MEANING OF RESPONSE STYLES

While the observation that response styles are largely stable is important in and of itself, it is relevant to dwell on the implications it has for the meaning of response styles. In other words: does the short term stability lend support to or does it invalidate specific theories of response styles? First, short term stability makes long term stability a theoretical possibility. That is, current findings do not contradict the interpretation of response styles as a learned behavior or even a trait (Hamilton 1968). Nevertheless, short term stability in this case is a necessary but insufficient condition for long term stability. What can be concluded is that response styles most probably have at least one cause that is stable over the period of filling out a questionnaire. Other than causes that are stable over the long run, some of the possibilities that might merit consideration are moods (see e.g. Schwarz 1997 for mood effects on the content level); anchoring of the scale meaning on specific response options (Marsh and Parducci 1977); and fatigue, (de)motivation and the resultant cognitive effort that is expended (Krosnick 1991). Note that each of these origins of response styles may

evolve over the questionnaire but can be reasonably expected to be rather constant over its course for most respondents. However, it is relevant to consider in more detail the plausible evolution over time of such causes and their effect on response styles. How mood will evolve is hard to predict and probably depends on a complex interaction of initial mood and questionnaire content. In the current study, initial mood was not controlled for and content was highly diverse; in other settings, however, it might be worth considering its impact.

Anchoring of a response scale here refers to assigning meaning to the response options by relating the extremes or other salient response options to specific reference stimuli to which the stimulus to be assessed can then be compared (Marsh and Parducci 1977; Parducci 1974). Since respondents typically keep in mind the last 10 to 20 stimuli as a reference (Wedell and Parducci 1988), anchoring can be expected to lead to response styles that gradually move over the course of a questionnaire. Empirically, such process would translate in an autoregressive effect. Anchoring is most relevant in situations where stimuli (commonly objects, but subjective states, values, etc. are also possible) are rated along a limited set of dimensions. This is consistent with the fact that the autoregressive effect was observed most strongly in H&T1, where objects were rated on 3 dimensions using ten-point rating scales.

In addition to the above, another process might result in an autoregressive pattern in response styles. The fact that a respondent selects a particular option will lead this option to be more accessible in memory afterwards. This might subsequently increase the probability of this same option being selected in answering the following items. While there is no reason to suspect that some response options (e.g. the extremes) would be more vulnerable to such effect, there is an indirect reason to suspect a stronger impact on certain response style measures. In particular, response styles that

are limited to a single response option will be most affected, followed by a response style defined by two response options. This is exactly what seems to be happening in the primary data set: MRS shows the strongest autoregressive effect, followed by ERS, while ARS and DRS show no autoregressive effect. Since this hypothesis is formulated post hoc, further investigation is necessary.

Finally, respondent fatigue and the related decrease in motivation and effort is a completely different matter than anchoring and accessibility. In the latter two processes (responses to) items in the questionnaire have an impact on the subsequent response style level. In the case of fatigue, however, it is usually assumed that a more autonomously driven process occurs: respondents 'grow' tired regardless of the specific stimuli rated or the specific responses given, suggesting that a latent growth model would be in place here. Other than autoregressive models, latent growth models estimate the gradual evolution of average and individual levels of a continuous variable (in this case response styles). Autoregressive SEM models do not necessarily include a mean/score component but focus on second-order moments, and merely imply that a respondent's relative position on a variable at time t is predictive of her/his relative position on this variable at time t+1 (Curran and Bollen 2001). Unfortunately, since each item had a unique position in the H&T and the primary data, it makes little sense to look for an evolution in mean or individual scores over the length of a questionnaire. To do this, one needs the assumption that the response style indicators would show identical means after controlling for position, an assumption that is not needed when using pure second-order moment based models as was done in the current study. It would therefore be interesting to investigate this matter based on data that have identical items in different positions within the

questionnaire. Findings by Kraut, Wolfson and Rothenberg (1975) suggest that one might expect an increase in MRS and a decrease in ERS over time.

CONCLUSION

Response styles in subsequent sets of contentwise unrelated items within a questionnaire are to a large extent caused by a common factor. Whether the relation to this factor is identical across sets (i.e. whether tau-equivalence holds), needs to be established for each data set, but in most cases this seems to be a valid assumption. In specific data sets the effect of the common factor needs to be complemented by an autoregressive effect. While the current findings are not conclusive in this regard, the autoregressive component may be strongest if respondents rate objects on a limited set of dimensions using rating scales with a high number of response options. Also, an autoregressive component may be present in general for ERS and MRS indicators. If present, the autoregressive coefficient can be reasonably expected to be time invariant in most cases.

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

In addition to the limitations and directions for future research touched upon in the discussion, three more such topics deserve discussion, related to the format of the items studied, the testing approach and the scope of stability.

First, the primary data studied made use of seven point Likert items only. As also noted by Greenleaf (1992a) and Baumgartner and Steenkamp (2001) it would be interesting to study how the use of different scale formats (e.g. five point scales, etc.) would affect response styles.

Second, the current study centered on an approach that does not and cannot result in a single test of significance and a single yes or no answer. An attempt was made to

ensure validity by making a balanced evaluation of a set of relevant criteria rather than relying on one decision rule. While this approach may be seen as lacking clarity by some, it appears the best way to guarantee meaningful results rather than one-time significant results. As Marsh et al. (2004) pointed out, although it would be nice to have 'golden rules' that provide researchers with definite and clear answers, there is no alternative to immersing oneself in the data and making well considered choices based on a combination of observations. In this study, this combination consisted of stand-alone model fit evaluation, nested and more broadly comparative fit evaluation, and assessment of model estimates, linked to a thorough search for theoretical views on response styles that were then translated into specific operational and testable models.

Moving beyond the time frame of a single data collection, it would be highly relevant to assess the long term stability of response styles. The short term stability of response styles enables the construction of measures of response styles that can be used to correct items in one and the same questionnaire. Similar measures that are valid and reliable over the long term would offer huge potential for improving the quality of panel data. If response styles prove to be sufficiently stable, measures could be constructed for members of data collection panels and included as default covariates in analyses. This would substantially decrease the risk of drawing conclusions driven by respondents' differences in reacting to questionnaire items rather than the content one intended to measure.

CHAPTER 6: THE LONG TERM STABILITY OF INDIVIDUAL RESPONSE STYLES (EMPIRICAL STUDY 3)

CHAPTER OUTLINE

The level of stability of response styles co-determines how strongly they may bias estimated self-report measures over time and/or the same measures' relationships with stable background variables. The current study investigated the stability of response styles based on data from the same respondents who filled out two questionnaires consisting of independent sets of random samples of questionnaire items. Between data collections, there was a one year time gap. The results provide convincing evidence that response styles have an important stable component, only a small part of which can be explained by demographics. The meaning and implications of these findings are discussed.

INTRODUCTION

Respondents to questionnaires have been found to show varying levels of response styles in their responses to closed-ended items (Greenleaf 1992a; Johnson et al. 2005). Regardless of content, individuals differ in their tendency to disproportionately use positive response options (acquiescence response style or ARS), negative response options (disacquiescence response style or DRS), midpoint response options (midpoint response style or MRS) and extreme response options (extreme response style or ERS). Consequently, item responses are a mixture of content and style. Since response styles cause consistency in individuals' responses, their presence leads to spurious correlations between item responses, and consequently, to overestimation of reliability (Green and Hershberger 2000). This is the case if reliability is assessed by estimating internal consistency as well as when it is assessed by estimating test-retest stability. Additionally, if response styles are stable personal characteristics, they lead to misestimation of the variances and covariances of self-report measures of variables (Baumgartner and Steenkamp 2001). If response styles are stable and systematically related to background variables like demographics, they also cause the misestimation of covariances of self-report measures with these background variables (Greenleaf 1992a). While the stability of response styles would be problematic in that it causes bias in results, it would also have its positive side. Specifically, if response styles are stable individual characteristics, this would offer interesting opportunities for correcting for them in panel research: once measured, response style indicators could be used as default covariates in later analyses to statistically correct for their effect. Given the above, it is of major importance to know to what extent response styles are stable within an individual over time. This question calls for an adequate research design meeting the following requirements. First, panel data with responses of the

same identifiable respondents to at least two questionnaires are needed. The data collections need to be separated far enough in time to ensure that transient influences (like mood, current life events, etc.) can be safely assumed not to be constant across the two situations. Moreover, to ensure that the stability of their responses is due to style and not to content, the questionnaires need to consist of different, independent sets of items, each of them consisting of a variety of unrelated items (Greenleaf 1992b). While the items should be heterogeneous in content, they should use the same format to be able to assess consistency in the response options selected. Such design is used in the current study to assess the stability of ARS, DRS, MRS and ERS over a one year gap in time.

CONCEPTUAL FRAMEWORK

There are two streams of research that are relevant in assessing the long term stability of response styles. First, some research links response styles to stable personal characteristics (on a cross-sectional basis), a link that logically implies a stable component to response styles. Second, though suffering from limitations in scope and methodology, some longitudinal research has been reported on response styles. Before discussing these studies, Rorer's (1965) influential critique on the response style literature is reviewed, since it will help clarifying some of the requirements that need to be met to assess response style stability.

RORER'S (1965) CRITIQUE OF THE RESPONSE STYLE LITERATURE

Based on his highly critical review of the literature, Rorer (1965) concluded that response styles are a myth. Up till 1965, no evidence seemed to have been provided that proved the existence of respondents' tendency to select some response category a disproportionate amount of the time independently of the item content. As Rorer

pointed out, showing that a response consistency exists when related or identical measures are answered twice does not necessarily imply the presence of response styles. To establish the existence of a stable response style, one needs to operationalize such response style as a stable tendency that applies to independent heterogeneous sets of items. Later research seems to have established the presence of response styles that at least generalize across different content domains (e.g. Bachman and O'Malley 1984; Baumgartner and Steenkamp 2001; Greenleaf 1992a, b; Ray 1979; Paulhus 1991). Evidence for long term temporal stability remains sparse if not non-existent, however, and even the short term stability of response styles has been questioned (Hui and Triandis 1985).

Basically, a distinction can be made between two major types of evidence in support of response style stability (Hamilton 1968). First, explicit test-retest investigations would provide direct evidence of temporal stability. Second, relations of response styles to stable personal characteristics indicate that at least the variance shared with these background variables is stable.

LONGITUDINAL STUDIES

Since the current study concerned long term stability, evaluations of reliability based on test-retest correlations and internal consistency between parts of the same cross-sectional data collection were less relevant (Baumgartner and Steenkamp 2001; Greenleaf 1992b; Hui and Triandis 1985). Hamilton (1968) listed several studies that assessed test-retest reliability of response styles across different data-collections. However, the time gap between test and retest ranged from 1 to 4 weeks only and, most importantly, in all cases the same questionnaire was used for both data collections. This makes it impossible to distinguish between style and content (Rorer 1965) and to rule out the possibility of artificial consistency (Feldman and Lynch

1988). Greenleaf (1992b) found that the aggregate distribution of ERS was stable over time. Unfortunately, the data did not allow an assessment of ERS stability on the individual level. Bachman and O'Malley (1984) did have longitudinal measures of response styles at the individual level. The authors found very high stability estimates for ARS and ERS: after taking into account (non-)reliability, the estimates of annual stability matched or exceeded those obtained for other common personal variables in the social sciences. However, here too content related consistency cannot be excluded as an alternative explanation of the stability, in that the stability coefficients were computed using repeated administration of the same sets of items. Also, the authors stressed that the items used for the study could be thought of as "*samples of agree-disagree items, but they are far from random samples*" (p. 502). Similar limitations apply to the interesting work by Motl and DiStefano (2002) and Horran, DiStefano and Motl (2003), in which the authors showed that method effects associated with negatively worded items in a self-esteem scale showed longitudinal invariance when the same scale was administered repeatedly to the same sample. Importantly, in this context, some research has suggested that retest effects may be present even when retest intervals are long (Ferrando 2002).

In sum, evidence on longitudinal stability of response styles, while thought provoking, is suggestive rather than conclusive, given the fact that content has not been controlled for in studies assessing the stability of response styles.

RELATIONS OF RESPONSE STYLES TO BACKGROUND VARIABLES

Complementing research that has tried to assess the longitudinal stability of response styles, some studies have documented relations between response styles and stable individual characteristics. Such relations, even if established cross-sectionally, would imply that the portion of variance a response style shares with a stable individual

variable is stable itself. Two such stable individual variables have been considered: (1) observable variables such as social demographics; (2) latent variables such as personality traits.

Demographics

In the literature on response effects and biases, the two most relevant demographics are age and education, the reason being that both have been related to cognitive functioning (Schuman and Presser 1981; Krosnick 1991; Knauper 1999). Education level is related directly to cognitive sophistication, in that people with higher cognitive sophistication may get higher levels of education, and that higher levels of education expose people more extensively to cognitive tasks and formalized ways of thinking (Krosnick 1991). In line with this, McClendon (1991b) hypothesized that lowly educated respondents are more readily influenced by cognitive mechanisms leading to ARS. The hypothesized effect could not be confirmed, according to the author most probably due to a faulty manipulation (McClendon 1991b). In another study, McClendon (1991a) did observe a negative relation between education level and ARS. Further, in a meta-analysis of the prominent Schuman and Presser (1981) studies, Narayan and Krosnick (1996) found evidence for an education effect on a wide range of response biases, including the levels of ARS, which were higher among the lowly educated. From their results, the authors concluded that respondents with lower levels of education were more likely to satisfice, i.e. to provide a satisfactory rather than an optimal response to the questions in a questionnaire. This also concurs with the early observation by Osgood (1941) that lowly educated respondents tend to simplify the task of responding to seven-point semantic differentials by only selecting the extremes and midpoints of the scale, leading to a trimodal (or even trichotomized) response distribution. Greenleaf (1992a) observed a negative relationship of both ARS

and ERS with education level. Marín, Gamba and Marín (1992) also found support for the negative association of ERS and education level.

Knauper (1999) showed that, while education may show significant relations to several response biases and effects, it is crucial to control for age in such analyses. While admitting that education may be related to cognitive sophistication, Knauper pointed out that in general education is negatively related to age, and that the observed relations may at least in part be due to a spurious effect. Age might well be the real explanatory variable, since increasing age is associated with a gradual decline in working memory capacity, which may make older respondents more prone to response effects and biases caused by cognitive limitations. Marsh (1996) found that method effects associated with negatively worded items are related both to age and verbal ability. Also, Mirowsky and Ross (1991) observed that ARS is related both to age and education. Both Marsh (1996) and Miroswky and Ross (1991) specified the function relating the response effects to age as a U form, where the effect declines from childhood to adolescence and then increases again at later ages. Hamilton (1968) posited a similar association for ERS stating that younger and older respondents have higher ERS levels. In a sample representing only the adult population, Greenleaf (1992a) found a positive relation between age and both ARS and ERS.

While age and education are considered the most relevant demographic antecedents of response styles by far, several researchers have stressed the importance of including gender as a covariate in studying response biases (Becker 2000; Hamilton 1968). Hamilton (1968) explicitly stated that response style research should always control for gender, since it has been found that females show higher levels of ERS. While there is no clear rationale for this finding, it is sufficiently consistent to consider it a potentially valid effect. Nevertheless, Greenleaf (1992a) found that females have

lower levels of ARS, but his data did not confirm the relation between gender and ERS.

Most commonly, researchers have not made the distinction between ARS and DRS, but have considered them as the opposite poles of the same underlying response style (e.g. Greenleaf 1992a; Cheung and Rensvold 2000). However, Bachman and O'Malley (1984) indicated the importance of investigating the relationship between ARS and DRS, since the two were related positively rather than negatively in their data. While the literature provides little base for formulating directed hypotheses on how DRS relates to demographics, there are clear indications that DRS is assumed to be higher among the highly educated, since the highly educated are expected to more thoroughly evaluate statements and also consider counter-evidence in this evaluation (Schuman and Presser 1981; McClendon 1991b). Taking into account Knauper's (1999) theorizing on the effects of age, it was hypothesized that DRS also is negatively related to age.

MRS has been studied rather sparsely. Often it is not relevant since even numbers of response options are used (Bachman and O'Malley 1984). At other times it is considered the opposite of ERS (e.g. Johnson et al. 2005). The available evidence seems to indicate MRS is indicative of respondent stable or transient cognitive limitations (Krosnick 1991; Kraut, Wolfson and Rothenberg 1975; Osgood 1941). In line with the arguments developed in the context of ARS and ERS, this led to the hypothesis that MRS is positively related to age and negatively related to education level.

Although the effects reported in most of the above studies were significant, the effect sizes of the relations often were modest, explaining less than 10% of the observed variance in response styles. Therefore the research question is adapted as follows.

Rather than investigating the presence of stable response style variance, the presence of stable response style variance will be studied in addition to the variance explained by demographics. Including the demographics as control variables will also allow further validation of the findings reported in the literature. It is important to investigate whether response styles have a substantial variance component after controlling for demographics because if this is not the case, it would suffice to discount the demographically caused response style effect from research findings, without further investigation of residual response style variance itself. In other words, controlling for demographics would suffice (for studies where the demographic effect is not the focus).

Latent stable background variables

Next to observable variables such as the above, response styles have been related to latent stable background variables. Hamilton (1968) provided both an overview of relations that have been observed as the main reason why the status of these findings is questionable, in that *"psychometric tests being correlated with ERS measures may themselves be influenced by response styles*" (Hamilton 1968, p. 198; also see Spector et al. 1997 for a similar critique). Moreover, if the measures of response styles and the background variables of interest are collected during the same data collection, both may be subject to common transient factors such as fatigue, cognitive limitations due to worries, etc. (Becker 2000). This would invalidate the presumed time invariance of the background variable measurement. Hence, the presence of a stable component to response styles apart from their variance shared with demographics has not been convincingly shown.

To conclude, the relation of response styles with latent stable individual variables is somewhat uncertain, while the relation with observable stable individual variables is

modest in effect size. If the latter component is the only stable component, this would mean that approximately 90% of response style variance is unstable, rendering untenable the view of response styles as individual trait variables. The question then remains how stable response styles are, and what proportion of their variance is explained by demographics and how much stable variance is present but unexplained. To address this issue, a longitudinal study is conducted consisting of two waves of data collection among the same respondents, each time using a questionnaire consisting of an independent random sample of agree-disagree items.

METHODOLOGY

Respondents were recruited from the panel of an online market research company. The sample was selected to represent a cross-section of the Belgian population in terms of age, gender and education levels. Data were collected in two waves. In between these two waves was a 12 month time lag. The questionnaires in both waves contained independent sets of agree-disagree items, specifically sampled to measure response styles. This method essentially reduced content to random noise, serving two goals at the same time. First, it guaranteed a sample of items representative of the items used in consumer research and applied psychological research. Second, it controlled for content without omitting it altogether.

ITEMS

For wave 1, from the marketing scales handbook by Bruner, James and Hensel (2001), 52 items were randomly selected from different scales. The 52 items had an average inter-item correlation of .07. For wave 2, the sampling frame was extended to not only include the Marketing Scales Handbook by Bruner, James and Hensel (2001), but also Measures of Personality and Social Psychological Attitudes by

Robinson, Shaver and Wrightsman (1991). From these two books 112 items from different scales were randomly selected. These items were put together in an uninterrupted random list making up the complete questionnaire. In this questionnaire, the average inter-item correlation equaled .13. Importantly, the items for wave 1 and wave 2 were independently sampled, resulting in two different sets of items. Hence, response patterns that were the same across both item sets cannot be attributed to the specific items and their content.

RESPONSE STYLE INDICATOR CALCULATION

In both waves, the items were divided into three sets, corresponding to three subsequent parts of the questionnaire. In wave 1, each set consisted of 17 or 18 items. In wave 2, each set consists of 37 or 38 items. In both waves, the three sets were used to compute three indicators for every response style (ARS, DRS, ERS and MRS). For ARS, the number of agreements was counted per set of items, weighting a seven as three points, a six as two points, and a five as one point. A similar method was applied to obtain DRS measures. ARS and DRS measures range from 0 through 3 and can be interpreted as the bias away from the midpoint due to ARS or DRS. If DRS is subtracted from ARS, this indicates the net bias. For example, a respondent with an ARS score of 1.5 and a DRS score of 1 has an expected mean score of 4 + 1.5 – 1 = 4.5 on a 7-point item due to the effect of ARS and DRS. ERS indicators were computed as the number of extreme responses (1 or 7) divided by the number of items. Similarly, MRS indicators were computed as the number of midpoint responses (4) divided by the number of items in the set. ERS and MRS scores can be interpreted as the proportion of respectively extreme and midpoint responses, and hence range from 0 through 1.

DEMOGRAPHICS

In both wave 1 and wave 2 the following demographics were measured. (1) Age was mean centered (mean = 42) and divided by ten to keep the variance in a range similar to that of the other variables in the model. (2) Education level was measured as the number of years of formal education, also mean centered (mean = 12.8). (3) Sex was indicated by a dummy variable, where male = 0 and female = 1.

RESPONDENTS

For the first wave, 3000 panel members of an Internet market research company were contacted. In total, 1758 responses were obtained, 1596 of which were unique respondents. 151 respondents did not finish the questionnaire completely. 1445 cases were retained for further analyses. In this sample, the average age was 42.6 (s=14.7), the average years of formal education equaled 6.77 (s=1.81), and 45.7% of respondents were female.

For the second wave, the 1372 still active panel members (out of 1445 respondents to wave 1) were contacted for participation. Special care was taken to optimize the response to the second wave, in line with recommendations by Deutskens et al. (2004). In total, 633 responses were obtained, of which 604 could be used for further analysis. In this final sample, the average age was 43.2 years (s=14.7), the average years of formal education equaled 6.98 (s=1.94), and 44.0% of the respondents were female. 104 respondents had one or more missing values. A comparison of demographics between respondents and non-respondents in wave 2 is included in the analyses reported below.

ANALYSES AND RESULTS

All analyses were performed using Full Information Maximum Likelihood (FIML) estimation to account for missingness (Enders 2006). Since the degree of non-normality was low (skewness < 2 and kurtosis < 7 for all but one observed variable) and since the alternative (robust) estimators yielded nearly identical results and substantively the same conclusions, the FIML results are reported (Curran, West and Finch 1996; Finney and DiStefano 2006).

The data were analyzed in several steps. First, for each wave separately, a confirmatory factor analysis (CFA) was conducted to assess the convergent and discriminant validity of the response style measurement model (Fornell and Larcker 1981). Second, to test for selectivity, it was investigated whether response style levels in wave 1 were predictive of non-response to wave 2 after controlling for demographics. Third, the focal model for this study was tested, linking the response style factor in wave 1 and wave 2 by a time invariant second order factor for each response style. In this mimic model, the second order response style factors were regressed on sex, age and education.

TIME SPECIFIC CFA'S

First, a CFA model with four factors was specified: ARS, DRS, ERS and MRS. Each response style had three reflective indicators. The unique factors of all first indicators of each of the four response styles were correlated. The same was done for the second and the third indicator of all response styles (Weijters, Schillewaert and Geuens 2005). This CFA model was fitted to the data for each wave separately.

In wave 1, all observed variables had skewness less than 2 and (excess) kurtosis less than one. The chi square test indicated significant misfit, $\chi^2(30, N=1573)=119.12$ ($p<.001$). The alternative fit indices showed good values, however (CFI = .995; TLI =

.988; RMSEA = .043), and the indices of local misfit (modification indices and standardized residual covariances) showed no systematic pattern. Therefore, it was decided to accept the model and its estimates as providing valid approximations of the data. As shown in Table 6-1, an evaluation of the factor loading estimates and factor correlations indicated a valid measurement model. Specifically, all factors had average variance extracted of over .50, indicating good convergent validity, and shared variances that were smaller than their average variance extracted, indicating good discriminant validity (Fornell and Larcker 1981). From Table 6-1, it is apparent that MRS was the most distinct response style, sharing little variance with the others, while ARS, DRS and ERS shared a substantial amount of variance.

TABLE 6-1: SHARED VARIANCE, AVERAGE VARIANCE EXTRACTED AND CORRELATIONS OF RESPONSE STYLE FACTORS

	Wave 1				Wave 2			
SV/AVE/*r*	ARS	DRS	ERS	MRS	ARS	DRS	ERS	MRS
ARS	0.67	*0.51*	*0.74*	*0.01*	0.65	*0.35*	*0.71*	*-0.50*
DRS	0.26	0.58	*0.65*	*0.03*	0.12	0.67	*0.62*	*-0.57*
ERS	0.54	0.43	0.78	*0.06*	0.50	0.39	0.83	*-0.14*
MRS	0.00	0.00	0.00	0.59	0.25	0.32	0.02	0.80

On the diagonals, average variance extracted (AVE) is reported; in the below-diagonal cells, the shared variance (SV, i.e. r^2) is reported (Fornell and Larcker 1981); in the above-diagonal cells, correlations (r) are reported

In wave 2, all but one observed variables had skewness below 2 and kurtosis below 7 (the exception was MRSt2a, kurtosis = 8.48). Since accounting for non-normality did not seem to influence the results to any significant extent, the regular FIML results were reported. While the chi square test was significant ($\chi^2(30, N= 604)=101.98$, $p<.001$), the alternative fit indices showed acceptable levels (TLI=.975; CFI = .991;

RMSEA = .063). Again, as shown in Table 6-1, the factor solution showed good convergent and discriminant validity, especially for MRS and ERS. Here too, all factors had average variance extracted of over .50, indicating good convergent validity, and shared variances that were smaller than their average variance extracted, indicating good discriminant validity (Fornell and Larcker 1981).

RESPONSE TO WAVE 2

It was investigated whether the response styles measured in wave 1 were predictive of response/non-response to wave 2, controlling for demographics. This was done for two reasons. First, in panel research, attrition is inevitable. It is important to investigate whether attrition is selective in such a way that it might bias the findings. Second, if response styles at time 1 would be predictive of response/non-response at time 2, this would suggest that response styles at time 1 were related to respondent motivation to participate in research. Such finding would also be relevant in providing guidelines on when to provide extra incentives for participation.

In order to respect the temporal order, response/non-response in wave 2 was regressed on the four response styles ARSt1, DRSt1, ERSt1 and MRSt1, and the demographics age, education and sex. To do so, a structural equation model was specified with as the independent variables: (1) the response styles modeled as latent variables as done in the CFA described above, freely covarying with (2) the demographics. As the dependent variable a dummy variable was used, where 0 indicated unit non-response to wave 2, and 1 indicated unit response to wave 2. This model was estimated by means of the WLSMV estimator in MPlus; this is a mean- and variance-adjusted weighted least square estimator (Muthén and Muthén 2004, 2006; Finney and DiStefano 2006). The WRMR was 1.020, indicating just acceptable fit (Yu 2002). Since the CFA specification had been validated before and since little extra variables

and restrictions were added, this model was accepted as a valid approximation of reality and the estimates were evaluated as reported in Table 6-2. From this table it appears that education level was positively related to the probability of participating in wave 2. Apart from that, no significant effects were observed. It can be concluded that response styles at time 1 were not predictive of response to wave 2. Hence, levels of response styles between respondents to wave 1 only and respondent to both waves can be plausibly accepted not to vary apart from the variance induced by their different levels of education.

TABLE 6-2: REGRESSION COEFFICIENTS FOR SELECTIVITY CHECK

IV[a]	B[b]	s.e.	t	Stdd B
ARS	0.120	0.542	0.221	0.029
DRS	-0.384	0.585	-0.657	-0.072
ERS	-0.450	0.927	-0.486	-0.071
MRS	0.001	0.953	0.001	0.000
AGE	0.007	0.024	0.272	0.010
EDU	0.067	0.019	3.596	0.120
FEMALE	-0.096	0.071	-1.341	-0.047

[a]IV= Independent variable;
Dependent variable is Non-response (=0) /Response (=1) to wave 2
[b]The regression weights are probit coefficients

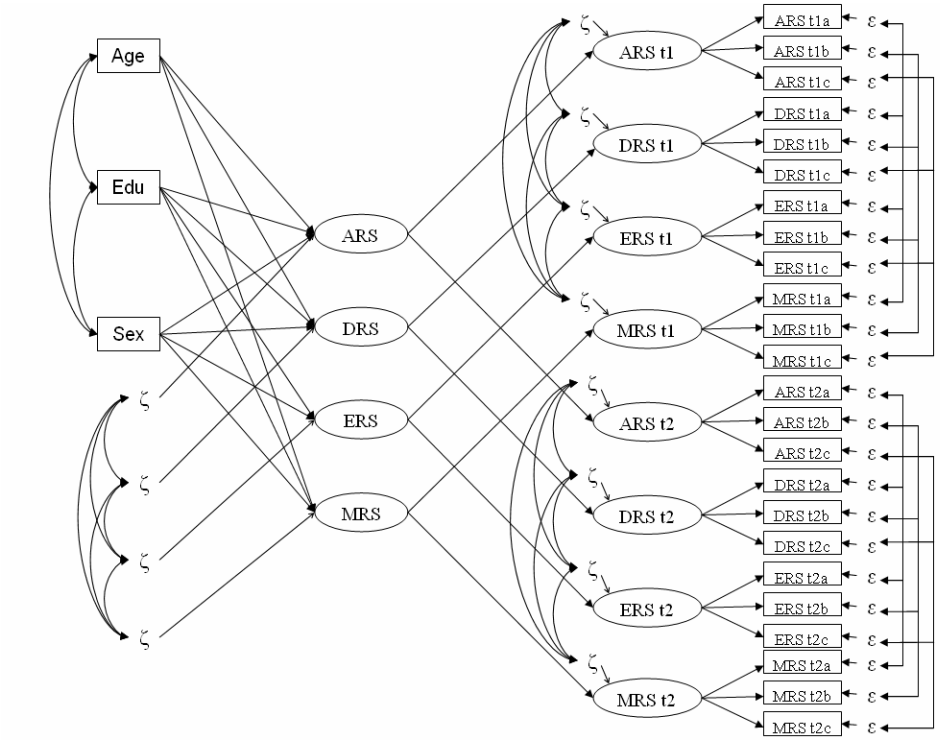
MIMIC MODEL OF TIME INVARIANT SECOND ORDER RESPONSE STYLE FACTORS

The focal model of this study was a mimic model (multiple indicators – multiple causes), in which response styles were specified as time invariant second order factors. The response style factors measured in wave 1 and wave 2 were their indicators. The demographics were the antecedents. Figure 6-1 depicts the mimic model. Note that the correlated uniquenesses for the observed indicators are omitted

from the figure (the details of these correlations are discussed in the time specific CFA's). At the first order level (time specific level), the disturbances were correlated because the response styles were expected to covary due to time specific factors. For example, a respondent might have been in a given mood or under time pressure when filling out questionnaire 1, but this effect might not have been present at time 2. At the second order level (the time invariant level), the response styles were correlated because the demographics were not expected to explain all the shared variance between the four response styles. More specifically, response styles might covary due to non-modeled common causes like stable individual traits. On the first order level, the factor loadings of one indicator per factor were set to one. On the second order level, both factor loadings per response style were set to one.

Figure 6-1

Mimic model of temporal stability of response styles



The chi square test of model fit was significant ($\chi^2(254, N=604)= 537.450, p<.001$), indicating statistically significant misfit of the model to the data. On the other hand, the alternative fit indices showed acceptable values (CFI = .980; TLI = .971; RMSEA = 0.043, 90 Percent C.I. = 0.038 to 0.048; Probability(RMSEA <= .05)=0.989). Also, the indices of local misfit indicated that potential misspecifications were statistically significant but substantially negligible.

The residual variances (or disturbances) of the response style factors on both the time specific first order level and the time invariant second order level are reported in Table 6-3. All residual variances were significantly different from zero at the .05 level. For the time invariant level, this indicates that the time specific response style factors shared an amount of stable variance other than that explained by the variance they shared with the demographic background variables. However, the time specific non-zero variances mean that the stable factor did not explain all the response style variance observed at one point in time. To obtain a clearer insight in the relative contribution of the respective variance components, the AVE's (average variance extracted) of the response style factors are presented in Table 6-3, both for the time specific and time invariant factors. On the time specific level, it is readily apparent that the different independent random samples of items all form the basis for reliable response style indicators, as shown by the AVE values (Table 6-3). This indicates that response style levels were stable at least at the time specific levels. At this level, the average response style indicator shared 68% of its variance with its time specific factor (see AVE in the Table 6-3). At the time invariant level, also remarkably high factor loadings were found: just over half of the variance in the average time specific response style factor was explained by its time invariant counterpart (see AVE in the

time invariant columns of Table 6-3). In the current data, DRS was the least stable over time, followed by ARS (both less than half of their variance was explained by the time invariant factor), while ERS and MRS had quite impressive levels of explained variance (58 and 57%; see Table 6-3).

Table 6-4 presents the structural regression weights and explained variances of the four response styles regressed on demographic variables. Just over half of the effects are significant at the .05 level. ARS is positively related to age. DRS is positively related to education level. Both MRS and ERS are negatively related to education level and positively related to age. Moreover, ERS is higher among females. The proportion of variance in the response style factors that is explained by the demographics varies from a low 2.3% for DRS to a maximum of 9.5% for ERS. MRS and ARS are somewhere in between, with respectively 6.7 and 6.1%.

TABLE 6-3: VARIANCE AND AVERAGE VARIANCE EXTRACTED (AVE) OF THE RESPONSE STYLE FACTORS

	Time invariant					Wave 1					Wave 2				
	AVE	s^2	s.e.	t	p	AVE	s^2	s.e.	t	p	AVE	s^2	s.e.	t	p
ARS	0.49	0.029	0.003	9.52	<0.001	0.65	0.055	0.005	10.18	<0.001	0.65	0.009	0.001	10.00	<0.001
DRS	0.44	0.018	0.002	8.55	<0.001	0.54	0.022	0.003	7.02	<0.001	0.66	0.012	0.001	8.64	<0.001
ERS	0.58	0.013	0.001	10.89	<0.001	0.75	0.009	0.001	7.07	<0.001	0.83	0.026	0.003	8.58	<0.001
MRS	0.57	0.006	0.001	10.15	<0.001	0.56	0.003	0.001	4.25	<0.001	0.80	0.020	0.003	6.02	<0.001

TABLE 6-4: STRUCTURAL REGRESSION WEIGHTS OF MIMIC MODEL

DV	IV	Estimate	S.E.	C.R.	P	Stdd	R^2
ARS	EDU	-0.010	0.005	-2.00	0.045	-0.11	0.061
	AGE	0.031	0.007	4.67	<0.001	0.25	
	SEX	0.038	0.020	1.95	0.052	0.11	
DRS	EDU	0.008	0.004	2.09	0.037	0.12	0.023
	AGE	0.006	0.005	1.05	0.293	0.06	
	SEX	0.016	0.016	1.02	0.308	0.06	
ERS	EDU	-0.009	0.003	-2.95	0.003	-0.14	0.095
	AGE	0.026	0.004	6.21	<0.001	0.31	
	SEX	0.032	0.012	2.55	0.011	0.13	
MRS	EDU	-0.008	0.002	-3.93	<0.001	-0.20	0.067
	AGE	0.010	0.003	3.59	<0.001	0.19	
	SEX	-0.003	0.009	-0.38	0.707	-0.02	

To get a better understanding of the effects reported in Table 6-4, it may be useful to estimate the mean scores of the four response styles. To do so, a model is estimated in which the factor mean of each time invariant response style is set to zero, as are the intercepts of all observed response style indicators. At the intermediate level, the time specific factor intercepts are freely estimated, thus guaranteeing an estimate of the average score that is based on the optimal weighting of the observed mean scores. The resulting estimates are presented in Table 6-5. The intercept constraints lead to a highly statistically significant increase in misfit ($\Delta\chi^2(16, N=604)=121.15$, $p<.001$), but a relative small deterioration of the alternative fit indices (TLI=.009; CFI=.007), overlapping RMSEA intervals (respectively P(.038<RMSEA<.048)=.95 and P(.044<RMSEA<.054)=.95), and acceptable overall model fit ($\chi^2(270, N=604)=658.601$; TLI=.962; CFI=.973; RMSEA = .049; 95% C.I.: .044-.054;

p(RMSEA<.05)=.648). Based on this, the means were accepted as reasonable estimates. Nevertheless, the mean response style levels were significantly different across the sets of items, a finding in line with Greenleaf's (1992a) remarks on how to create sets of items that are parallel with regard to their response style levels.

TABLE 6-5: MEAN ESTIMATES OF RESPONSE STYLES

	Wave 1		Wave 2		Difference
	Mean	s.e.	Mean	s.e.	t
ARS	0.87	0.02	0.86	0.02	0.94
DRS	0.59	0.01	0.64	0.01	3.94
ERS	0.22	0.01	0.22	0.01	0.13
MRS	0.19	0.01	0.21	0.01	2.31

The residual correlations between the response styles on the time invariant second order level (i.e. the correlations capturing the shared variance not explained by shared antecedents, in this case demographics), were .25 for ARS and DRS, .71 for ARS and ERS, .57 for DRS and ERS, -.46 for MRS and ARS, -.43 for MRS and DRS, and -.03 for MRS and ERS. Apart from the MRS-ERS correlation, all of these are significant at the .05-level.

DISCUSSION

In the current study, response styles were measured over two waves of data collection using independent random sets of items. The time between the two waves was one year. Consequently, some respondents did not respond to the second wave of data collection. However, response style levels of respondents in wave 1 were not predictive of their response/non-response to wave 2 after controlling for demographics. It was found that demographics were predictive of participation to

wave 2. In particular, respondents with higher education levels had higher probabilities of participating in wave 2. For this reason, and to estimate the effects of demographics on response styles, a mimic model was specified of four response styles, using education, age and sex as the antecedents of acquiescence response style (ARS), disacquiescence response style (DRS), extreme response style (ERS) and midpoint responding (MRS). ARS, DRS, ERS and MRS were specified as latent factors acting on two levels: the time specific level of response styles is a result of a time invariant response style factor; complemented by a time specific unique disturbance (non-modeled situational variables). The time invariant response style factors were regressed on demographics, which were modeled as time invariant covariates (making abstraction of the one year increase in age and other potential changes).

On the time invariant level, ERS was strongly positively related with both ARS and DRS. ARS and DRS were positively related too, but to a lesser extent. This indicates that ARS and DRS, rather than opposites of the same pole, may to some extent be indicative of respondents' willingness to choose sides on the issues presented to them and to differentiate their responses accordingly. Not all ARS and DRS variance should therefore necessarily be equated with directional bias, a point also raised by Bachman and O'Malley (1984) and Greenleaf (1992b).

The data further showed that, after controlling for demographics, MRS was negatively related to ARS and DRS and non-significantly related to ERS. This also concurs with the above observation that ARS, DRS and ERS may be indicative of differentiation. MRS and ERS did not constitute opposites of the same dimension, but were nearly orthogonal dimensions. Thus, it is essential not to operationalize these response styles by one measure, as is sometimes done.

In sum, MRS seems to indicate a tendency not to differentiate; ARS and DRS are to some extent determined by a tendency to differentiate, and to some extent by a tendency to use extreme directed responses in doing so. The latter is captured by ERS, which is nearly independent of non-differentiation (MRS). While the four response styles share some variance with one another, they cannot be completely reduced to a more limited set of factors. For example, if a higher order factor indicating differentiation were specified, ARS and DRS would load equally and positively, while MRS would load negatively, but the factor would only explain a very limited proportion of the response style factors' variances. Further, if a higher order factor were proposed linking ERS, ARS and DRS, this factor would not be able to account for the different correlations between ARS and ERS on the one hand, and DRS and ERS on the other hand. Consequently, to obtain a response style profile of a respondent or group, all four response styles are necessary and form complementary non-redundant dimensions.

Each of the response styles was significantly affected by some specific demographics. The explained variance was rather modest, with R squares below 10% in all cases. In heterogeneous samples, the response style differences across demographic groups may seriously bias results though. For example, consider an average respondent L (for 'low education') with 6 years of formal education (only primary school) as opposed to an average respondent H (for 'high education') with 18 years of formal education (postgraduate). The expected levels of response styles for L as compared to H would be .14 higher for the net effect of ARS-DRS (i.e. average score difference on a seven point scale), but, what is more alarming, 10.8% higher on ERS and 9.6% higher on MRS, corresponding to the respective proportions of extreme and midpoint responses. Similarly, comparing an average respondent Y (for young) aged 20 years to an

average respondent S (for senior) aged 70, would result in expected levels of response styles for S as compared to Y that would be .12 higher due to the net effect of ARS and DRS, as well as 13% higher on ERS and 5% higher on MRS. To obtain the predicted scores of a lowly educated seventy-year old versus those of a highly educated twenty-year old, it suffices to add the above effects. It is obvious that, while the directional bias due to ARS and DRS may be moderate, the distributions will look quite dissimilar, with the seventy-year old lowly educated respondents showing a nearly trichotomized response distribution. Such response pattern has been noted by Osgood (1941), who also linked it to low education.

While the above effects are worrisome when comparing the extremes of the demographic spectrum, demographics account for only a minor proportion of the variance in response styles. However, the current results provide conclusive evidence that in addition to the stable component of response styles explained by demographic differences, there also is a much bigger proportion of stable response style variance not being explained by these background variables. Based on the current study, this stable portion of response styles cannot be related to specific variables. While the literature has suggested some possibilities, no convincing evidence is available. A major obstacle in proving the link between response styles and some stable individual characteristic, like personality traits, is that such traits are most commonly measured by means of self-reports, which can reasonably be assumed to be contaminated by response styles, resulting in circular causality. Also, the established stability of these same trait measures might be due at least in part to response style stability. Nevertheless, current results clearly indicate the necessity to solve this issue.

While the current study did not link response styles to explanatory variables other than demographics, the observed effects combined with the correlations between response styles, provide some clear clues on the meaning of response styles.
MRS and ERS both positively relate to age[11], and negatively to education level, suggesting an association with cognitive limitations. A similar profile is obtained for ARS, but less so. While ARS probably has a cognitive limitation component, it also has a component related to differentiation. Hence the positive correlation with DRS. DRS in turn, is positively related to education level, confirming its status as the consequence of critical thought and differentiation rather than a directional bias.
It is notable that researchers commonly have been most preoccupied with ARS or the net effect of ARS-DRS (Ray 1979; Watson 1992; McClendon 1991a, b; Billiet and McClendon 2000). The reason for this attention for ARS is probably that the bias that may be caused by directional response styles is most obvious and easily understood. At the same time, researchers who have criticized response style research have most commonly focused on ARS, arguing that it is (1) non-existent or unstable (Rorer 1965, in "The great response style myth"), or (2) that its biasing effect is rather limited (Schimmack, Böckenholt and Reisenzein 2002). However, it is ERS that shows the highest stability and the strongest relationship with demographics in the current data. This concurs with previous findings by Peabody (1962), who observed that ERS most probably is a stable response style, while observed directional differences (in agreement levels) are more closely related to content rather than to style.

[11] Since our sample is limited to adults, our data do not contain the age brackett where ERS may decline over age, i.e. from childhood to adolescence (Marsh 1996; Hamilton 1968). We confirmed the linearity of the observed effects by studying scatter plots of the estimated factor scores by age.

LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

While a major contribution of the current study is the establishment of longitudinal stability of response styles over a one year time period, it would be interesting to study longitudinal data that allow one to track stability and change of response styles over several years. Such design would allow the study of growth curves of response styles over the life cycle.

A key opportunity for future research lies in the challenge of measuring stable individual traits, e.g. personality traits, in a way that guarantees the absence of response style bias. This would allow researchers to investigate how such traits are linked to response styles. Given the current observation that at least 90% of the stable variance in response styles is unexplained, this is one of the priorities for response style research in the near future.

CHAPTER 7: ASSESSING RESPONSE STYLES ACROSS MODES OF DATA COLLECTION (EMPIRICAL STUDY 4) [12]

CHAPTER OUTLINE

The current study compares levels of response styles across three modes of data-collection: paper and pencil questionnaires, telephone interviews and online questionnaires. Using Means And Covariance Structures (MACS), data collected by different modes are found to show differences in response styles. Specifically, telephone data have lower levels of midpoint responding. The potential bias the observed response style differences may cause are illustrated and discussed.

[12] A previous version of this paper is available as Vlerick Leuven Gent Management School working paper 2004/20 (Weijters, Geuens and Schillewaert 2004) and Ghent University FEB working paper 05/349. This study was also presented at the Marketing Science conference 2004 in Rotterdam.

INTRODUCTION

Imagine a researcher wants to compare the levels of satisfaction among online and offline customers of a retailer that uses both the online and the offline channel. Based on practical considerations, use of a multi-mode survey combining online and offline data collection would most probably be an option to address the question at hand. The whole set-up would be useless, however, if the online and offline data were not comparable in terms of how the item responses reflected the underlying construct, satisfaction.

It has become common practice to frame this issue of comparability in terms of measurement invariance. While measurement invariance testing has received quite a lot of attention, also to a growing extent in cross-mode contexts (see, e.g., Deutskens, de Ruyter and Wetzels 2006; Ferrando and Lorenzo-Seva 2005), the possibility that violations of measurement invariance may be due to response styles has not. Cheung and Rensvold (2000) have shown that measurement invariance may be violated as a consequence of group differences in levels of response styles and have demonstrated how measurement invariance tests can be used to assess differences in response styles between groups. While these groups have often been samples from different cultures, the problem of measurement invariance translates directly to the cross-mode situation.

Using the assessment of measurement invariance to detect response style differences has some important limitations, however. First, assessment of measurement invariance is content specific, in that it relates to the equivalence of the particular construct and its indicators under investigation. Hence, acceptance of invariance of a specific measure does not carry any information that generalizes beyond this measure. Second, as Little (2000) has argued in reaction to Cheung and Rensvold's (2000)

article, establishing measurement invariance does not rule out response bias, especially when such bias is uniform across different indicators. For example, if a factor is operationalized by means of three items, each of which is contaminated by ARS to a similar extent, this bias will most probably not result in rejection of the hypothesis of scalar invariance, but may very well show up as an apparent latent mean difference. Third, testing for measurement invariance is inherently diagnostic in nature, and does not offer corrective measures when invariance is found to be violated. On the other hand, once measures of response styles have been created, they can be used to correct observed scores for the bias due to response styles. Finally, it is not clear why a cross-group comparison should be limited to only two response styles, namely Net Acquiescence Response Style (NARS) and Extreme Response Style (ERS), while several other response styles have been identified, most notably Midpoint Response Style (MRS). Also, Acquiescence Response Style (ARS) and Disacquiescence Response Style (DRS) are two related but separate response styles and can hence not be reduced to NARS (Bachman and O'Malley 1984). All of these response styles (i.e. ARS, DRS, ERS and MRS) have been found to potentially cause bias in measurement of constructs (Baumgartner and Steenkamp 2001).

Because of the above reasons, a direct assessment of response style differences between modes of data collection would be preferable to the indirect assessment via invariance tests of measurement parameters. Therefore, the objective of the current paper was to compare levels of response styles across modes of data collection. The potential differences with classical measurement invariance tests are illustrated and discussed.

Investigating response bias across different modes of data collection is highly relevant. Recently, multiple modes of data collection or mixed-modes have become

increasingly popular in survey practice (de Leeuw 2005). It is crucial for contemporary survey research to investigate whether different modes of data collection bring along different levels of response styles. This is an important issue for both practical and academic research with repercussions on the optimal choice of a data collection procedure.

Especially with regard to the growing importance of the Internet and web surveys (Gunter et al. 2002; Johnson 2001; Griffis, Goldsby and Cooper 2003; Deutskens 2006), such comparison would enrich the understanding of the comparability of various research methods. Although researchers have identified a wide range of possible (dis)-advantages of web surveys, the focus of previous research is mainly on response rate, response speed, costs, representativeness of samples, anonymity and confidentiality (Deutskens et al. 2004; Gunter et al. 2002; Ployhart et al. 2003; Simsek and Veiga 2001; Thompson et al. 2003; Truell 2003).

The current paper compares offline self-administered questionnaire data, telephone interview data and online self-administered data in terms of systematically measured response styles, using a highly diverse set of commonly used questionnaire items among three subsamples of respondents who responded to the same questionnaire via a different mode of data collection. Additionally, as an illustration, it is tested how a cross-mode comparison of a substantive construct measure may be biased by response styles and whether or not this is detectable by means of the classical measurement invariance tests.

THEORETICAL FRAMEWORK

RESPONSE STYLES

In survey studies, researchers assume that the responses to questionnaire items reflect a respondent's true position towards the content of the question. This is not always the case though. The presence of random error has been generally accepted and is often accounted for by using multi-item scales (Churchill 1979), possibly combined with Structural Equation Modeling (Fornell and Larcker 1981). The effect of systematic error, on the other hand, poses more serious problems to the validity of survey research and has not been as widely recognized or investigated as would be warranted by its potential biasing effects (Baumgartner and Steenkamp 2001; Podsakoff, Mackenzie, Lee and Podsakoff 2003). Often, respondents seem to be prone to response styles, defined as "*[behavior patterns] where the individual tends to select disproportionately a particular response category regardless of item content*" (O'Neill 1967). Based on the impact they have on observed scores, one could distinguish between two major types of response styles: unidirectional and bidirectional. Unidirectional response styles refer to a respondent's preferred use of positive, neutral or negative response options. The net result of these styles is a shift in the within-subject mean (Greenleaf 1992a; Net Acquiescence Response Style in Baumgartner and Steenkamp 2001). There are three such unidirectional response styles: Acquiescence Response Style (ARS), i.e. the tendency to disproportionately use positive response categories; Disacquiescence Response Style (DRS), i.e. the tendency to disproportionately use negative response categories; and Midpoint Responding (MRS), i.e. the tendency to disproportionately use the midpoint of a scale. Bidirectional response styles, on the other hand, refer to a respondent's tendency to use response categories that are present on both sides of the response

option spectrum. This category consists of only one response style: Extreme Response Style (ERS), the tendency to use the most extreme response options on both the left and the right hand side of the scale (Baumgartner and Steenkamp 2001; Greenleaf 1992a, b). The net result of this bidirectional style is a change in the within-subject standard deviation (Greenleaf 1992a)[13].

As demonstrated by Cheung and Rensvold (2000), response styles affect observed scores and their relation to the latent variables they reflect. More specifically, in a measurement model where observed variable x is a linear function of latent variable ξ and unique factor δ, with intercept τ, such that $x = \tau + \lambda\xi + \delta$, higher (lower) ARS inflates (deflates) measurement intercept τ, and higher (lower) ERS inflates (deflates) factor loading λ. Consequently, if groups have different levels of response styles, this will lead to between-group differences in measurement intercepts and loadings. However, to be able to compare groups in terms of latent means, metric and scalar invariance have to be satisfied (Little 1997; Vandenberg and Lance 2000): metric invariance refers to the condition in which the measurement slopes λ are equal across groups, while scalar invariance refers to the condition where, in addition to metric invariance being established, the measurement intercepts τ are equal across groups (Steenkamp and Baumgartner 1998). As Cheung and Rensvold (2000) point out, inter-group differences in response styles may threaten metric and scalar invariance and render inter-group comparisons impossible. However, while it is by now generally

[13] Note that Baumgartner and Steenkamp (2001) show that ERS also has an effect on the expected mean score of a scale. However, this effect is conditional on the mean deviation from the midpoint, which is closely related to the idea of an interaction effect between the latent score and ERS proposed by Cheung and Rensvold (2000) and implicitly applied by Greenleaf (1992a).

acknowledged that measurement invariance is a necessary condition for meaningful inter-group comparisons (Meredith 1993; Ployhart and Oswald 2004; Steenkamp and Baumgartner 1998; Vandenberg and Lance 2000), it is not a sufficient condition (Little 2000). Both uniform bias due to ARS and/or ERS, as well as bias due to response styles other than ARS and ERS may go unnoticed in invariance testing. Further, measurement invariance needs to be established for each measure in each measurement situation and hence has little generalizability. An additional limitation of studying response style differences by means of invariance testing is that the latter procedure is limited to a measure specific diagnosis of the problem, and does not allow for correction of bias if such bias is observed. To counter these limitations, it is necessary to make a more direct assessment of response styles, by creating measures of the response styles themselves, rather than by assessing their impact indirectly via their biasing effect on measurement parameters.

A MACS OPERATIONALIZATION OF RESPONSE STYLES

Based on a thorough review of the literature, Podsakoff et al. (2003) conclude that multi-indicator multi-method factor measurement of sources of bias has important advantages over models using single indicators and/or single method factors. While previous research has used multiple indicators for response styles (Baumgartner and Steenkamp 2001) and has modeled one specific response style, ARS, as a latent variable (Billiet and McClendon 2000; Welkenhuysen-Gybels, Billiet and Cambré 2003), the current paper proposes and applies the use of multiple specifically designed indicators for simultaneous measurement of several response styles. The discussion that follows, clarifies why such procedure is more than a straightforward extension of existing approaches. For a valid operationalization of response styles, several

requirements need to be simultaneously addressed. First, the operationalization needs to represent a complete profile of unidirectional and bidirectional response styles. A reduced set, focusing on item intercepts and loadings alone (i.e. the parameters that are commonly tested for invariance) may miss important sources of bias and does not capture the full scope of behavioral phenomena of interest (i.e. the underlying response styles). Second, as is the case for all latent constructs, response styles need to be invariant across groups in order to be comparable in a meaningful way. While the necessity of measurement invariance has been generally acknowledged, it has not been applied to measurement of response styles, which is remarkable in light of the methodological focus of the research domain. To be able to assess invariance, there is a need for multi-indicator measures of response styles[14]. The use of multiple rather than single indicator measures for each response style is necessary not only for invariance testing, but also to account for measurement error and to enable correct assessment of convergent and discriminant validity of the response style measures (Podsakoff et al. 2003). A specific issue with response style measures is that they usually are based on different mathematical transformations of the same data. For example, ARS and DRS measures based on the same item set will have a structural tendency to correlate negatively. It is not a conceptual necessity, however, that the response styles themselves are negatively related (Bachman and O'Malley 1984). By

[14] Note that this requirement is independent of the relation between measurement invariance and response styles as discussed by Cheung and Rensvold (2002; see above). These authors state that measurement non-invariance may be indicative of response styles. Here it is stated that measures of response styles need to meet the condition of measurement invariance in order to be valid and useful for group comparions of response style levels. These propositions, while superficially similar, relate the same concepts but in a different way.

correlating indicators based on the same item sets while at the same time correlating the response style factors, it is possible to obtain a more truthful estimate of the correlation between the response styles rather than between response style indicators. This may become clearer later on, when the measurement model is discussed in more detail. Evidently, the use of multiple indicators also facilitates the assessment of internal consistency. Further, it makes it possible to explicitly model the unique variances of indicators, which is important in this context, since it is crucial to abstract only the variance that is not specific to a certain subset of items.

A further requirement is that response style measures should be based on a representative sample of heterogeneous items. Researchers often use convenience samples of items to measure response styles, usually because secondary data are analyzed (e.g., Bachman and O'Malley 1984). The use of a random sample of items is preferable because maximum heterogeneity of the content of the items ensures that the observed response tendencies are not contentwise related (Greenleaf 1992a), and because only the use of a representative sample of items allows one to generalize findings across all items in the same population. Previous cross-mode comparisons are limited in this regard (as discussed below).

RESPONSE STYLES ACROSS MODES OF DATA COLLECTION

Notwithstanding the availability of several modes of data collection and the growing success of the Internet in this respect (Johnson 2001), little research is available that addresses the impact of mode of data collection on response styles.

Jordan, Marcus and Reeder (1980) compared telephone and household interviews, and found more acquiescence and extremeness in the telephone interviews. Kiesler and Sproul (1986) compared electronic and paper mail self-administered surveys in terms of the contents of responses to a specifically health related questionnaire. They found

that in the electronic surveys, people tended to show less inhibition in their responses, but mainly concluded that their results "*show considerable similarity of response between the paper and electronic survey but not so much that the two may be considered interchangeable without further research"*. The measures used by Jordan et al. (1980) and Kiesler and Sproul (1986), however, were constructed ad hoc and related to the specific content of the questionnaire. Indeed, both Jordan, Marcus and Reeder (1980) and Kiesler and Sproul (1986) measured response styles in the content domain of health related issues, which may be a domain that is particularly sensitive to biases that are response set-based (i.e. content related; Rorer 1965) rather than response style based (i.e. non-content related; Rorer 1965). This implies that the major advantage of a direct assessment of response styles, i.e. the generalizability beyond a specific content domain, is not realized. The limitation of topic specificity also applies to other mixed-mode studies on comparability of different modes with regards to different aspects (for an overview of such comparisons of online and mail surveys, see Deutskens, de Ruyter and Wetzels 2006).

Further, in the studies by Jordan et al. (1980) and Kiesler and Sproul (1986), only limited subsets of response styles were studied, using operationalizations that were suboptimal (i.e. not meeting standards set by, e.g. Rorer 1965; Bentler, Jackson and Messick 1971; Greenleaf 1992b). Consequently, the question remains open whether and to what extent mode of data collection systematically affects (non-content related) response styles and the need exists for a comparison of modes of data collection using a thorough operationalization of all relevant response styles based on a diverse and broad sample of commonly used scale items.

The topic of mode comparability is becoming especially important since substantive questions need to be answered concerning the generalizability of conceptual models

from an offline to an online context (see for example Szymanski and Hise 2000; Venkatesh, Smith, and Rangaswamy 2003). Often, respondents in the offline and online settings are easier to reach respectively by means of mailed paper surveys and e-mails linking to online questionnaires respectively. To save costs, researchers may also want to use online questionnaires for as many respondents as possible and complement the mode with another mode to cover the whole population of interest, including those who are not online.

HYPOTHESES

The current study aimed to compare self-administered paper and pencil questionnaires (P&P), telephone interviews (Telephone), and self-administered online questionnaires (Online). The P&P mode is considered the reference group to which the other two modes are compared. The P&P mode and the Telephone mode differ from one another in several important aspects. While perception of the items / response to the items is visual / manual in the P&P mode, it is auditory / vocal in the Telephone mode. However, since the response options to a series of Likert items are identical for all items, it is not very plausible that response order effects will occur. Depending on the mode of data collection, primacy and recency effects have been observed in this regard (Krosnick and Alwin 1987), but only for response options that were idiosyncratic to one question (i.e. a specific list of option is read for each specific question), which is not the case here. A probably more influential difference between P&P and Telephone is the presence of an interviewer in the latter condition. The interviewer's presence may motivate respondents to provide an answer other than the midpoint, since a midpoint response might be experienced as non-satisfactory (Ayidiya and McClendon 1990). Moreover, while the P&P questionnaires are self-administered and consequently self-paced, in the Telephone mode an interviewer is

largely in control of the process, possibly speeding up the process to some extent, if only because silences on the phone may be experienced as awkward. Time constraints have been found to increase the levels of ARS (McGee 1967). Hence, the telephone mode would be expected to lead to higher ARS. In line with Jordan et al. (1980) it is also hypothesized that ERS is higher in the Telephone mode. Since the Telephone mode will probably lead respondents to be biased towards acquiescence (Jordan et al. 1980), a negative effect of Telephone mode on DRS is posited.

The above suggests the following hypotheses:

H1a: The Telephone mode of data collection has a higher level of ARS than the P&P mode.

H1b: The Telephone mode of data collection has a lower level of DRS than the P&P mode.

H1c: The Telephone mode of data collection has a higher level of ERS than the P&P mode.

H1d: The Telephone mode of data collection has a lower level of MRS than the P&P mode.

Unlike the Telephone mode, the Online mode of data collection is very similar to P&P in most respects, including visual perception of the questions, manual response to the questions, and self-administration. The latter aspect implies that the respondent decides on the speed with which the items are read and responded to. Given these similarities and the tentative conclusions by Kiesler and Sproul (1986), the null hypotheses are posited for the response style comparison between P&P and Online data collection:

H2a: The P&P and Online mode of data collection have equal levels of ARS.

H2b: The P&P and Online mode of data collection have equal levels of DRS.

H2c: The P&P and Online mode of data collection have equal levels of ERS.

H2d: The P&P and Online mode of data collection have equal levels of MRS.

These null hypotheses are especially relevant because they represent the ideal case and the implicit working hypothesis of cross-mode research that does not explicitly test for response style differences (e.g., Venkatesh, Smith, and Rangaswamy 2003).

EMPIRICAL STUDY

Random assignment of respondents to modes of data collection is not a viable strategy to address the current question. Much of the differences between modes may be due to situational, uncontrollable variables (Ferrando and Lorenzo-Seva 2005) and an overly controlled setting would impede external validity and would risk making the study irrelevant. As a consequence, the most valid design seems to be a quasi-experiment using balanced samples. Balancing should be based on the variables that have been identified as key antecedents of response styles. These are age (Knauper 1999; Greenleaf 1992a; Hamilton 1968; Mirowsky and Ross 1991), education level (Shulman 1973; Hamilton 1968; Greenleaf 1992a; McClendon 1991a; Narayan and Krosnick 1999) and gender (Hamilton 1968; Greenleaf 1992a).

The empirical study is reported in two parts. In Part 1, a cross-mode comparison is made of the levels of response styles. To illustrate the relevance of the observed differences, in Part 2 a cross-mode comparison is made of the scores on a latent construct, with or without correction for response styles.

PART 1: DIAGNOSIS OF CROSS-MODE DIFFERENCES IN RESPONSE STYLES

A multi-group cross-mode MACS is specified and tested that allows for the assessment of response style measurement invariance across modes of data collection as well as for a comparison of levels of response style bias across modes of data

collection. A randomly selected construct is included in the questionnaire with the aim of illustrating a proposed correction procedure for response styles. The latter topic is discussed later. First the cross-mode mean differences in response styles are investigated to test the above hypotheses.

METHODOLOGY

Respondent sampling

Data were collected among three samples of respondents, using identical questionnaires across three modes of data collection: (1) Self-administered Paper and Pencil questionnaire (P&P): N=655, recruited by means of a random walk procedure[15] (response rate 58.0%); (2) Telephone interview (Tele) among a sample taken from the general population: N = 496 (response rate 32.0%); (3) Self-administered online survey among the online panel of an online market research company, recruited by means of a personalized e-mail (Online): N=1445 (response rate 48.2%)[16].

[15] For each day of data collection, each data collector received one randomly generated address, covering city, suburb and countryside. From this start address, they followed a predefined procedure explaining how to select the next address. Questionnaires were collected two days later.

[16] Note that the response rate in the telephone mode was lower than in the other modes due to higher refusal rates in this group, a widely acknowledged phenomenon, also in cross-mode designs similar to the one reported here (e.g. Jordan, Marcus and Reeder 1980, p. 212; McClendon 1991b). Hence, this should be considered a weakness of the telephone mode rather than a weakness of the current study. Overall, the obtained response rates compared favorably to average response rates to consumer surveys reported in top marketing journals (as charted in a meta-analysis by Anseel, Lievens and Vermeulen 2006). Hence, it seems safe to conclude that the current sample represents the population of interest, i.e. respondents to consumer surveys. Also note that respondents to later reminders or respondents with lower average response rates have been found to show similar data quality as do other respondents (Andrews 1984).

With the aim of obtaining comparable samples, three equally large samples were resampled from the above groups, balancing for age, education level and sex. This procedure ensures that observed differences in response styles cannot be attributed to demographic differences. Additionally, it leads to comparable sample sizes, thus guaranteeing similar levels of power for mean difference tests across all combinations of the three modes. Since the telephone sample was the smallest group, it was used as the target level group in computing sampling probability weights. As intended, the resulting samples showed no significant differences on the three demographic variables in chi square and ANOVA tests (respective p-values for age, education and sex were .993, .856 and .434). The respective balanced samples for P&P (N=501), Tele (N=496), and Online(N=535) had average ages of 46.3 (s=13.9), 46.3 (s=13.0), and 46.2 (s=13.4); average years of formal education of 12.5 (s=2.7), 12.6 (s=2.6), and 12.6 (s=2.6); percentages of females of 64.9%, 65.7%, and 62.1%. It was tested whether the results were robust against fluctuations in sampling. This proved to be the case.

Questionnaire and item sampling

From the compilation of multi-item scales by Bruner, James and Hensel (2001), 52 unrelated items were randomly selected from different scales. All items were adapted to a seven point Likert scale. To be able to assess the impact of response styles on a substantive measure (see below), a multi-item measure of trust in frontline employees (TRUST) in a clothing retail context was included. The construct was measured by means of four items taken from Sirdeshmukh, Singh, and Sabol (2002). For this measurement, respondents were asked to think back of their latest such encounter. The TRUST items were grouped in one block.

Response style indicator calculation

The 52 randomly selected items had an average inter-item correlation of .07. The item series was randomly split into three sets (a number of indicators that balances stability with parsimony; Little et al. 2002), each of which was used to calculate an indicator for each response style using equations 1 through 4 below. All sets consisted of 17 or 18 items. This allowed computing three indicators each for ARS, DRS, ERS, and MRS by applying the following formulas (Bachman and O'Malley 1984; Baumgartner and Steenkamp 2001; Hui and Triandis 1985)[17]. For each set of k items:

(1) $$ARS = [f(5)*1 + f(6)*2 + f(7)*3]/k$$

(2) $$DRS = [f(1)*3 + f(2)*2 + f(3)*1]/k$$

(3) $$ERS = [f(1) + f(7)] / k$$

(4) $$MRS = f(4) / k$$

In these formulas, f (o) refers to the frequency of response option o. Consequently, ARS and DRS can be interpreted as the bias away from the midpoint of a response scale due to acquiescence and disacquiescence. The net effect of both response styles is easily obtained as ARS – DRS. MRS can be read as an estimate of the proportion of midpoint responses, ERS as an estimate of the proportion of extreme responses.

MACS Model and data analysis

Calibration of factor structure

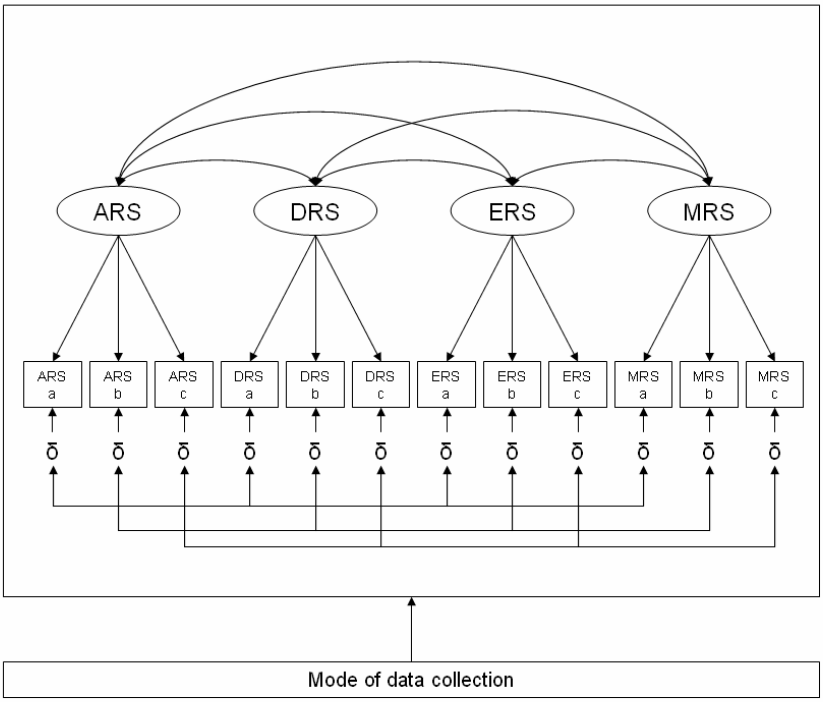
First the response style model was calibrated on an online hold-out sample (N=500), which had a similar demographic profile as the three validation samples (Telephone,

[17] In line with the definition of response styles, response style measures were based on sets of unrelated items and operationalizations were used that are not content related; more specifically, for ARS and DRS, the methods labeled 'ARS1' and 'DRS1' in Baumgartner and Steenkamp (2001) were used.

P&P and Online). The response style model consisted of a MACS model in which ARS, DRS, ERS and MRS were freely covarying latent constructs. Each factor had three indicators. Across response styles, the indicators that were based on the same sets of items had correlated error terms to take into account the shared variance due to basing measures of response styles on the same items (see Figure 7-1)[18].

Figure 7-1

MACS for cross-mode mean comparison



[18] Such model corresponds to a covariance matrix of the indicators in which not only the main diagonal (containing the variances) is systematically higher than the other values, but also the diagonals of each of the submatrices corresponding to covariances of different style indicators based on the same item sets.

The model showed good fit to the data (χ^2 (30) =25.29, p=.843; TLI=1.00; CFI=1.00; RMSEA=.000; RMSEA 90% C.I. =.000 - .030). The convergent and discriminant validity was evaluated by the method proposed by Fornell and Larcker (1981). The results of this analysis are shown in Table 7-1. As required, every factor's average variance extracted (AVE) was higher than all the proportions of shared variance (SV) with any other factor. ERS showed very high convergent validity. On the contrary, DRS had the lowest convergent validity, with an AVE slightly below .50. However, as mentioned, its AVE was higher than all it's SV's. Moreover, all three DRS indicators showed standardized loadings that were in a similar acceptable range (.68, .63 and .70), and none of the indices of local misfit (modification indices and standardized residuals) was significant, as can be expected given the good overall model fit. Therefore, the measurement model was accepted as a valid representation of the response style measures.

TABLE 7-1

DISCRIMINANT AND CONVERGENT VALIDITY ANALYSIS

SV/AVE/*r*	ARS	DRS	ERS	MRS
ARS	0.54	*0.28*	*0.70*	*-0.55*
DRS	0.08	0.45	*0.65*	*-0.38*
ERS	0.50	0.42	0.80	*-0.11*
MRS	0.30	0.14	0.01	0.59

The diagonal shows the average variance extracted (AVE). Below the diagonal, shared variance (SV) is reported. Above the diagonal, the correlation coefficients are shown.

Cross-mode comparison

Response styles were compared across three modes (paper and pencil, telephone and online) by specifying a multi-group MACS. Nested models were specified to test for measurement and structural invariance. A review of the measurement invariance literature led to the following procedure to assess whether the subsequent null hypotheses of invariance should be rejected (Cheung and Rensvold 2000; Jöreskog 1971; Vandenberg and Lance 2000; Little 1997; Meredith 1993; Ployhart and Oswald 2004; Steenkamp and Baumgartner 1998). First, the chi square difference test was evaluated (Jöreskog 1971). Since the sample sizes employed in the current study were well above 200, the chance of rejecting the model based on chi square values could be expected to be substantial (Marsh, Balla and McDonald 1988). If chi square was insignificant, the invariance hypothesis was accepted. If it was significant, the change in CFI was evaluated (Comparative Fit Index; Bentler 1990): a decrease in CFI equal to or higher than .01 led to rejection of the null hypothesis of invariance (Cheung and Rensvold 2002). Additionally, in cases where the chi square difference test was significant, it was evaluated to what extent indicators of local misfit, modification indices (M.I.'s) and standardized residuals (s.r.'s), showed consistent patterns of significant values (Steenkamp and Baumgartner 1998; Little 1997). If the decrease in CFI was smaller than .01 and the local misfit indices did not show consistent patterns, the hypothesis of invariance was accepted.

TABLE 7-2

FIT INDICES FOR NESTED MODELS TESTING CROSS-MODE MEASUREMENT INVARIANCE OF RESPONSE STYLES

	Chi square test			Chi square difference			Alternative fit indices				
Model	χ^2	df	p	χ^2 diff	df diff	p diff	TLI	CFI	RMSEA	LO 90	HI 90
A. Unconstrained	179.1	90	< 0.001				0.988	0.994	0.027	0.021	0.033
B. Metric invariance	206.9	106	< 0.001	27.8	16	0.033	0.988	0.994	0.027	0.021	0.032
C. Scalar invariance	294.4	122	< 0.001	87.5	16	< 0.001	0.983	0.989	0.033	0.028	0.037

DF =degrees of freedom; χ^2 diff = χ^2 difference test; DF diff=degrees of freedom of the χ^2 difference test

FINDINGS CROSS-MODE RESPONSE STYLE COMPARISON

The MACS model was fitted to the data using the ML estimator. Skewness for all indicators was below 1, kurtosis below 2; hence it was concluded that the normality assumption was approached to an acceptable extent (a common cutoff criterion is skewness < 2 and kurtosis < 7; Finney and Distefano 2006). The model test results are presented in Table 7-2.

Although the chi square value for the unconstrained model (model A) was significant, the alternative indices had acceptable values (see Table 7-2) and there were no indications of particular misspecifications. The model was gradually constrained further by imposing subsequent levels of invariance. To evaluate invariance, the procedure outlined above was implemented. Imposing metric invariance (model B), resulted in a slight decrease in fit, as evidenced by the chi square difference test which is significant at the .05 but not the .01 level. The alternative fit indices remained stable, and there were no indications of local misfit induced by the constraints. Therefore, metric invariance was accepted.

Imposing scalar invariance (model C), resulted in a statistically significant deterioration in fit (see Table 7-2). The decrease in CFI was less than .01, however, and the RMSEA confidence intervals of model C and B overlapped. The indices of local misfit indicated that the misspecifications were relatively small and randomly dispersed throughout the model. Moreover, releasing one or more individual constraints did not substantially improve fit (neither did it influence the results reported below to any significant extent; this was verified). As a consequence, scalar invariance was accepted.

TABLE 7-3

LATENT MEANS IN THE PARTIAL STRUCTURAL MEAN INVARIANCE MODEL

	P&P		Tele					Online				
	Mean	s.e.	Mean	s.e.	E.S.	t		Mean	s.e.	E.S.	t	
ARS	0.89	0.02	0.96	0.03	0.29	3.66	***	0.86	0.02	-0.14	-1.86	
DRS	0.71	0.02	0.71	0.02	-0.01	-0.03		0.67	0.02	-0.24	-2.91	**
ERS	0.31	0.01	0.30	0.01	-0.07	-0.95		0.28	0.01	-0.18	-2.57	*
MRS	0.19	0.01	0.15	0.01	-0.47	-6.38	***	0.21	0.01	0.12	1.46	

P&P=paper and pencil; Tele = telephone. ARS = acquiescence response style; DRS = disacquiescence response style; ERS = extreme response style; MRS = midpoint responding; s.e. = standard error of the mean estimate; E.S. = effect size of the mean difference with the P&P mode. * = significant at the .05 level; ** = significant at the .01 level; *** = significant at the .001 level.

Based on the scalar invariant model, the latent response style means could be compared (Steenkamp and Baumgartner 1998). To obtain estimates in the original scale of the response style indicators, the intercept of the highest loading indicator of each response style was set to zero and the latent factor mean was freely estimated. The resulting mean estimates, standard errors and t difference tests are given in Table 7-3. Additionally, Table 7-3 provides an estimate of effect size (Thompson and Green 2006), expressing the mean difference of both the Telephone and the Online group with the P&P group scaled in standard deviations of P&P (which served as the reference group). Most importantly, the Telephone group showed a lower level of MRS, in support of H1d. The difference was highly significant and substantial (nearly half a standard deviation of the P&P reference group). Further, the Telephone group had a higher level of ARS (H1a). Finally, the Online group showed two significant differences with the P&P group, in that it had lower levels of both DRS and ERS (contradicting Hypotheses 2b and 2c).

DISCUSSION RESPONSE STYLE COMPARISON

The results reported above show that response styles can be simultaneously operationalized as multi-indicator latent constructs in means and covariance structures (MACS) that have measurement invariance across the three modes of data collection under study: P&P, Telephone and Online.

It is interesting to relate the current results to findings by Jordan, Marcus and Reeder (1980). Using a narrow set of items (32 items related to health care) on a different format (4-point Likert items) and, consequently, more weakly operationalized response styles (Jordan et al. 1980, p. 216), these authors found indications of more ARS and ERS in a telephone survey as compared to a door-to-door survey. Since the number of response options was even, MRS was not measured in this study. The

current results suggest that MRS shows the most substantial difference between modes, with Telephone mode having a lower MRS level, in line with Hypothesis 1d. In this group, the probability of respondents choosing the neutral point of a scale is markedly smaller than in the other modes. The responses are shifted to the positive side, as reflected by the slightly higher ARS level as hypothesized (H1a). While the current data do not allow to conclusively explain the mechanism underlying this phenomenon, a plausible interpretation (discussed above) flows forth from the interaction with an interviewer that is present in the Telephone mode but not the P&P and Online modes. In particular, it is suggested that respondents may feel pressed to provide an opinionated response rather than a midpoint response, leading to lower MRS. Additionally, the presence of an interviewer might increase the perceived and/or real time pressure, which in turn leads to higher ARS (in line with McGee 1967).

As for the difference between Online and P&P, it was found that the former group had significantly lower levels of DRS and ERS, thus rejecting hypotheses 2b and 2c. While the effect sizes indicated an effect of moderate size, the statistical significance was less than those for the Telephone group. Nevertheless, the whole response style profile of the Online group pointed to a moderate way of responding, with the highest MRS and the lowest ARS, DRS and ERS. Possibly, these respondents (who are part of a panel) were most experienced in answering questionnaires and approached the task in a more routine driven way than did the other respondents. Note that the net effect of ARS and DRS led to a nearly identical expected score for the Online and P&P groups (see below). In terms of spread, on the other hand, the expected response distribution for the Online group has less heavy tails (as shown by the lower ERS value).

It is relevant to bring to mind the scaling of the response styles as shown in Table 7-3. The Telephone MRS score of 0.15 indicates that in the Telephone mode, on average 15% of respondents will select the middle response option in response to a random item, as opposed to respectively 19% and 21% in the P&P and Online groups. In other words, approximately one fourth of the midpoint responders in the P&P or online groups might have chosen a different (probably more favorable) option in the Telephone mode. This is a substantial difference, especially when taking into account the effect that MRS can have on observed scores (Baumgartner and Steenkamp 2001). The other observed differences, though some were significant, are less substantial. The Telephone group showed higher levels of ARS, and the Online group showed lower levels of DRS. Translated to expected observed scores (by considering NARS = ARS-DRS and adding NARS to the midpoint, i.e. 4), these results indicate that the average item score in the P&P, Telephone and Online modes would be 4.18, 4.25 and 4.19. This indicates that in the Telephone mode, scores would be expected to be inflated due to the combined effect of ARS and DRS.

PART 2: IMPACT OF RESPONSE STYLES ON A SUBSTANTIVE CONSTRUCT

In this section, the above findings are translated into hypotheses concerning expected score differences on a substantive construct and made more concrete by means of an empirical illustration. In particular, scores are studied on a latent variable measured in the same data collection as the response style indicators discussed above, and adapted to the same seven point rating format: Trust in Front Line Employees (henceforth labeled TRUST). This scale has shown good reliability and validity in several studies (Sirdeshmukh, Singh and Sabol 2002; Zeithaml, Berry, and Parasuraman 1996; Dodds, Monroe, and Grewal 1991). In the current data set, apart from being included in the same questionnaire, these items were entirely unrelated to the response style

measures: the content of the items did not overlap with any of the items in the response style indicators and the items themselves were not used in computing the response style indicators. Any relationship between the observed response style levels and the four items can therefore only be attributed to shared response style bias.

HYPOTHESES

To guide the evaluation of the illustration, based on the above findings, the following expectations can be formulated regarding the TRUST item response frequency distributions. It is anticipated that the telephone group will show the lowest frequency of midpoint responses (low MRS; see Table 7-3). This will most probably be accompanied by higher levels of moderate agreement (slightly higher ARS, but no specifically high ERS; see Table 7-3). While the online and P&P group can be expected to be more similar to one another than to the telephone group, the online group can be hypothesized to show somewhat lower levels of disagreement (lower DRS; see Table 7-3) and less heavy tails of the frequency distribution (lower ERS; see Table 7-3) than the other two groups.

ANALYSIS AND FINDINGS

The observed scores of the four items in question are shown in Figure 7-2. These bar charts clearly visualize how response style differences between modes may bias cross-mode comparisons of observed scores. As expected, the telephone group showed drastically lower frequencies of the middle response and slightly higher frequencies of favorable responses, more specifically moderately favorable responses (rather than extremely favorable responses). If the response style data had not provided clear predictions on the cross-mode differences in response distributions, the observed scores would most probably have been ascribed to real content related differences and

(post hoc) explanations might have been provided for the observations. An important question is whether the bias due to response styles would have become apparent in a measurement invariance analysis. To probe this issue, the data were subjected to a multi-group CFA in which metric and scalar invariance were checked for.

Figure 7-2

Bar charts of cross-mode frequency distributions (response percentages) for TRUST items

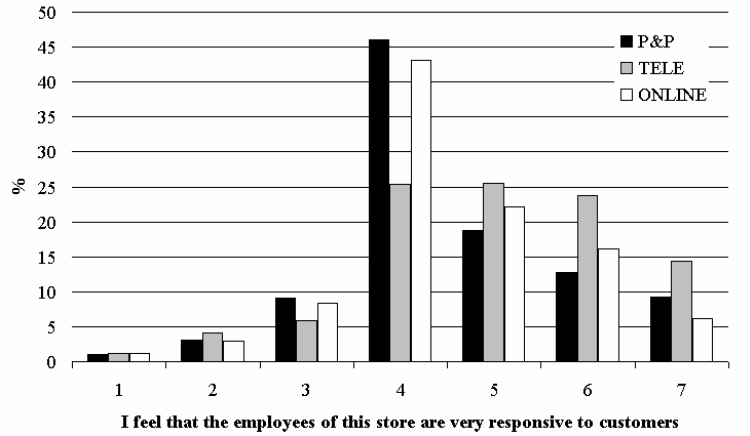


TABLE 7-4

MEASUREMENT INVARIANCE FIT TESTS FOR TRUST FACTOR

		Chi square				Chi square difference				Alternative fit indices		
	Model[a]	χ^2_{S-B}	S-B factor	df	p	χ^2_{S-B} diff	S-B factor	df diff	p diff	TLI	CFI	RMSEA
Uncorrected	A	40.5	2.12	6	< 0.001					0.956	0.985	0.106
	B	55.8	1.73	12	< 0.001	7.9	1.33	6	0.248	0.972	0.982	0.085
	C	67.7	1.49	18	< 0.001	4.7	1.03	6	0.578	0.979	0.979	0.074
Corrected	A	539.6	1.03	260	< 0.001					0.982	0.987	0.046
	B	547.4	1.04	266	< 0.001	9.1	1.39	6	0.168	0.987	0.982	0.046
	C	553.7	1.04	272	< 0.001	7.4	1.00	6	0.285	0.987	0.982	0.045

[a]Models: A = Unconstrained model; B = Metric invariance model; C = Scalar invariance model.

Uncorrected model test

To account for non-normality, the mean-adjusted ML estimator in Mplus was used (MLM; Satorra and Bentler 2001; Muthén and Muthén 2004, 2006). The resulting model fit indices and chi square difference tests are reported in Table 7-4 (taking into account the MLM adjustment factor, labeled S-B factor after Satorra and Bentler 2001).

The initial model showed a significant chi square value and rather high RMSEA, but acceptable values on the TLI and CFI. The rather high RMSEA value seems due to the fact that relatively many parameters are freely estimated while they are rather similar across the three groups: RMSEA imposes quite a substantial penalty for complexity; hence the improvement in fit for models with increasing levels of invariance (see below). There was no indication of particular misspecifications. Imposing metric and scalar invariance did not lead to a significant increase in misfit and even resulted in an improvement of the relative fit indices. Based on such data, it would be plausible for a researcher to accept metric and scalar invariance. The measurement invariance tests showed little indication of systematic bias on the indicator level. This confirms the earlier statement, based on Little (2000) that measurement invariance is no guarantee against response style bias. A logical next step based on the available data would be to test for mean differences between groups. While there was no reason to expect true differences in TRUST between the three groups, the previous response style findings suggested that probably a mean difference would appear. In particular, the Tele group can be expected to show a higher mean than the other groups, due to the slightly higher ARS combined with the substantially lower MRS: for scales with average scores above the midpoint, MRS leads to a decrease in the observed mean (Baumgartner and Steenkamp 2001). Indeed,

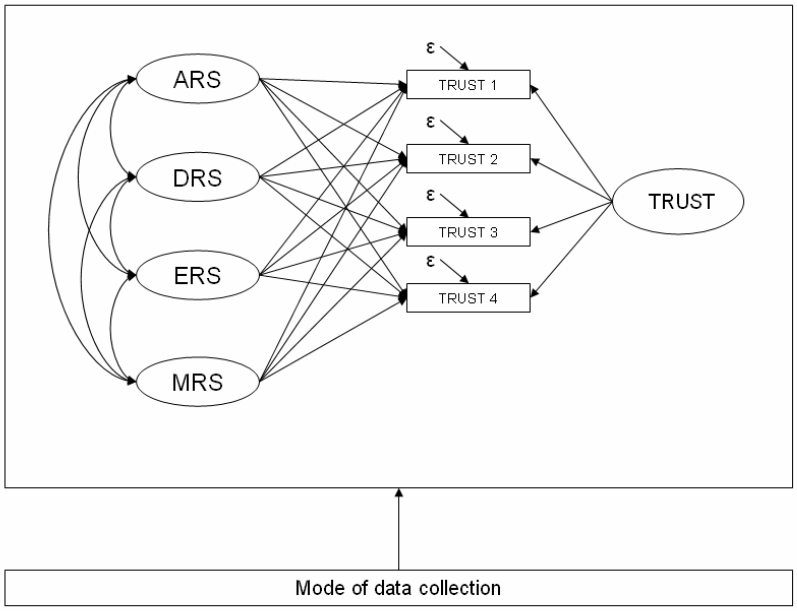
this is what happened. In a model where the mean of the P&P group was set to 0 while the means of the Telephone and Online groups were freely estimated, the following estimates were obtained. For Telephone, the mean was 0.49 (s.e. = 0.07; t = 6.85, $p < 0.001$), for Online it was 0.02 (s.e. = 0.06; t = 0.23, p = 0.388).

Response style corrected model

The above results are compared to those obtained from the TRUST factor measurement model when it was corrected for response style bias. The correction was realized by regressing the TRUST indicators on the response styles in one simultaneously estimated model, as shown in Figure 7-3.

Figure 7-3

Measurement model corrected for response styles*



*The indicators for the response style measures are not shown.

This procedure is an extension of the multiple linear regression method proposed by Baumgartner and Steenkamp (2001). These authors regressed observed scores on response style measures by means of multiple regression analysis. The obtained residuals can than be considered to be corrected scores and can be used for subsequent analyses. Applying the regression of the observed scores on response style factors in the same model that is used for the factor mean evaluation has the following advantage(s) over working with regression residuals. First, the substantive model and the measurement correction model are estimated simultaneously, so that one can evaluate and integrate the results of both models. Specifically, the loadings on both the response style factors and the substantive factor can be compared and the relative contribution of both types of factors to the variance in the observed score can be assessed. Second, the proposed procedure does not need the assumption that the response styles are measured without error, an assumption that is implicitly made when using response style scores as the independent variable in a multiple regression analysis. The residuals resulting from such regression contain error variance from the response style estimates. Similarly, the multiple regression residuals method assumes measurement invariance of the response style factors across groups, an assumption that can be explicitly tested with the current approach. The main disadvantage of the currently proposed method is that the resulting model is complex and requires the estimation of a large number of parameters. For the current purposes, however, correct estimates are more important than ease of implementation.

The model used here thus combines the four-item factor measurement model for TRUST with the four response styles model (as depicted in Figure 7-1, assuming scalar invariance as established above) by loading/regressing the four TRUST items on the four response style factors. This model reflects the conceptualization of the

response style bias affecting the measurement items rather than the latent construct (Podsakoff et al. 2003). Since the items were closely related, one regression weight was estimated for the four items together for each group and response style. In other words, all four items in the Telephone group had the same regression weight on ARS, for example. They did have a different weight for DRS, however, as they did for ARS in the Online group. Consequently, the regression of the observed scores on the response styles required the estimation of twelve additional parameters, i.e. one regression weight per response style (four in total) per group (three in total). It was confirmed that the substantive conclusions did not depend on the choice for this specific restriction.

The tests for invariance of the TRUST loadings and intercepts are shown in the lower half of Table 7-4. While the model showed statistically significant misfit (a significant chi square test), the alternative fit indices compared favorably to commonly used cutoff criteria (Hu and Bentler 1999) as well as to those obtained for the uncorrected model. Moreover, the model estimates had meaningful values and the indices of local misfit provided little reason to suspect model misspecification. As further shown in the lower half of Table 7-4, measurement invariance seems a plausible assumption in the corrected model, as it appeared to be in the uncorrected model. Most importantly, however, the corrected model allows assessing how response styles have biased the estimates in the uncorrected TRUST measurement model. This is done by comparing the standardized factor loading estimates and AVE's of TRUST in both models, as shown in Table 7-5. As expected, the results indicate that the factor loadings are inflated due to response styles. While the overestimation by 9 to 11% in the Online and P&P groups might be considered acceptable by some, the overestimation of loadings by 17% in the Telephone group is clearly problematic (Bandalos 2006). In

line with this, the AVE for TRUST in the Telephone group dropped from .62 to .45 after correcting for response styles. This finding indicates that an important part of the variance shared by the indicators is due to response styles, not content. Without the response styles diagnosis, one would have been easily led to erroneously accept the apparently high convergent validity of the TRUST factor.

TABLE 7-5

COMPARISON OF CORRECTED AND UNCORRECTED FACTOR STRUCTURE

	Uncorrected model		Corrected model		Bias	
	Loading	AVE	Loading	AVE	Loading	AVE
P&P	0.82	0.69	0.74	0.56	11%	22%
Tele	0.78	0.62	0.67	0.45	17%	36%
Online	0.84	0.71	0.77	0.60	9%	19%

Loading = average standardized factor loading; AVE = average variance extracted.
Bias = ((uncorrected estimate – corrected estimated) / corrected estimate).

The mean levels on the TRUST factors were also compared across the modes in the corrected model. In this model, the respective mean estimates (and standard error) for the Telephone and Online groups respectively were 0.39 (s.e. = 0.30; t = 1.284, p = 0.175) and 0.05 (s.e. = 0.26; t = 0.057, p = 0.398). Contrary to the finding in the uncorrected model, the mean difference of the Telephone group was no longer significantly different from the P&P reference group. This finding is due to two interrelated corrections: the mean estimate is deflated by subtracting the bias due to response styles, and additionally the lower reliability of the TRUST factor is reflected in the larger standard error of the mean estimate. While the latter phenomenon may seem undesirable, it is clear that the apparent reliability of the mean estimate in the uncorrected model is artificial and does not provide a valid foundation for inferences.

Similar deteriorations in reliability have been observed by Watson (1992) and Mirowsky and Ross (1991) and are in line with theoretical expectations (Green and Hershberger 2000). To conclude, the apparent convergent validity of TRUST in the Telephone group and the apparent mean difference of this group with the other modes seem to be due to response style bias. This makes sense in light of the absence of any appealing a priori reason to expect substantive differences between the three samples.

DISCUSSION

The above case illustrates (1) how response styles may inflate factor loadings and thus artificially create nice looking factor structures, as proved to be the case in the Telephone group; (2) how response styles may lead to spurious mean differences between modes of data-collection; (3) a possible method for implementing a response style correction within a MACS model; and (4) that measurement invariance tests are sometimes not fit to discover response style bias. The finding that response styles inflate factor loadings and bias mean estimates is not new (e.g. Baumgartner and Steenkamp 2001). However, the current study clearly demonstrates that such effects may bias cross-mode comparisons. Different modes of data collection may show different levels of apparent convergent validity and artificial cross-mode mean differences may be caused by response styles. Establishing measurement invariance, while undoubtedly useful, does solve this problem. Based on the current findings, it is argued that it may be necessary to base response style indicators on information that is not also used in the substantive model of interest. It is important to note the advantages that the approach discussed in the current paper offers over alternative procedures. A common approach in a MACS context has been to use the indicators of a substantive model to also operationalize a method or response style factor (Billiet and McClendon 2000; Podsakoff et al. 2003). Such approach has several problems.

First, a common method factor corresponds to a measure of net acquiescence response style (NARS = ARS - DRS). Consequently, differences in response styles other than ARS and DRS may go unnoticed. Second, the absence of response style factor specific indicators leads to a problem of indeterminacy, in that variance shared by indicators of the same construct may be attributable to the construct and/or response styles while there is no way of knowing which of the two is the true source. While this issue may partly be addressed by the use of scales containing reversed items (Billiet and McClendon 2000), in many instances such scales are not available (Baumgartner and Steenkamp 2001). Moreover, there is good reason to believe that respondents' responses to reversed items are inconsistent for reasons other than acquiescence (Wong, Rindfleisch, and Burroughs 2003).

As touched upon above, one can also regress observed scores on response style scores using multiple regression analysis. While the currently proposed method may seem like a straightforward extension of such approach, it is important to note the advantages the use of multiple indicators in a MACS framework bring along in this context. In addition to accounting for measurement error and allowing for measurement invariance tests, this method allows one to simultaneously assess the relative contribution of the response style factors and the substantive factors to the observed scores' variance.

Finally, it was illustrated in the current study that measurement invariance tests are not necessarily effective in diagnosing response style differences (Little 2000). In addition, if measurement invariance tests do find differences in loadings and/or intercepts, this methodology does not provide any tools for correction. In other words, measurement invariance testing is limited to diagnosis, and does not offer corrective methods. The current procedure does allow for such correction, but it should be noted

that measures that are heavily contaminated by response style variance will have low consistency after correction, which is reflected in higher standard errors (Mirowksy and Ross 1991). Clearly, designing studies in such a way as to avoid response style bias is preferable by far to trying to solve the contamination post hoc.

CONCLUSION

In this paper, a comparison was made of levels of response styles across three common modes of data collection, using a means and covariance structure (MACS). Among other things, the MACS approach allows for better estimates of relevant parameters and assessment of measurement invariance of response style measures across groups of respondents. The model was applied to a data set consisting of balanced samples of respondents in three modes of data collection, (1) self-administered paper and pencil questionnaires (P&P), (2) telephone interviews (Tele), (3) self-administered online questionnaires (Online), using measures of four response styles: acquiescence response style (ARS), disacquiescence response style (DRS), extreme response style (ERS), and midpoint responding (MRS). The operationalization shows measurement invariance across the three groups, which makes it an appealing method for use in similar settings. The findings of the mean comparison are important and show that telephone interview data should be handled with caution, in that they show systematic bias as compared to other data. This conclusion is in line with findings by Jordan, Marcus and Reeder (1980) in a different context and using a more limited set of measures. It is apparent from the current data that telephone interviews result in lower MRS, and slightly higher levels of ARS. Telephone survey participants seem to tend to use rating scale options away from the midpoint. As for the Online data, slightly lower levels of DRS and ERS were found,

which tentatively seems to point towards a more moderate way of responding to items.

The present findings need to be taken into account in future research that aims to compare theoretical models across online and offline contexts. For such comparisons, the use of self-administered paper and pencil questionnaires and self-administered online questionnaires is recommended, and not telephone interviews. It is not suggested that Telephone interviews are invalid as such, but rather that they are probably incomparable to the visual/manual self-administered formats. Moreover, it is advisable to test for response style differences between modes of data collection before proceeding to the actual comparisons between online and offline measurement and structural models.

Based on this research, the following procedure for cross-mode marketing research is recommended. (1) Include a set of unrelated items in a questionnaire, or try to distill these from parts of the questionnaire that are not needed for the research question at hand. The latter is often possible when several research topics are pooled in one questionnaire. (2) Diagnose response styles by means of the 4-style typology ARS, DRS, ERS, MRS (as illustrated in Figure 7-1). (3) If significant differences in response style levels are apparent from the previous step, include response style factors in the model (as illustrated in Figure 7-3).

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

Some limitations of the current study provide opportunities for future research. First, like most response style research (e.g., Greenleaf 1992a; Baumgartner and Steenkamp 2001), the scope is limited to one type of measurement scale. All items used and discussed in this paper are seven-point Likert items. It might be interesting to study how scale format is related to response styles, and – possibly in a later stage – how

scale format interacts with mode of data collection. It might be argued that seven point scales are too complicated for use in the Telephone mode. However, different lines of research have led to the recommendation to use seven point items (e.g. Cox 1980), usually without specifying the specifics of data collection. The main advantage of seven point scales is that they produce scores that can be reasonably treated as interval scaled data (Bollen and Barb 1981), although this approach has been questioned by some (e.g. Babakus, Ferguson and Jöreskog 1987). However, as Cronbach (1950) suggested, part of the reliability of rating scales with many options may be due to the increasing response style variance they induce. This issue needs further clarification.

Also, it would be enlightening to study what causes different levels of response styles in different data collection settings. The current study focused on establishing the presence of a mode effect on response styles, but did not determine the causal process that led to this difference. To further probe this issue, experimental work is called for, implementing independent manipulations of the factors that are confounded in the modes of data collection as they are used in real life, like visual versus auditory presentation, self-administration versus interviewer interaction, and self-paced versus interviewer-paced timing.

CHAPTER 8: RESPONSE STYLES AS SATISFICING STRATEGIES (EMPIRICAL STUDY 5)

CHAPTER OUTLINE

The current study focused on four response styles and how these styles relate to the optimizing-satisficing dimension, where optimizing was operationally defined as time-intensive differentiation of responses to items that are homogeneous in form but heterogeneous in content. The relationship between each response style and optimizing was allowed to vary across respondents, such that response styles could be satisficing strategies, optimizing strategies, or both (but then for different groups of respondents). Two major satisficing strategies were observed, one combining stylistic extreme and midpoint responding, the other concentrating mainly on midpoint responses. Important implications for the meaning of observed responses to questionnaire items are discussed.

INTRODUCTION

In survey research, not every response is equally informative. Even if two respondents give identical responses to the same item, these responses may carry completely different meanings. Much depends on the process that led to the observed response. One respondent may have put quite some effort into understanding the question, bringing to mind relevant information, integrating this information into an overall judgment, evaluating the acceptability of the judgment and finally reporting it in the form required by the questionnaire (Tourangeau, 1984). Another respondent might well skip these processes on the whole and give a midpoint response, regardless of the specific content of the questionnaire item. The former respondent can be said to be optimizing, while the latter is said to be satisficing (Krosnick 1991). Satisficing and optimizing can be thought of as the polar opposites of the same continuum (Krosnick 1999) and are henceforth used to refer to the same dimension.

The more respondents are satisficing, the less their responses are driven by content, and the more their responses reflect the respondents' response styles (Jackson and Messick 1958; Jackson 1967), defined as individual difference variables reflecting disproportionate use of specific response options regardless of content (Rorer 1965). It is not clear, though, which specific response styles respondents resort to when satisficing. It is important to know how response styles are related to satisficing because this would enable better prediction and understanding of the potential effects of respondent motivation, as well as a better diagnosis of which response styles are problematic in questionnaire data, in that they are indicative of suboptimal information processing. Based on such diagnosis, it could be decided to disregard certain respondents and/or to statistically correct for the response styles (Greenleaf 1992a). Therefore, the main objective of the current study was to investigate the

relationship between satisficing and the use of the following response styles: Acquiescence Response Style (ARS), Disacquiescence Response Style (DRS), Extreme Response Style (ERS) and Midpoint Response Style (MRS).

It is plausible that different individuals use different satisficing strategies (i.e. strategies aimed at minimizing the investment of time and cognitive effort from the part of the respondent). For example, some respondents may simplify their task by using only the midpoint and both extremes (MRS and ERS), while others may stick to agreeing with items regardless of content (ARS). Differences in satisficing strategies are not directly observable but reveal themselves in the association (captured, e.g., by a regression function) between a given response style and the satisficing-optimizing dimension. The functional form of this association may therefore be different across individuals, depending on the individuals' satisficing strategies. Consequently, estimating a single relationship between satisficing and a specific response style for all respondents might be inadequate and potentially misleading (Wedel and DeSarbo 1995). For that reason, the current study investigated heterogeneity in the relations of the four response styles ARS, DRS, ERS and MRS to the satisficing-optimizing dimension. This enabled the distinction between different satisficing strategies. To this end, structural equation mixture modeling (SEMM; Jedidi, Jagpal and DeSarbo 1997) was used.

CONCEPTUAL BACKGROUND

THE PSYCHOLOGY OF SURVEY RESPONSE

Elaborating on an earlier proposal (Tourangeau 1984), Tourangeau, Rips and Rasinski (2000) discussed a model that outlines the psychological processes involved in responding to a survey question. While these processes need not occur in a fixed

sequence and it is not even needed that all of them occur, they are presented in the most logical order. (1) In the comprehension stage, the respondent attends to the instructions and the question, and creates an internal cognitive representation of these stimuli by activating relevant concepts and identifying what information is being sought. (2) In the retrieval stage, the respondent retrieves from long term memory the information that is needed to provide an answer to the question. (3) In the judgment stage, the respondent integrates the material that was retrieved from long term memory into an overall judgment. (4) In the response stage, the judgment is translated into one of the available response options, edited for acceptability, and reported. In-depth execution of all these processes is a demanding task. In this regard, Tourangeau et al. (2000, p. 8) remarked "*Although some processes may be mandatory, others are clearly optional – a set of cognitive tools that respondents can use in constructing their answer. Exactly which set of processes they carry out will depend on how accurate they want their answer to be, on how quickly they need to produce it, and on many other factors.*"

SATISFICING THEORY

The latter issue is the focus of satisficing theories. Feldman and Lynch (1988) and Feldman (1992) posited that responses to questionnaires are subject to the principle of cognitive economy. This principle states that respondents will not use resources in the development of judgments, beliefs, etc. unless some reason for doing so exists. If respondents minimize the amount of resources they invest in formulating a response to a questionnaire item, they are said to be satisficing. If they put in the resources required to arrive at an optimal response, they are optimizing (Krosnick 1991). Feldman (1992) showed that response formulation can be flexible, allowing the respondent a considerable degree of freedom in the amount of effort s/he is willing to

spend. For a response to be maximally determined by content, in-depth execution is required of the processes of comprehension, retrieval, judgment and response formulation. This takes time and considerable cognitive effort (Narayan and Krosnick 1996). Consequently, respondents may resort to a more shallow processing strategy, such that the impact of item content is diminished. The less a response is driven by content, the more it is driven by an individual's response style (Jackson and Messick 1958; Jackson 1967). The question then becomes what specific style respondents resort to when satisficing. With that question in mind, four tendencies that have been defined as prevalent response styles in the literature seem relevant: Acquiescence Response Style (ARS), Disacquiescence Response Style (DRS), Extreme Response Style (ERS) and Midpoint Response Style (MRS). These styles are behavioral tendencies, while Optimizing is conceptualized as their potential motivational antecedent. Before discussing response styles as satisficing strategies, an operational definition of the optimizing-satisficing dimension is introduced that allows empirically studying the relation between the styles and this dimension.

AN OPERATIONAL DEFINITION OF OPTIMIZING

Since the defining aspect of optimizing is the cognitive process that happens between perception of the stimulus (the item) and performance of a response, optimizing cannot be directly observed or measured. Therefore, it is proposed to operationally define optimizing as the co-occurrence of two necessary components of the process. One component, Time-On-Task (TOT), is related to the resources invested by the respondent. The other component, response differentiation (DIFF), is related to the resulting response pattern. Both of these are discussed in turn.

Schaeffer and Presser (2003, p. 68) mention conserving time and energy as the respondent's purpose of satisficing. In other words, time and energy are main inputs

that can vary as a function of satisficing. Since mental energy is currently impossible to measure directly, the most reasonable alternative indicator for optimizing on the input side is time, more specifically Time-On-Task (TOT). In research on the psychology of survey response, TOT has traditionally been used as an indication of the effort that is exerted by the respondent in formulating a response (Osgood 1941; Matell and Jacoby 1972; Tourangeau, Rips and Rasinski 2000, p. 94-95). It is reasonable to consider a certain level of TOT as a necessary condition for optimizing, in that optimizing requires respondents to go through the extended process of comprehension, information retrieval and information integration. Respondents usually do not report a readily available response (labeled the "file drawer model"; Wilson and Hodges 1992), but more often than not construct a judgment based on several elements retrieved from memory, an activity that necessarily takes time (Tourangeau, Rips and Rasinski 2000). Though a necessary condition, a certain level of TOT is not a sufficient condition for optimizing, in that high TOT may also be due to other factors related to an individual's speed (perceptual and cognitive capabilities and/or transient factors like situational demands on cognitive resources).

Hence, to ensure that a respondent is optimizing, it does not suffice to observe her/his TOT. Therefore, response differentiation is proposed as an output related criterion that complements TOT. Response differentiation (DIFF) can be defined as the extent to which a respondent provides diverse responses to items that are homogeneous in form but heterogeneous in content. The extent to which respondents differentiate their responses to heterogeneous items has been opposed to satisficing by Herzog and Bachman (1981). These authors observed that straight line responding (as an extreme case) and responding without much differentiation (as a moderate case) tend to increase with decreasing levels of respondent motivation due to fatigue. Clearly, if a

respondent answers a series of unrelated items, high differentiation between their responses can be expected in case they truthfully respond to each individual item. Differentiation can hence be considered a necessary condition for optimizing. Again however, it is not a sufficient condition, in that respondents may respond randomly to a series of items. Random responding has been identified as a time saving strategy adopted by some respondents (Krosnick 1991; Drolet and Morrison 2001), though it seems to be much more uncommon as a strategy than is low differentiation (Herzog and Bachman 1981; Knowles 1988; Drolet and Morrison 2001; Kraut et al. 1975).

In sum, both TOT and DIFF are necessary conditions to classify response behavior as optimizing, but they are not sufficient conditions. Taken together, however, it is implausible that high levels of both could signify anything other than optimizing or random variation. The latter possibility can be accounted for by focusing on the common variance shared by multiple indicators.

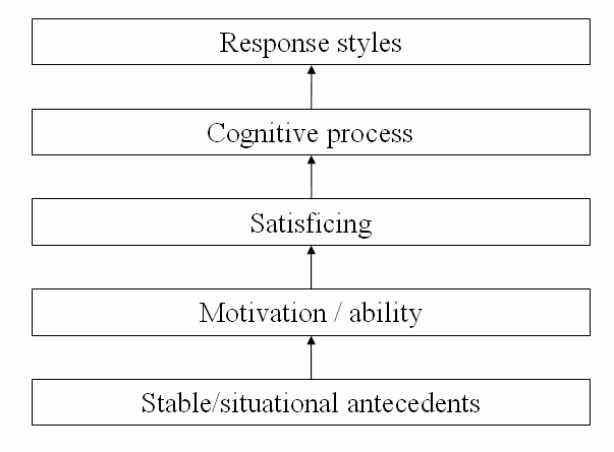
In sum, the following operational definition of optimizing is proposed: optimizing is the time-intensive differentiation of responses to items that are homogeneous in form but heterogeneous in content. The next section discusses how response styles are expected to relate to optimizing.

RESPONSE STYLES AS SATISFICING STRATEGIES

Since little evidence presents itself that empirically relates response styles directly to satisficing, the hypotheses are built on indirect evidence. In particular, a causal chain is assumed in which response styles are the consequence of a cognitive process driven by satisficing, in turn dictated by a respondent's motivation and/or ability, which finally relates to stable background variables and/or situational effects. Schematically, this chain can be summarized as in Figure 8-1.

Figure 8-1

Causal schema of satisficing and response styles



Evidence linking the stable/situational antecedents to response styles may be indicative of a satisficing process. As mentioned above, different respondents may have different satisficing strategies. Consequently, some response styles may be hypothesized to have both a positive and a negative relation to optimizing. The idea is not to suggest that both relations co-exist within the same respondent, but rather that a positive relation may be present for some respondents, a negative relation for others. Contradictory hypotheses are therefore not considered to be mutually exclusive.

Since ARS and DRS are commonly treated together in the literature, these response styles are discussed together. For example, Cheung and Rensvold (2000) and Greenleaf (1992a) consider the net effect of ARS and DRS, labeled Net Acquiescence Response Style (NARS) by Baumgartner and Steenkamp (2001). ERS and MRS have also been discussed jointly in parts of the literature (e.g., Johnson et al. 2005), and are discussed together here as well.

ARS AND DRS

It is quite common in the literature to relate ARS to satisficing due to lack of cognitive sophistication (Krosnick 1991; Knowles and Nathan 1997) or superficial processing (Couch and Keniston 1960), sometimes in rather belittling or negative terms. Peabody (1966), for example, blamed scale designers for the presence of ARS in certain scales due to presenting complex statements "to those who are simple-minded". In line with this, education level, a common proxy for cognitive sophistication (Schuman and Presser 1981; Narayan and Krosnick 1996), has been negatively related to ARS (Gage, Leavitt and Stone 1957; Greenleaf 1992a; Jackson and Pacine 1961; McClendon 1991a; Mirowsky and Ross 1991; Narayan and Krosnick 1996; Schuman and Presser 1981; Watson 1992). Mirowsky and Ross (1991) also found that ARS was highest among younger and older people, which may indicate that ARS is at least partly due to limitations in working memory capacity (Knauper 1999). Based on the above evidence, the following hypothesis is put forward:

H1a: ARS is negatively related to optimizing.

Some studies have failed to replicate the relation of (N)ARS with cognitive ability and/or education level (Bachman and O'Malley 1984; McClendon 1991b; Ray 1979). Further, it has been found that NARS is not related to the serial position of items in a questionnaire (Knowles 1988; Kraut, Wolfson and Rothenberg 1975), although an initial study by Clancy and Wachsler (1968) was inconclusive in this regard. This indicates that ARS and DRS are not related to respondent fatigue, as pointed out by Kraut et al. (1975). Greenleaf (1992a) states that NARS is closely related to content driven responding. Further, Bachman and O'Malley (1984) note that ARS and DRS correlate positively. According to these authors, this suggests that these styles do not

merely represent a directional bias, i.e. a preference for either favorable or unfavorable responses, but that both probably also are related to differentiated response behavior. In other words, a respondent who carefully and truthfully answers all questions will have a non-zero level of ARS and DRS, and more careful content-driven responses may result in higher levels of both response styles. Hence the two following hypotheses:

H1b: ARS is positively related to optimizing

H2a: DRS is positively related to optimizing

There seems to be little indication that DRS might be a satisficing strategy. It seems to be the one response style that is considered to be related to criticalness and thoughtful processing of item content in a consistent way, and hence does not seem to be a viable satisficing strategy for any respondent (Couch and Keniston 1960; Elliot 1961). Hence, no such hypothesis is proposed.

MRS AND ERS

Kraut, Wolfson and Rothenberg (1975) found that MRS increases with the serial position of items, while ERS decreases. These results indicate that fatigue (i.e. a situational antecedent of motivation, and hence satisficing) may lead respondents to increasingly stick to the middle option while making less use of extreme response options. This is in line with suggestions that MRS may among others be due to lack of interest (Schuman and Presser 1981). Furthermore, although Kalton, Roberts and Holt (1980) observed no effect of demographics on MRS, Narayan and Krosnick's (1996) meta-analysis showed MRS to be negatively related to education level. Thus, these lines of research seem to converge on the conclusion that MRS most probably is a satisficing strategy, while ERS may be related to optimizing.

H3a: ERS is positively related to Optimizing

H4a: MRS is negatively related to Optimizing

On the other hand, Osgood (1941) observed that on a seven point scale both midpoint responses and extreme responses take relatively less time to be formulated as compared to more moderate reactions. The author interpreted this as midpoint and extreme responses requiring little cognitive effort. Further, Osgood formulated the impression that such response pattern combining MRS and ERS seems to be more prevalent among less cognitively sophisticated individuals. In line with this, older respondents are believed to have higher ERS levels (Hamilton 1968; Greenleaf 1992a). Arthur and Freemantle (1966) interpret ERS as due to the absence of temperance of initial impulsive responses, which again points to lack of cognitive processing. Based on this evidence, the following hypothesis is proposed:

H3b: ERS is negatively related to Optimizing

Little empirical results link MRS to optimizing and it seems that MRS may be related to satisficing alone, and no such hypothesis is proposed.

As touched upon above, the hypotheses contain propositions that are in direct contradiction with one another (H1a versus H1b and H3a versus H3b). While such opposite predictions could be viewed as mutually exclusive and hence competing hypotheses, this investigation starts from the view that different respondents may use different satisficing strategies. In other words, it is seen as a possibility that some of the alternative hypotheses may both be true, albeit for different individuals. For that reason, this study investigated the presence of latent classes defined by different latent regressions between response styles and satisficing/optimizing.

PROFILING VARIABLES

Previous research has found effects of age, education level and gender on response styles. These variables will therefore be included as antecedents of the different

satisficing strategies. This allows profiling respondents with different satisficing strategies in terms of these key demographics.

METHODOLOGY

DATA

Items and respondents

A questionnaire was designed consisting of a randomly selected set of items, which made it particularly well-suited for measuring response styles. More specifically, 112 items were sampled from the Marketing Scales Handbook by Bruner, James and Hensel (2001) and Measures of Personality and Social Psychological Attitudes by Robinson, Shaver and Wrightsman (1991) and brought together in a questionnaire using seven point Likert scales. All respondents received the items in the same order. Data were collected from an online consumer panel, resulting in a 41.7% response rate. The sample represented a cross-section of the Belgian online population in terms of age, gender and education level. 511 cases were useful for further analysis. In this sample, the average age was 43.5 years (s=14.7), on average, respondents had had 7.0 (s=2.0) years of formal education after primary school, and 44.4% respondents were female.

RESPONSE STYLE INDICATOR CALCULATION

For all respondents alike, the items were divided into three sets, corresponding to subsequent parts of the questionnaire: the first part consisted of page 1 to 3 (48 items), the second of page 4 and 5 (32 items), the third of page 6 and 7 (32 items). The three sets were used to compute as many indicators for every response style (ARS, DRS, ERS and MRS). For ARS, the agreements per set of items were counted, weighting a seven as three points, a six as two points, and a five as one point. A similar method

was applied to obtain DRS measures (cf. Baumgartner and Steenkamp 2001). ARS and DRS indicators reflect the expected deviation from the midpoint due to ARS or DRS respectively if means would be computed based on the item responses. ERS indicators reflect the number of extreme responses (1 or 7) divided by the number of items. Similarly, MRS indicators reflect the number of midpoint responses (4) divided by the number of items in the set.

MEASURE OF OPTIMIZING

An optimizing variable was constructed based on the operational definition of optimizing as time-intensive differentiation of responses to items that are homogeneous in form but heterogeneous in content. The operationalization of optimizing needs to make use of observed outcomes rather than direct measures. Hence, it is appropriate to model optimizing as a latent construct. The indicators need to represent the co-occurrence of TOT and DIFF, which can most easily be achieved by using the product of TOT and DIFF for a set of items, and using the product terms as the indicators of Optimizing (OPTIM).

Operationally, the following measures are proposed for TOT, DIFF and OPTIM, given a random set of items that are homogeneous in form but heterogeneous in content. The indicators were first computed per page (of which there were seven in the current data) and than averaged to obtain an indicator per item set (of which there were only three).

First, TOT is the number of minutes spent on answering the items. As a proxy for TOT, the time was used during which the web questionnaire page was open on the computer of the respondent. Some respondents might have left open a page while doing something else. To avoid that such observations excessively altered the frequency distribution, a plot was created of all percentiles for each time measure (the

X axis showed the percentile number, from 1 to 99; the Y axis had the observed value, from 0 through the maximal observation). For all pages, the plot of percentiles clearly showed a sudden jump around the 98^{th} percentile from 7 minutes to 20 minutes. This was taken as an indication that time after this point is disproportionately longer than time taken by other respondents and could be assumed to be spent not only on the task of responding. All values beyond this point were set to the percentile value after which the sudden increase occurred.

For each of the three item sets, a differentiation (DIFF) indicator was computed as $\Pi(1 + f(o))$, where $f(o)$ is the frequency of endorsing response option o taken across a set of items and Π refers to the product taken over all response options o. The indicators were then rescaled to a 0-1 range by subtracting the minimum possible value and dividing by the maximum possible value.

For each page, the TOT indicator was multiplied by DIFF. To obtain specific OPTIM indicators per item set (rather than per page), the natural log of the average page indicators plus one was taken (to compensate for the skewing effect of taking product terms; Tabachnick and Fidell 1996), resulting in three OPTIM indicators. The three OPTIM indicators had a Cronbach's alpha of 0.85. Further evidence of internal consistency and validity are provided by the analyses reported below and in Appendix 8-1. Moreover, no alternative interpretations of the measure present themselves, as the combination of both differentiation and time investment quite clearly point towards optimizing (Krosnick 1991, 1999; Schaeffer and Presser 2003). Appendix 8-1 shows that DIFF is not by design related to any of the response styles under study. This appendix also discusses the scaling of OPTIM.

MODEL

The research question addressed in this study calls for the use of Structural Equation Mixture Modeling (SEMM) for the following reasons (cf. Jedidi, Jagpal and DeSarbo 1997). First, as discussed above, substantive theory supports the model in which each of the above response styles is a function of OPTIM. Second, both response styles and OPTIM are latent variables. Third, a priori segmentation is not feasible. Finally, there are clear reasons to believe that the regression functions are heterogeneous (as apparent from the mutually exclusive hypotheses above).

For each response style in isolation, a latent regression model is specified in which the response style is regressed on OPTIM. The regression parameters are class specific. Class membership is modeled as a function of the profiling variables: age, education level and sex.

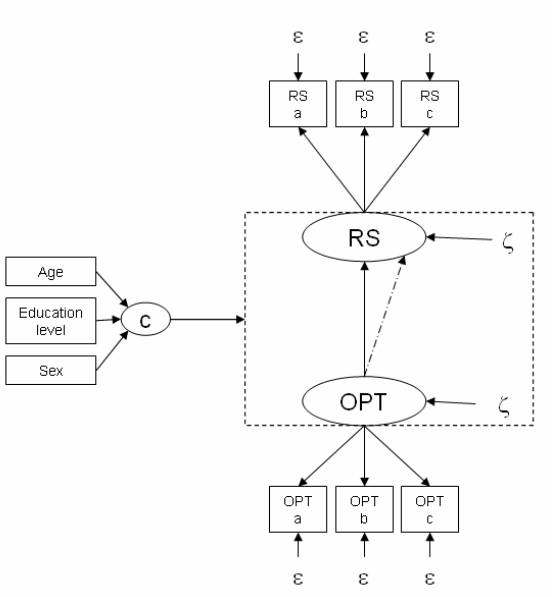
There is no reason to expect a strictly linear relationship: the hypotheses only propose generally increasing or decreasing associations, which may well level off after a given point. Therefore, the quadratic term of OPTIM is included as an antecedent, a practice that has been recommended when linearity is not explicitly hypothesized (Ganzach 1997). Including nonlinear effects in the model also is a safeguard against extraction of classes when no such latent groups are present (Bauer and Curran 2004, p. 22). Inclusion of a quadratic effect is more parsimonious and more meaningful than would be estimating only linear effects with an extra latent class to account for the nonlinearity (Rindskopf 2003).

In the model equations below, variables that have a class specific distribution are denoted by subscript k; individually varying variables are denoted by subscript i. For reasons of computational feasibility (more specifically to obtain convergence and to avoid local maxima), and because the research hypotheses were response style

specific, the model was estimated for each response style separately. Below, RS refers to the response style under study. The model is graphically shown in Figure 8-2.

Figure 8-2

Single response style Structural Equation Mixture Model



The broken arrow indicates a non-linear quadratic effect

The observed indicators y_i of response styles and OPTIM are a linear function of their related latent variables:

$$y_i = \tau + \Lambda\eta_{ik} + \varepsilon_i \qquad (1)$$

where τ is a vector of item intercepts, Λ contains factor loadings, η_{ik} is a vector of latent scores on RS and OPTIM, and ε_i contains the residual score not explained by the factors RS and OPTIM. The latent variables, in turn, are a function of the latent class variable c_i and a residual ζ_{ik}. The equation also contains a regression of RS on OPTIM and OPTIM squared:

$$\eta_{ik} = Ac_i + B_{1k}\eta_{ik} + B_{2k}\eta^2_{ik} + \zeta_{ik} \qquad (2)$$

where η_{ik} is a vector of latent factor scores on RS and OPTIM, A is a weight matrix that results in different means of RS and OPTIM across classes, c_i assigns class membership, B_{1k} and B_{2k} contain regression weights, η^2_{ik} contains the quadratic term of OPTIM, and ζ_{ik} is a residual term. The model assumes that the residuals are normally distributed with mean zero. Finally, class membership is modeled by a multinomial regression of class on demographics (x_i):

$$Ln[P(c_{ik}=1|x_i)/P(c_{iK}=1|x_i)]=\alpha_k+\Gamma_k x_i \quad (3)$$

where K is the last class, arbitrarily chosen as a reference class, α_k is an intercept term and Γ_k contains regression weights.

It would not be useful to extract classes in which the latent variables would have different meanings. Therefore, the measurement parameters in equation (1) are assumed to be equal across classes, i.e. scalar invariance is assumed. The structural regression weights, the means of OPTIM and RS, and the variance of OPTIM (see equation 2) are allowed to freely vary across classes.

Hence, the following parameters are class-specific: the mean of OPTIM and RS, the regression weights of RS on OPTIM and OPTIM squared, the variance of OPTIM, and the regression weights of class membership on age, education level and sex, and k-1 class membership variables (where k refers to the number of classes).

RESULTS

ANALYSIS

For the data-analyses reported below, the robust Maximum Likelihood (MLR) mixture estimation in Mplus 4.1 was used (Muthén and Muthén 2006). A high number of random starts was tried (with a minimum of 200 initial and 20 fully iterated starting

values) and only results for which the highest Likelihood was replicated were accepted.

TEST OF RESPONSE STYLE SPECIFIC MODELS

To determine the number of classes, the model for each RS using 1 to 4 classes was estimated. As indicators of the true number of classes in the data, two criteria were used that have been commonly applied and that have recently been validated by means of a Monte Carlo study in the context of latent class analysis and mixture growth modeling (Nylund, Asparouhov and Muthén 2006). First, the lowest value of Bayes' Information Criterion (BIC) was taken as an indication of the optimal number of classes (Jedidi et al. 1997; Lubke and Muthén 2005). The formula to obtain BIC is -2LL + q*ln (n), where LL is the log likelihood of the estimated model, q is the number of freely estimated parameters, and n is the sample size. In a Monte Carlo study, Nylund et al. (2006) found that BIC by far outperforms other information criteria in determining the true number of classes (in latent class analysis and mixture growth modeling). Before this, others have also indicated that BIC is less sensitive to sample size than the AIC, and that BIC seems not to share other indices' tendency to overextract classes (Bauer and Curran 2004; Jedidi et al. 1997; Gagné 2006). In addition to BIC, a statistical test of whether k as a number of classes is a better representation of the data than is k-1 (which is the null hypothesis), was provided. The usual chi square difference test is not valid in this context. The reason for this is that setting a class probability to zero means setting a parameter to the boundary of its allowable space: in such cases -2LL does not follow a chi square distribution (Lo, Mendell and Rubin 2001). A solution for this is provided by the Lo-Mendell-Rubin Likelihood Ratio Test (LMR-LRT; Lo, Mendell and Rubin 2001). The LMR-LRT uses an approximation to the likelihood ratio test distribution to compare models with

different numbers of classes. Nylund et al. (2006) recommend using this test as a first step in the class enumeration problem. If necessary, that is if there is doubt, one can additionally apply the bootstrap likelihood ratio test (BLRT; McLachlan and Peel 2000), which uses bootstrap samples to estimate the distribution of the LL difference test statistic.

Given the current objectives, measures of entropy were of secondary interest. The aim of the current study was not to assign respondents to classes but to identify the classes themselves. Focusing on the optimization of entropy would primarily lead to classes with a high degree of separation, which is most easily realized by extracting classes with highly different mean vectors (Gagné 2006). The interest of the current study, however, predominantly was in the different regression weight vectors. Consequently, the entropy measure was used after the class enumeration decision merely for post hoc evaluation of the degree of separation between classes. The proposed strategy given the objectives is in line with recommendations by Jedidi et al. (1997) and Bauer and Curran (2004).

The results of the analyses with different classes for ARS, DRS, ERS and MRS are presented in Table 8-1.

TABLE 8-1

MODEL FIT BY NUMBER OF CLASSES

	K	LL	q	BIC	p^b
ARS	1	230.7	26	-299.3	
	2	278.6	29	-376.4	0.069
	3	296.9	38	-356.8	0.141
	4	No convergence			
DRS	1	418.7	26	-675.3	
	2	449.2	29	-717.5	0.003
	3	465.3	38	-693.6	0.203
	4	478.6	47	-664.0	0.107
ERS	1	1277.9	26	-2393.7	
	2	1422.3	29	-2663.8	0.000
	3	1451.1	38	-2665.3	0.115
	4	1464.9	47	-2636.6	0.437
MRS	1	1591.3	26	-3020.4	
	2	1645.9	29	-3111.0	0.000
	3	1661.9	38	-3086.8	0.216
	4	No convergence			

K = number of classes; LL= log likelihood; q = number of parameters; BIC = Bayesian Information Criterion; p^b = Lo-Mendell-Rubin adjusted LRT test for k-1 (H_0) versus k classes.

For ARS, DRS and MRS, both the minimal BIC value and significant LRT probabilities led to the conclusion that there are two classes in the data, although for ARS the LRT test had only a marginally significant p-value (i.e. .05<p<.10). An inspection of the model estimates provided further support for the presence of two classes for the ARS model, however, as will be discussed in more detail below. For ERS, the BIC value and the LRT test pointed towards a three class and a two class solution respectively, although the BIC difference between the two- and the three-

class solution was trivial. To address the uncertainty, an additional Bootstrap Likelihood Ratio Test was carried out based on 100 bootstrap draws. The resulting approximate p-values were 1.000 for the null hypothesis of 2 versus 3 classes, and 0.000 for the null hypothesis of 1 class versus 2 classes. Thus, the results pointed out the two-class solution as the optimal model.

In sum, these findings indicated the presence of two latent classes defined by separate regression functions for each response style on OPT. The specific estimates for each latent class are given in Table 8-2. Since the meaning of the quadratic effect is dependent on the scaling of the independent variable, it may be helpful to look at the scatter plots in Figure 8-3a through 8-3d. The entropy values for ARS, DRS, ERS and MRS were .60, .57, .73 and .69. In reading these results, one should keep from reification of the classes and remember that the classes and class memberships are model specific. Class membership indicates that a given individual observation most probably is drawn from a specific multivariate distribution. Class assignments are hence far from deterministic.

From the results in Figure 8-3a and in the left hand columns of Table 8-2, it is apparent that over two thirds of respondents showed a positive relation between ARS and OPT (H1b; ARS C1 in Table 8-2). Hence, for a majority of the respondents in this study ARS was an optimizing strategy, which means that higher levels of favorable responses for these respondents are not problematic, but on the contrary indicate a more content driven (time-intensive and differentiated) response pattern. Younger respondents have a higher probability of being in this first ARS class (see the estimate for $B_{c\text{-}age}$ under the ARS model in Table 8-2). For the remainder group, ARS was not significantly related to OPT (ARS C2 in Table 8-2), although the scatter plot (Figure 8-3a) suggested a negative relation, which would indicate that a substantial number of

respondents engaged in stylistic acquiescence responding when minimizing time and effort. In sum, the evidence in support of H1a is non-conclusive. Age is positively related to the probability of belonging to the latter class.

A somewhat similar pattern emerged for DRS: a majority of respondents tended towards higher levels of disacquiescence when optimizing their responses to the questionnaire (H2a; DRS C1 in Table 8-2). Lower levels of education were related to a higher probability of belonging to this group (see the estimate for $B_{c\text{-}edu}$ for class 1 in the DRS model). The remainder group of respondents showed no significant relation between optimizing and DRS (DRS C1 in Table 8-2) and the scatter plot showed a rather diffuse pattern for this class.

TABLE 8-2

MODEL ESTIMATES BY CLASS AND RESPONSE STYLE

	ARS						DRS					
	C1			C2			C1			C2		
	71.2%			28.8%			70.6%			29.4%		
RS regression	Est.	s.e.	t	Est.	s.e.	t	Est.	s.e.	t	Est.	s.e.	t
$B_{RS\text{-}opt}$	0.573	0.094	6.11	-0.589	0.474	-1.25	0.711	0.085	8.33	-0.149	0.225	-0.66
$B_{RS\text{-}opt^2}$	-1.781	0.523	-3.41	0.154	0.831	0.19	-1.877	0.321	-5.85	-0.164	0.343	-0.48
Mean OPT	0.000			0.155	0.082	1.89				0.187	0.051	3.65
Mean RS	0.000			0.364	0.041	8.91				0.303	0.036	8.37
Var OPT	0.026	0.008	3.34	0.072	0.021	3.47	0.027	0.005	5.12	0.069	0.017	3.99
Var RS	0.019	0.004	5.33	0.019	0.004	5.33	0.017	0.003	5.10	0.017	0.003	5.10
Class 1												
antecedents	Est.	s.e.	t				Est.	s.e.	t			
$B_{c\text{-}Age}$	-0.382	0.132	-2.89				-0.102	0.120	-0.85			
$B_{c\text{-}edu}$	0.050	0.106	0.48				-0.221	0.081	-2.72			
$B_{c\text{-}female}$	0.457	0.354	1.29				0.503	0.325	1.55			
Intercept	0.836	0.536	1.56				0.770	0.373	2.06			

	ERS						MRS					
	C1			C2			C1			C2		
	42.9%			57.1%			57.5%			42.5%		
Regression	Est.	s.e.	t	Est.	s.e.	t	Est.	s.e.	t	Est.	s.e.	t
$B_{RS\text{-}opt}$	-0.529	0.053	-10.06	0.157	0.064	2.46	0.062	0.025	2.43	0.807	0.381	2.12
$B_{RS\text{-}opt^2}$	0.674	0.170	3.96	-0.320	0.147	-2.18	-0.028	0.077	-0.36	2.940	0.689	4.27
Mean OPT	0.000			-0.151	0.030	-5.07	0.000			-0.254	0.024	-10.38
Mean RS	0.000			-0.178	0.025	-7.01	0.000			0.123	0.047	2.61
Var OPT	0.059	0.009	6.19	0.022	0.005	4.54	0.038	0.006	6.35	0.016	0.003	4.68
Var RS	0.004	0.001	5.16	0.004	0.001	5.16	0.001	0.000	4.25	0.001	0.000	4.25
Class 1												
antecedents	Est.	s.e.	t				Est.	s.e.	t			
$B_{c\text{-}Age}$	0.310	0.083	3.78				-0.284	0.085	-3.34			
$B_{c\text{-}edu}$	0.008	0.056	0.14				0.283	0.067	4.19			
$B_{c\text{-}female}$	0.016	0.248	0.07				-0.629	0.254	-2.48			
Intercept	-0.351	0.242	-1.45				0.583	0.214	2.73			

B_{YX} refers to the regression weight with Y as the dependent, X as the independent variable. Regressions are linear for the response styles on OPT, logistic for class membership on demographics. Est. = estimate; s.e. = standard error; RS = (Acquiescence, Disacquiescence, Extreme, Midpoint) Response Style; OPT = Optimizing; C = Class

Figure 8-3a: Scatter plot of ARS by Optimizing (RS specific model)

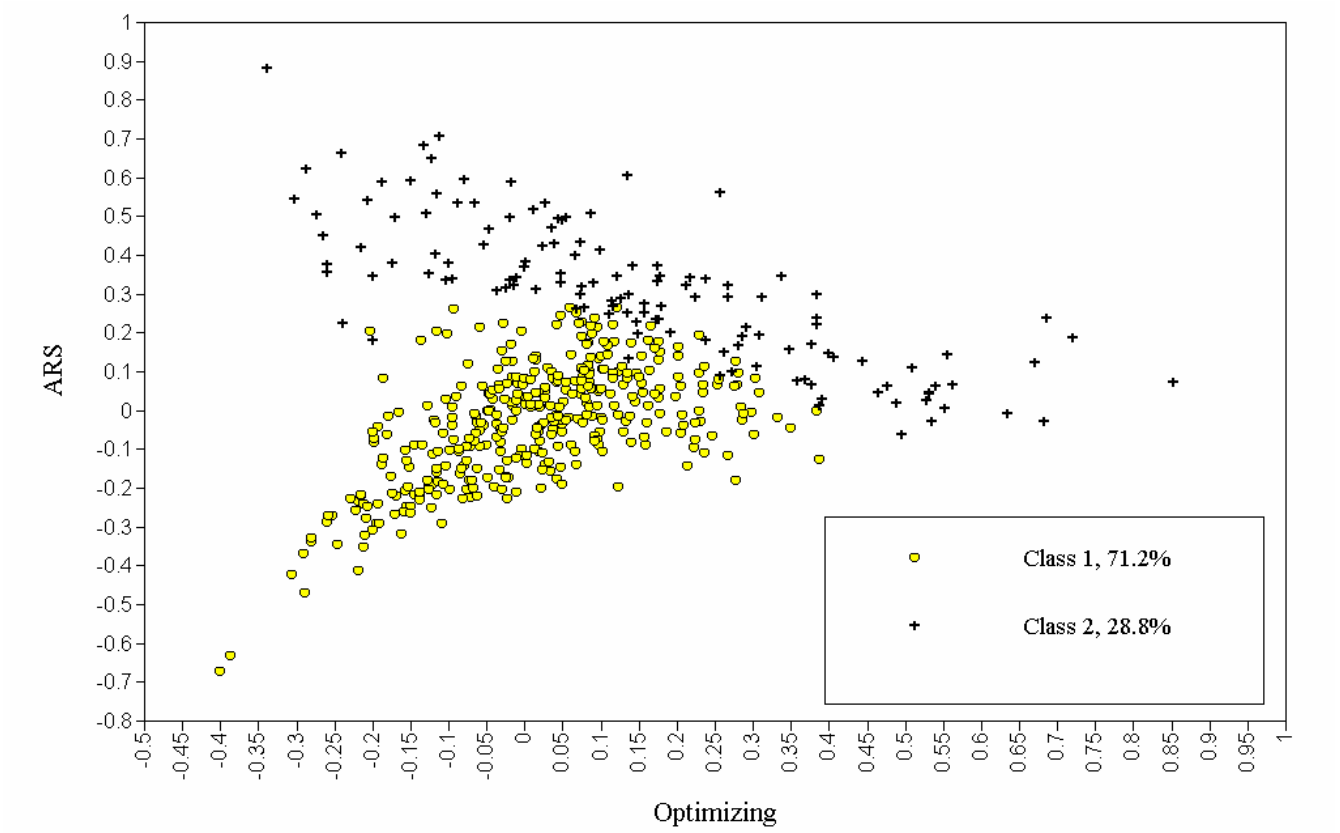


Figure 8-3b: Scatter plot of DRS by Optimizing (RS specific model)

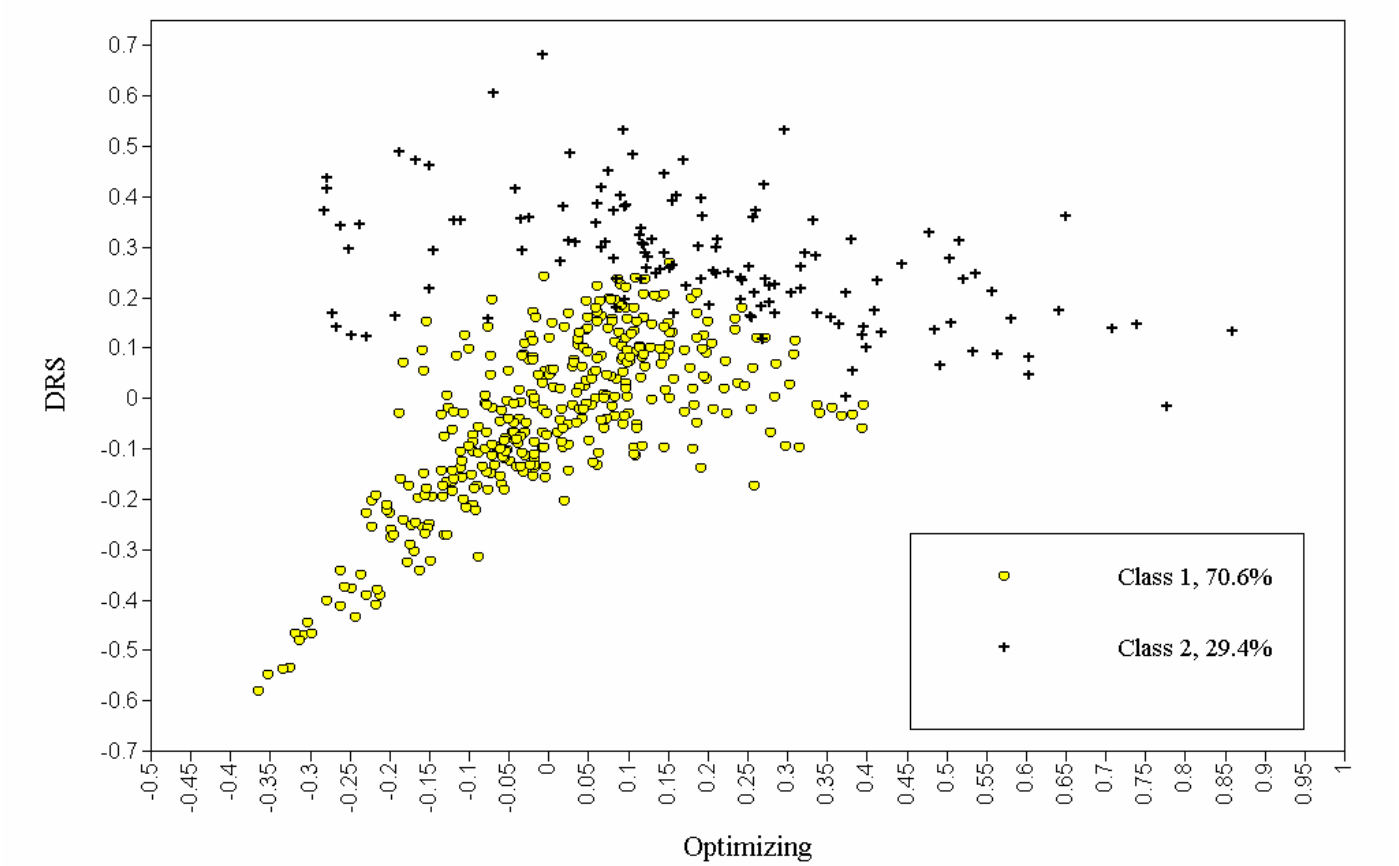


Figure 8-3c: Scatter plot of ERS by Optimizing (RS specific model)

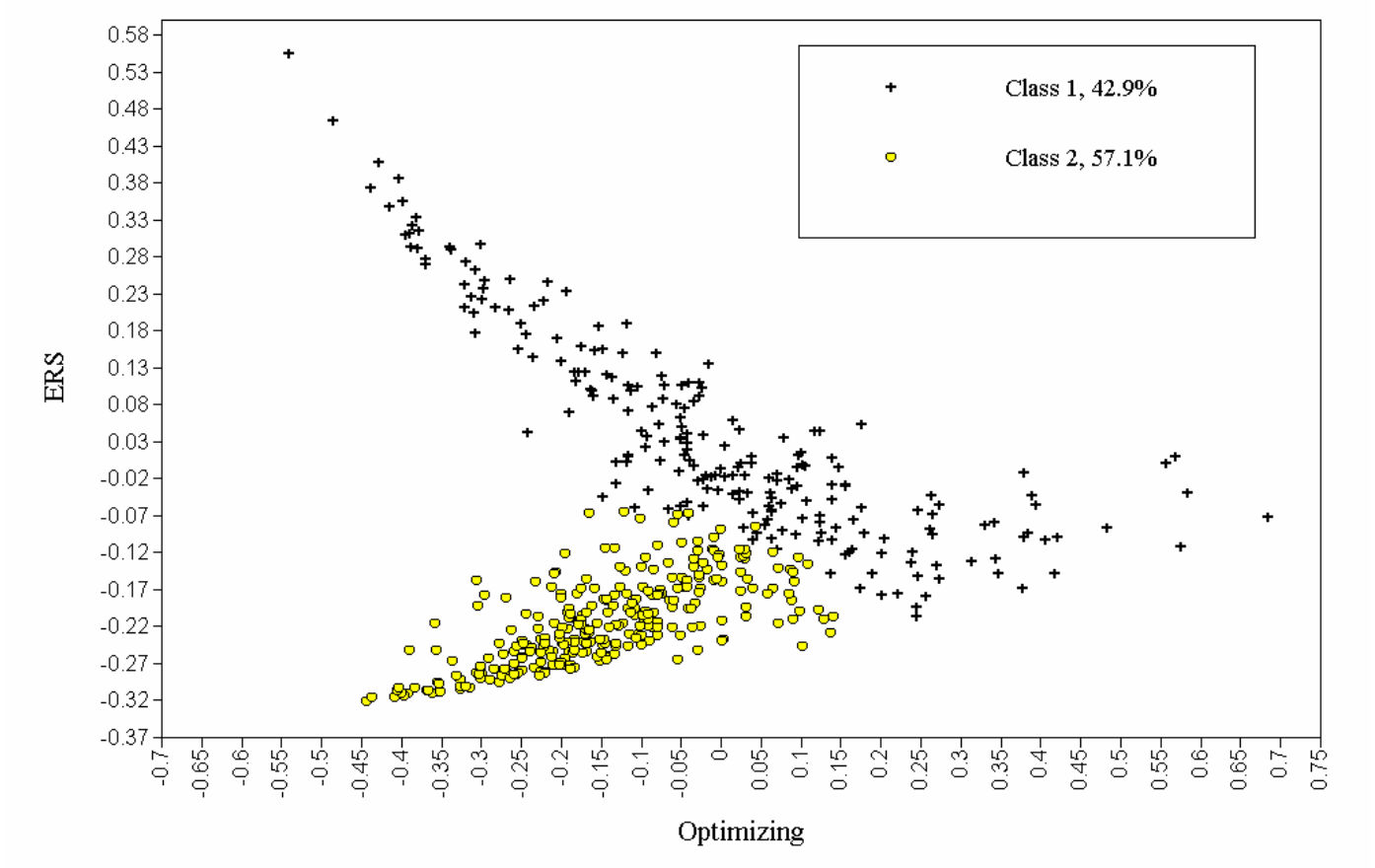
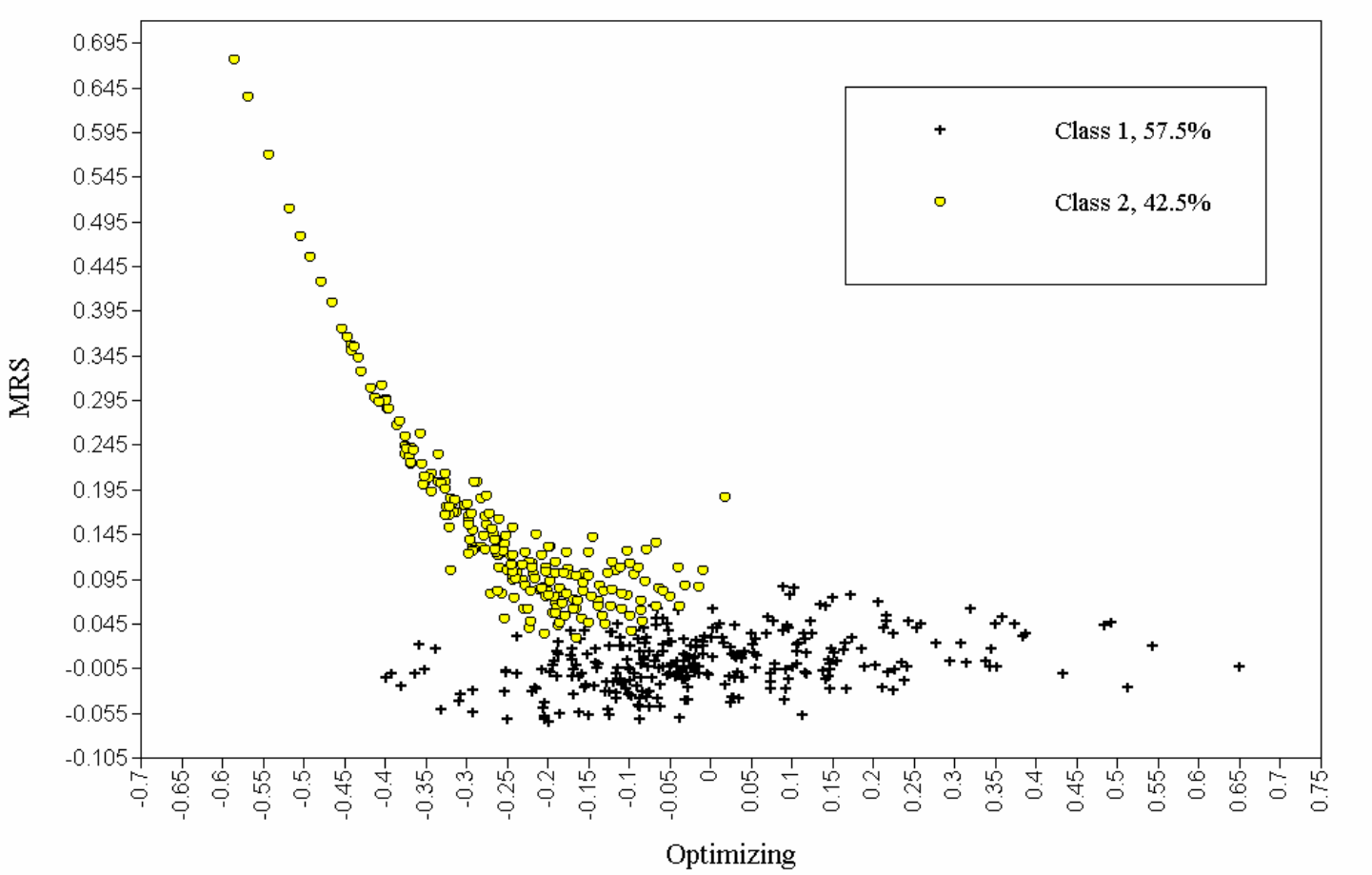


Figure 8-3d: Scatter plot of MRS by Optimizing (RS specific model)



The relation between ERS and OPTIM was captured by two regression functions each of which describes a sizable portion of the sample. For 57% of the respondents, ERS went up slightly when optimizing (H3a), while for the remainder 43% ERS clearly was a satisficing strategy (H3b). Thus it seems that for ERS in particular, it would be misleading to consider the response style as a mere nuisance factor in all cases. For some, it may be a means of differentiating responses to heterogeneous questions, while for others it may be a way of simplifying the survey task. Respondents for whom ERS served as a satisficing strategy tended to be older (see $B_{c\text{-}age}$ for C1 in the ERS model in Table 8-2).

Finally, MRS showed a pattern that was distinct from the other response styles. For 42.5% of the respondents (MRS C2 in Table 8-2), MRS was a satisficing strategy (H4a). Remarkably, this is the same 42% that was satisficing most strongly (see OPT means under the MRS model in Table 8-2), while the other 57.5% of the respondents had a relatively higher OPT score and showed a weak but positive MRS-OPT relation (MRS C1 in Table 8-2). Thus, it may be incorrect to assume that MRS is never due to optimizing. Nevertheless, the negative relation with optimizing seems dominant. Respondents have a higher chance of having a positive MRS-OPT relations if they are younger, have higher education levels and are male (see $B_{c\text{-}age}$, $B_{c\text{-}edu}$ and $B_{c\text{-}female}$ for Class 1 of the MRS model in Table 8-2). Obviously then, respondents that are older, female and have lower levels of education have a higher probability of using MRS as a satisficing strategy.

TEST OF THE FULL ADEM MODEL

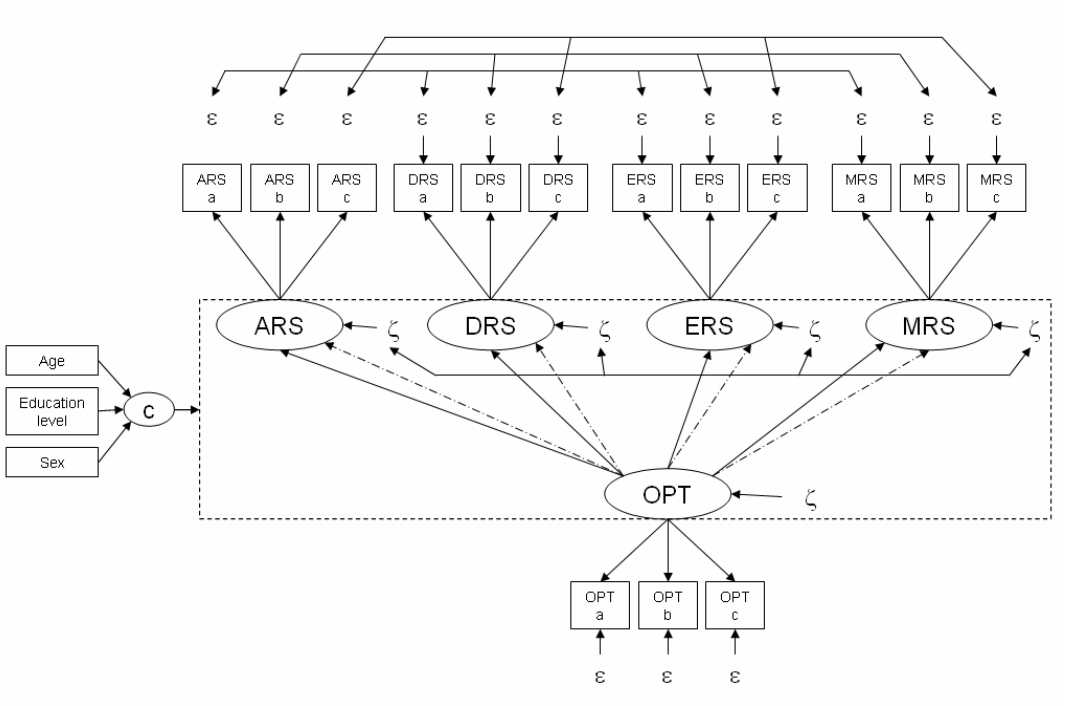
While the above findings have addressed the research hypotheses, it would be relevant to draw a profile of different satisficing strategies in terms of all four response styles simultaneously, rather than using separate models for all response styles. To explore

the relations between the four independently estimated classes, as a first step a cross-classification of the estimated class memberships across the response style specific models was made. Specifically, four dummy variables were created for membership of ARS Class 1, DRS Class 1, ERS Class 2 and MRS Class 2. The resulting phi coefficients are .26 for ARS-DRS; .52 for ARS-ERS; .53 for DRS-ERS; .26 for ARS-MRS; .33 for DRS-MRS; and .26 for ERS-MRS. All these coefficients are significant at the .001-level. The cross-classification suggests that the latent classes, though obtained in independent analyses, are related. Specifically, respondents from ERS-class 2 (ERS as optimizing) seem most probable to also belong to ARS-class 1 (ARS as optimizing), DRS-class 1 (DRS as optimizing), and MRS class 2 (MRS as satisficing). Most other respondents belong to ERS-class 1 (ERS as satisficing), ARS and DRS classes 2 (ARS/DRS as satisficing) and MRS class 1 (MRS as a neutral or optimizing style). In other words, ARS, DRS and ERS optimizing are positively related among one another, while being negatively related to MRS optimizing. Based on these indications a two-class structural equation mixture model was estimated in which ARS, DRS, ERS and MRS were simultaneously regressed on OPTIM. This model was labeled the ADEM model (for ARS, DRS, ERS, and MRS) and is depicted in Figure 8-4. For the two-class model, with 99 free parameters, LL=4864.658, BIC = -9111.92, LMR LRT $p< 0.0001$, entropy = .772. Almost half (49.2%) of the respondents were assigned to class 1, the remainder 50.8% to class 2.

These indices compared well to a three class solution: with 121 free parameters, LL=4917.237, BIC = -9079.874, entropy = .638; LMR LRT $p = 0.4009$. This comparison provided additional support for the presence of two classes in the data. The estimates for the two class model are given in Table 8-3, the scatter plots in Figure 8-5a, b, c and d.

Figure 8-4:

ADEM Structural Equation Mixture Model



In Figure 8-4, broken arrows indicate quadratic effects

TABLE 8-3

PARAMETER ESTIMATES FOR THE ADEM TWO CLASS MODEL

		Class 1			Class 2		
		49.2%			50.8%		
		Est.	s.e.	t	Est.	s.e.	t
Regression weights	$B_{ARS\text{-}opt}$	-0.376	0.074	-5.06	-0.018	0.489	-0.04
	$B_{ARS\text{-}opt^2}$	0.411	0.226	1.82	-2.400	0.937	-2.56
	$B_{DRS\text{-}opt}$	-0.113	0.065	-1.74	0.929	0.51	1.82
	$B_{DRS\text{-}opt^2}$	0.161	0.210	0.77	-0.756	0.863	-0.88
	$B_{ERS\text{-}opt}$	-0.519	0.049	-10.64	0.605	0.242	2.50
	$B_{ERS\text{-}opt^2}$	0.782	0.171	4.56	0.821	0.489	1.68
	$B_{MRS\text{-}opt}$	-0.174	0.029	-5.99	0.700	0.293	2.39
	$B_{MRS\text{-}opt^2}$	0.331	0.095	3.47	3.923	0.721	5.44
Means	ARS	0.000			-0.073	0.068	-1.08
	DRS	0.000			0.026	0.082	0.31
	ERS	0.000			-0.098	0.039	-2.52
	MRS	0.000			-0.008	0.027	-0.31
	OPT	0.000			-0.202	0.024	-8.50
Variances	ARS	0.026	0.003	7.66	0.026	0.003	7.66
	DRS	0.021	0.003	7.96	0.021	0.003	7.96
	ERS	0.005	0.001	5.79	0.005	0.001	5.79
	MRS	0.005	0.001	7.23	0.005	0.001	7.23
	OPT	0.052	0.008	6.47	0.013	0.003	4.63
Class 1 antecedents	Intercept	-0.111	0.184	-0.60			
	$B_{c\text{-}Age}$	0.267	0.084	3.17			
	$B_{c\text{-}edu}$	-0.023	0.061	-0.38			
	$B_{c\text{-}female}$	0.100	0.232	0.43			

B_{YX} refers to the regression weight with Y as the dependent, X as the independent variable. Est. = estimate; s.e. = standard error; RS = (Acquiescence, Disacquiescence, Extreme, Midpoint) Response Style; OPT = Optimizing

Figure 8-5a: Scatter plots of ARS by Optimizing (Full ADEM model)

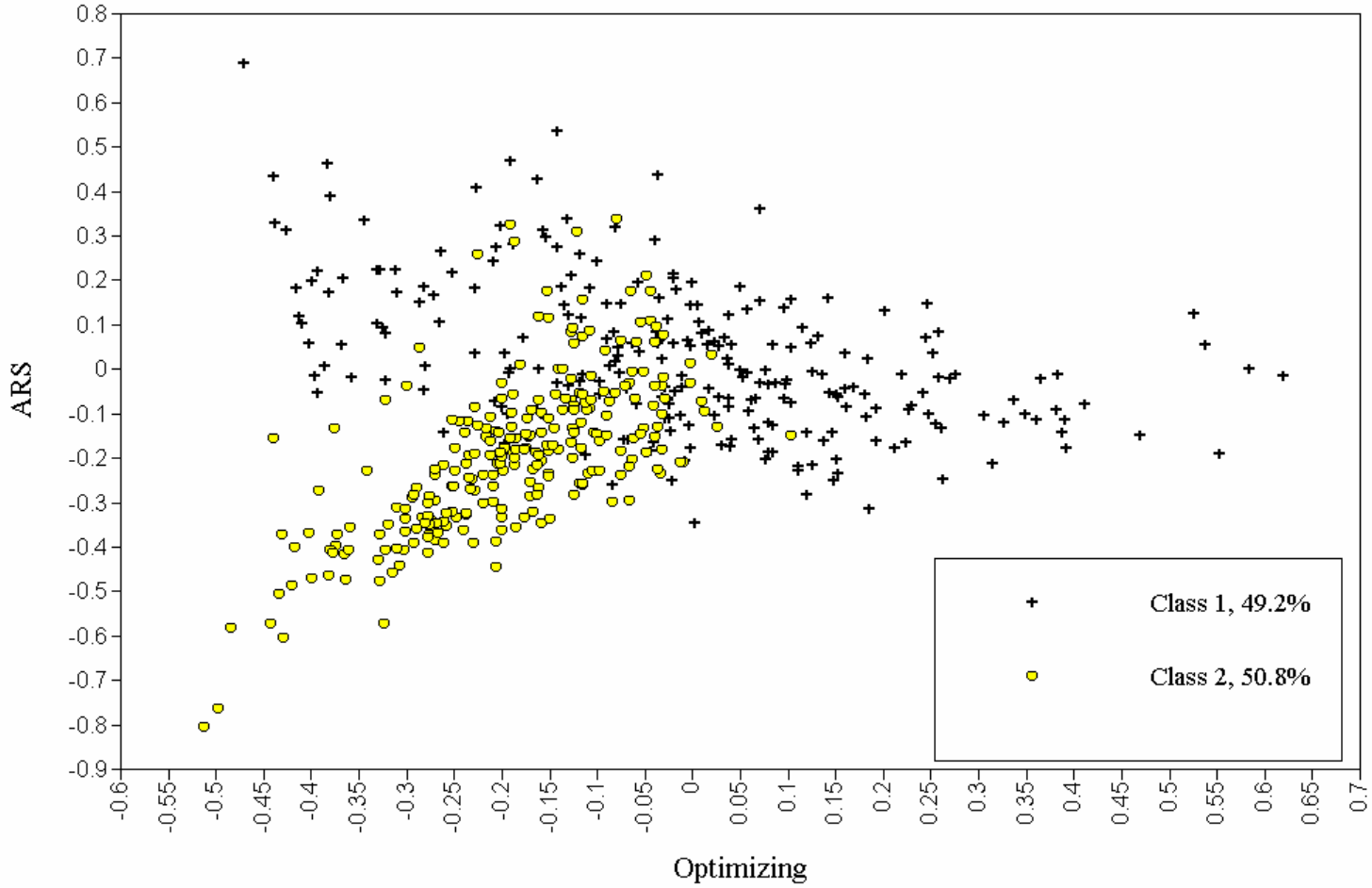


Figure 8-5b Scatter plots of DRS by Optimizing (Full ADEM model)

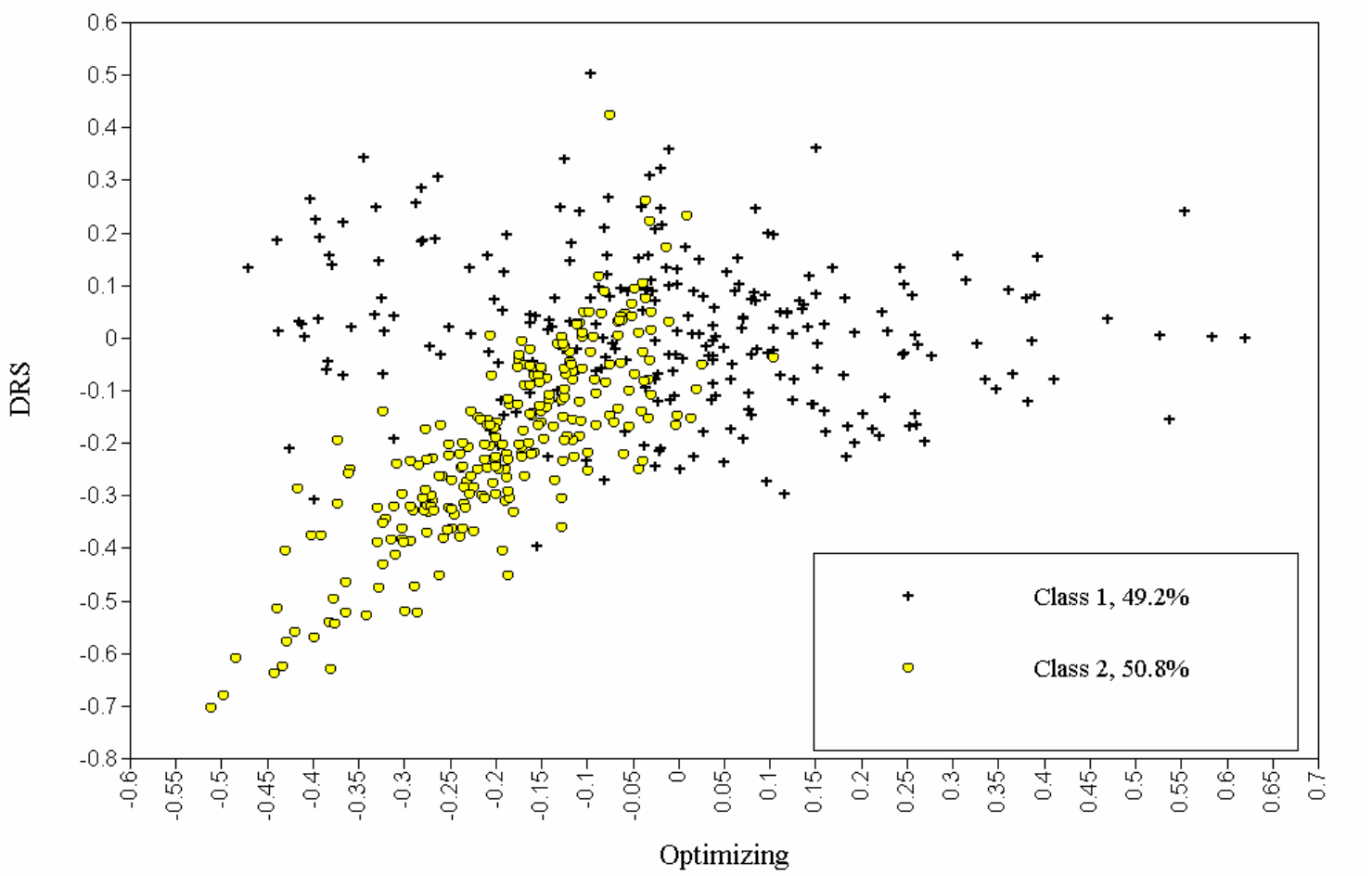


Figure 8-5c: Scatter plots of ERS by Optimizing (Full ADEM model)

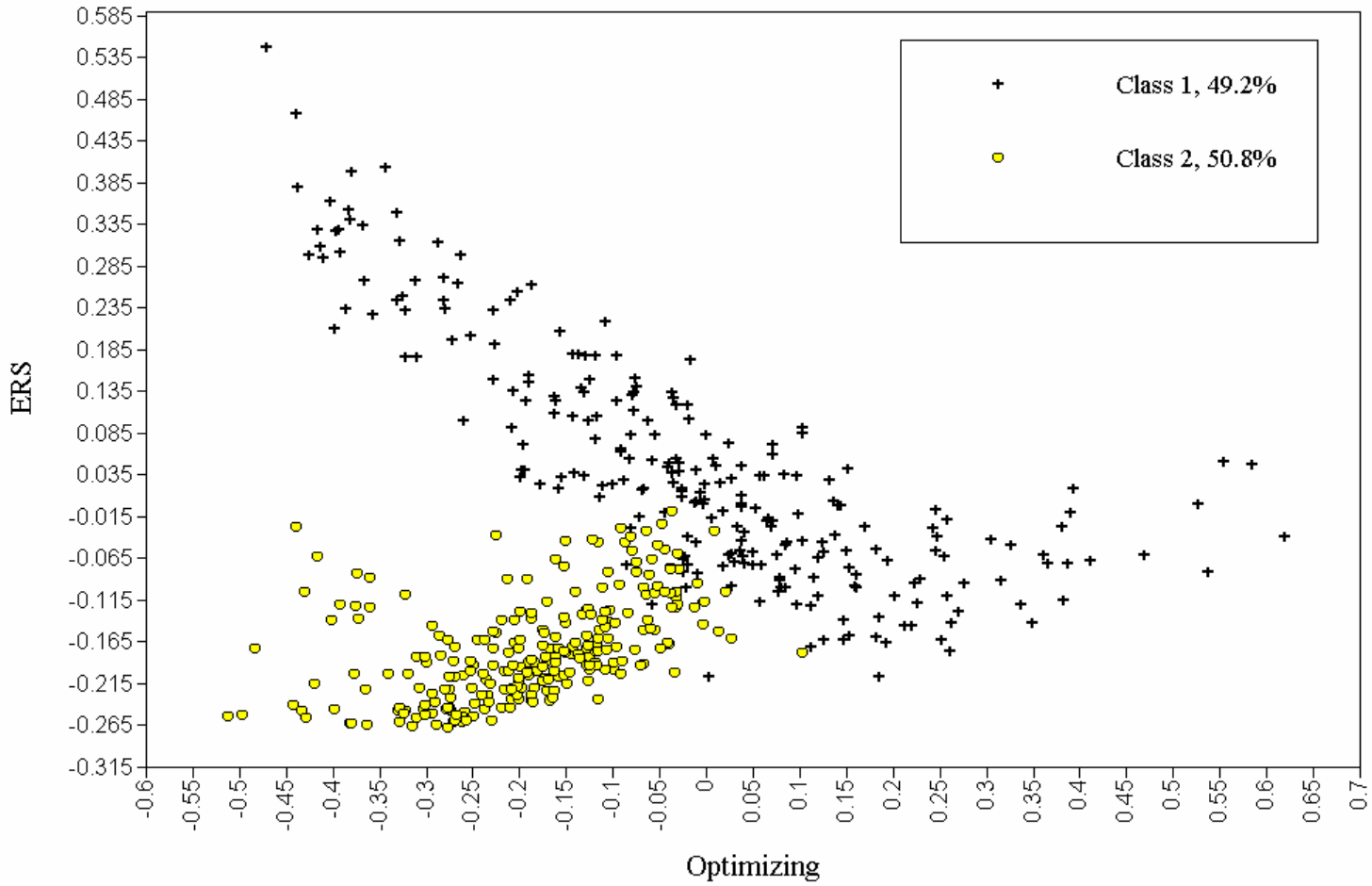
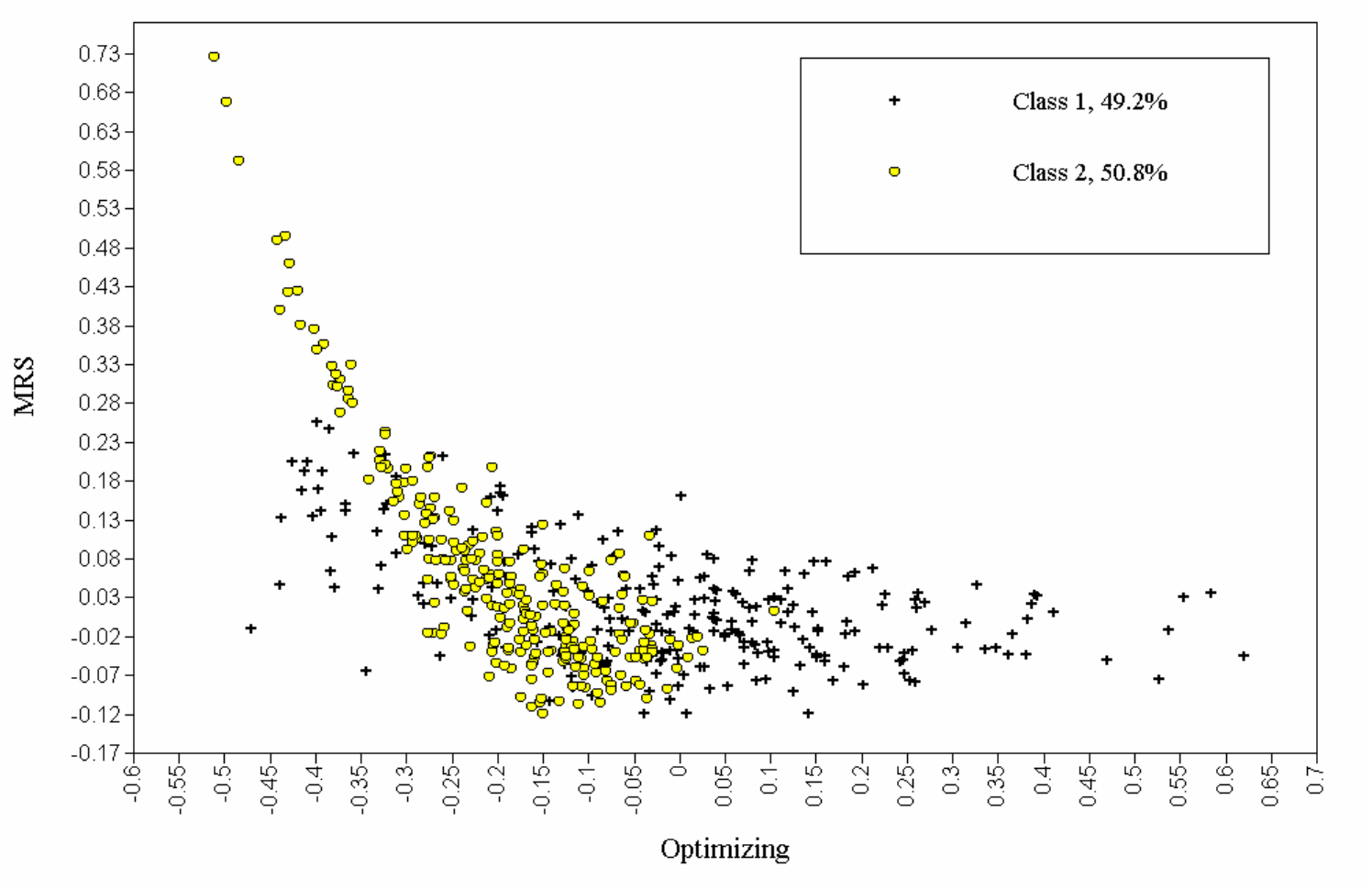


Figure 8-5d: Scatter plots of MRS by Optimizing (Full ADEM model)



From the plots the ERS optimizing class (ADEM class 1) and ERS satisficing class (ADEM class 2) that also emerged from the earlier analyses are readily recognizable.
In ADEM class 1, ERS was a satisficing strategy, as were ARS and DRS (though the latter only marginally significantly). This is apparent from the negative relationship between these 3 response styles and optimizing. For class 1, ERS showed the most clear-cut regression scatter plot. ERS seems to be the driving variable behind the latent class segmentation. The observed association of ARS and DRS with OPTIM may well be due to ERS: every extreme response expresses either extreme agreement or extreme disagreement by definition and this way directly affects the ARS and DRS scores. Further, for class 1, MRS showed a weakly negative relationship with OPTIM. Note that, though the regression weights seemed to be in the same range as those for class 2, the quadratic effect of OPTIM weighted heavier in class 2 due this class being situated predominantly on the negative side of the OPTIM dimension. Given the estimates in Table 8-3, for OPTIM scores below -.18, the quadratic effect of MRS becomes stronger than its linear effect. Consequently, the positive quadratic effect captures the declining trend visible in the left most part of the scatter plot.
In ADEM class 2, ERS, ARS and DRS were all optimizing strategies, in that their association with OPT was generally positive (see Figure 8-5a, b, c). Note again that the quadratic effect for ARS might be misleading at first sight (Table 8-3, class 2, B estimates for ARS). MRS appeared as an outspoken satisficing strategy in this class, given its negative association with OPT (see Figure 8-5d).
Age was a significant antecedent of class membership, in that older respondents had a higher chance of belonging to class 1. In other words, older respondents tended to satisfice by stylistically checking both the extreme response options and the midpoint,

while younger respondents more often tended to satisfice by stylistic midpoint responding only.

IMPACT OF SATISFICING STRATEGIES ON RESEARCH

To assess how the two overall satisficing strategies affect observed scores, six equally sized segments were created by splitting the two ADEM classes into an optimizing segment, a medium segment and a satisficing segment, as shown in Table 8-4.

TABLE 8-4

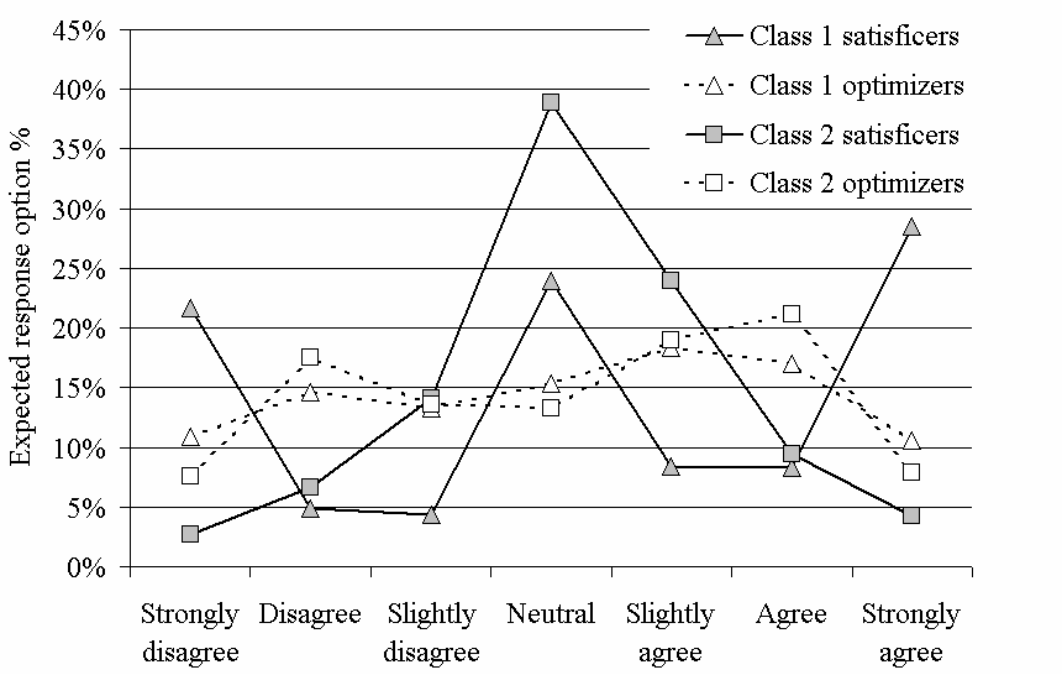
SEGMENTATION BY ADEM CLASS AND OPTIMIZING LEVEL

		ADEM satisficing strategy class	
		Class 1	Class 2
Optimizing level	Low	Class 1 satisficers	Class 2 satisficers
	Medium	Class 1 midgroup	Class 2 midgroup
	High	Class 1 optimizers	Class 1 optimizers

Expected response frequency distributions for the optimizing and satisficing segments of both classes are given in Figure 8-6 (class membership is based on most likely posterior class assignment).

Figure 8-6

Response profiles for optimizers versus satisficers by latent ADEM class



For each of both classes, this graph indicates the most likely response frequency distribution across a wide range of items, regardless of content, when respondents are optimizing versus satisficing. Clearly, the distributions look dramatically different across the satisficing segments in class 1 and class 2. For respondents who are optimizing, response frequencies look largely the same across classes.

This also implies that across the two satisficing segments, a given response option seems to have a different meaning. Henceforth, for ease of reference class 1 satisficers are labeled trident satisficers, class 2 satisficers are labeled central satisficers. For example, a class 1 satisficer who agrees with a survey item is more likely to endorse the extreme 'strongly agree' response than is a class 2 satisficer who agrees to the same item. The latter is more likely to check the midpoint unless s/he agrees really strongly. It is interesting to further elaborate this point. The reader should keep in

mind that these response frequency distributions are based on a set of 112 heterogeneous items probing a wide variety of different constructs. It is plausible that the scales from which the items were drawn have acceptable levels of discriminant validity, since all scales have been subjected to a thorough validation process (Bruner, James and Hensel 2001; Robinson, Shaver and Wrightsman 1991). Assume that the respondents' latent scores on these constructs take on independent normal distributions, such that on average the distribution of latent scores within a single individual across the items should approach a normal distribution itself (since they are similar to random draws from independent normal distributions). Hence, the observed response distribution can be seen as resulting from mapping a normal distribution onto a seven-point response format. Within a given segment of respondents, the proportions of each response style can then be considered to reflect the portion of the normal distribution (from any latent construct) that is mapped onto a given response option. For example, on average 22% of the trident satisficers (class 1) selected response option 1, 'fully disagree'. This indicates that on average, all '1' responses for this segment reflect a position somewhere in the portion of a latent construct's distribution between minus infinity and the z-value of -.783 (corresponding to the z-value left of which lies 22% of the normal probability density function). Given the near-symmetry of the expected distribution in all segments, there is no reason to expect substantial directional bias. This is important, in that no artificial mean differences are to be expected across segments for scales that have a mean near the midpoint of the scale. However, trident satisficers (class 1) will show inflated (deflated) scores on scales that have a mean higher (lower) than the response scale's midpoint, while the reverse is true for the central satisficers of class 2 (Baumgartner and Steenkamp 2001).

In addition, the satisficing strategies may be a source of heterogeneity in the way constructs are translated into item responses. To clarify this point, a set of random variables was generated for a sample consisting of similar proportions of the six segments identified as in the data used above, and the same demographic profile per segment as in the data set used above (N=10000; SPSS 12.0.2). The simulation is illustrative and does not aim to investigate this matter conclusively. First, two standard normal variables were generated, representing a latent construct and a unique variance that together constitute an observed indicator score. The weighted sum of both (such that each explains half of the variance in the resulting indicator) was mapped onto a seven point scale by applying the logic explained above. That is, for each of the six segments (trident/class 1 satisficers, etc.), the appropriate thresholds were defined to map the normal distribution onto seven response options. Regressing the indicator on the latent construct per segment then resulted in the expected item response function, an example of which is given in Figure 8-7a.

Figure 8-7a

Expected indicator score as a function of latent construct

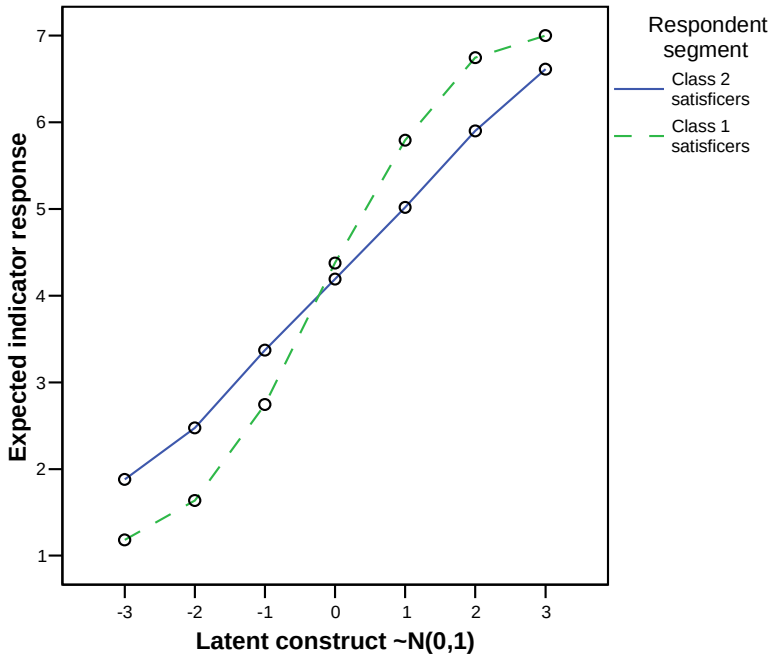
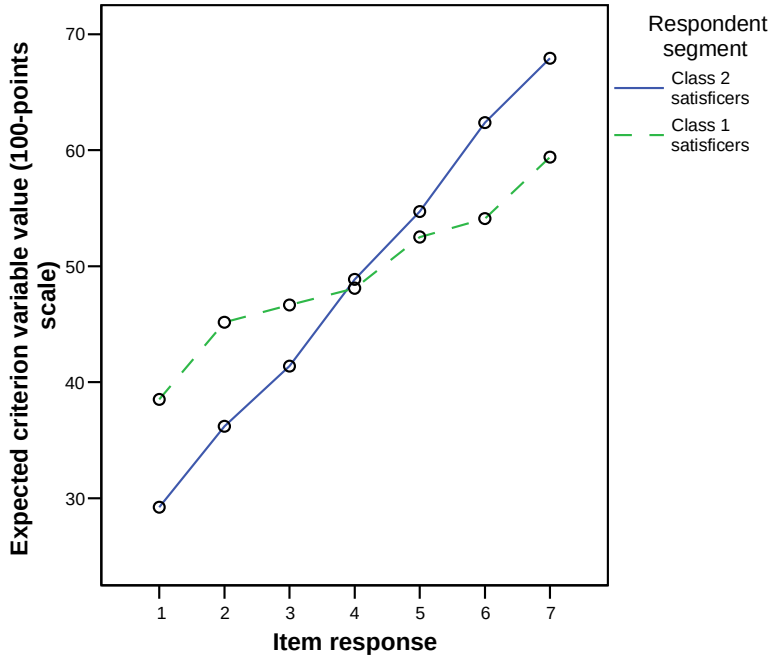


Figure 8-7b

Expected criterion variable score as a function of indicator



Note that only the two satisficing segments are shown, but that all other segments would just gradually fill up the range in between the lines (ordered by level of ERS). This graph clearly illustrates the nonlinearity resulting from disproportionate extreme responding. Also, it reflects the observation that the class 2 satisficers are very unlikely to endorse an extreme response, even if the underlying latent score is extreme (e.g., 3 standard deviations away from the mean). Most importantly, within the framework of the assumptions outlined above, the graph shows how the same response option has a different meaning for different respondents. For example, a '6' response for a trident (class 1) satisficer may correspond to the same level of the latent construct as does a '5' response for a central (class 2) satisficer. Clearly, Likert item responses should not be interpreted at face value. Also, creating ordinal categories of respondents (for example segments that have negative – neutral – positive attitudes) based on self-reports is dangerous in this regard, in that central satisficers (class 2), who are younger on average, will self-evidently be over sampled in the middle category.

The response function discussion might give the impression that ERS results in responses that carry more information, in that the full range of response options are used in responding to items. However, consider the following fictitious situation, similar to the setting investigated by Mittal and Kamakura (2001). Consumers' loyalty intentions are measured on a single item seven-point measure. Actual behavioral loyalty is independently measured on a 100-point scale (e.g. representing the percentage of purchases of a specific brand as a proportion of the total product category purchases). As described above, the intention measure would reflect a mapping of a latent construct on a response scale according to the respondent segment specific mapping function. On the other hand, the behavioral measure can be

reasonably expected to be a function of the latent construct as well (i.e. intention leads to behavior), but this function will be independent of the segment specific mapping function, since the behavior was not self-reported. Note that the latter function need not be identical across segments for the current argument to apply, but it is implausible that it has an identical or even related structure to the item response mapping function. To illustrate what happens in this setting, again a standard normal variable was generated. Based on this simulated latent variable, an item score was constructed according to the segment specific item response function, reflecting a self-report measure of intention. Also, a criterion variable score was constructed that did not follow a segment specific mapping function, reflecting a behavioral loyalty score. In both cases the latent variable explained half of the variance in the dependent variable (indicator or criterion variable); a residual normal variable explained the other half. Figure 8-7b shows the relation between the self-report measure and the criterion variable. Obviously, a small change in scores for the central (class 2) satisficers carries more information than it does for the trident (class 1) satisficers.

DISCUSSION

Acquiescence Response Style (ARS) was found to be positively related to optimizing for a majority of respondents. For a second class of respondents, there was no significant relation between ARS and OPTIM. Age was negatively related to the probability of belonging to the former class, indicating that for younger respondents it is more likely that ARS is part of an optimizing strategy.

Like ARS, Disacquiescence Response Style (DRS) made part of an optimizing strategy for most respondents. For the remainder group of respondents, who had higher education levels on average, there was no significant relation between DRS and OPTIM. Possibly, for respondents with higher education, the effort needed to

disconfirm statements in a questionnaire is less than for respondents with a lower education.

Extreme Response Style (ERS) showed a remarkable dichotomy in its relation to OPTIM, in that for a first class of respondents ERS was a satisficing strategy, while for a second class of respondents it was part of an optimizing strategy. Age related to higher probabilities of belonging to the former class. In other words, older respondents are more likely to satisfice by stylistic extreme responding than are younger respondents.

The current results concerning ARS, DRS and ERS relate directly to the conclusion by Greenleaf (1992a) that NARS (i.e. ARS-DRS) consists mainly of an information component, while ERS has both an information and a bias component[19]. It seems that rather than ERS having two components, it has different meanings for two classes of respondents, bias and satisficing related for one class, content and optimizing related for the other. Further, the positive effect of age on ERS may be conditional on respondents' satisficing, that is, this effect may only be present if respondents do not exert the effort required of them to provide optimal responses to the questionnaire items.

The analysis of Midpoint Response Style (MRS) also resulted in two classes of respondents. A first class could be characterized as younger, more probably male and having higher education levels. This class had higher levels of optimizing and showed a slight positive association between MRS and OPTIM. The other class demonstrated

[19] Actually, Greenleaf (1992a) studied standard deviation, not ERS. However, Baumgartner and Steenkamp (2001) find a strong correlation between these two and use them as indicators of the same style.

a clear-cut negative relation between MRS and OPTIM, indicating that for these respondents MRS was a satisficing strategy.

While the investigation of the response styles considered in isolation provided interesting insights into their meanings, the current study went further, and classified respondents based on their full response style profile. Such profile related the four styles (ARS, DRS, ERS and MRS, ADEM in short) to optimizing. Two major classes were found. A first class consisted of 49.2% of the current sample and showed a strong negative relation between ERS and OPTIM. That this effect generalized to ARS and DRS was due mainly to extreme responses (which necessarily are either positive or negative). Though less outspoken than for ERS, MRS also showed a significant negative relation with OPTIM. Older respondents had a higher chance of belonging to this class.

A second class, consisting of 50.8% of the sample, showed a more single-minded focus on MRS as a satisficing strategy.

An interesting observation regarding these two classes is that their response style levels are nearly identical if they are optimizing. Only when respondents are satisficing, the differences in response strategy became clear, as illustrated in Figure 8-6. When satisficing, class 1 respondents showed a response frequency distribution with three peaks, reflecting high ERS and MRS. This pattern was labeled trident satisficing. Class 2 satisficers were labeled central satisficers because their response frequency distribution became narrowly concentrated around the midpoint.

The presence of two such disparate satisficing strategies has important implications. Most importantly, current findings suggest that the same response may have very different meanings across respondents. As illustrated in Figure 8-7a (based on a simulation), a latent score one standard deviation above the mean could be mapped as

a 6 on a seven-point scale by a trident satisficer, while being mapped as a 5 by a central satisficer. Similarly, a self-reported attitude or intention may translate in dramatically different levels of related behavior for different segments of respondents (cf. Figure 8-7b). Consequently, for a researcher trying to predict behavior, it would be highly relevant to know the ADEM response style profile of the respondents under investigation. This would enable one to predict the functional form of the relation between criterion variable and self-report measure a priori, rather than having to derive it post hoc from the actually observed self-reports and behavioral data (as was done by Mittal and Kamakura 2001 for several demographic segments).

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

Since the current study was carried out in a specific sample in one European country, it would be very enlightening to replicate the investigation in different samples in a cross-cultural context. It is a plausible hypothesis that cross-cultural differences in response styles are moderated by optimizing.

Another limitation to the generalizability of the current findings is that data were collected using one specific item format. Seven point Likert items have been recommended by experts from diverse research streams (Cox 1980), but may be contaminated by response styles in particular ways. Especially the trident satisficing pattern is most probably very specific to this format (although it could operate in the equally popular five point likert items). It might well be that the gain in information transfer capacity of this type of scale is offset to a large extent by the heterogeneity in the way respondents use this scale. Moreover, the heterogeneity is hard to detect, requiring both specific item sets and rather complicated analyses. This issue definitely deserves further investigation.

In the current study, a newly proposed operationalization for OPTIM was used. It could be argued that the Time-On-Task aspect of the proposed measure is confounded with attitude accessibility, in that fast responses have been linked to accessible attitudes (Krosnick 1993). However, this possibility was countered by measuring both differentiation and time-on-task over a random sample of items that were highly diverse in terms of content. It is very unlikely for a respondent to have similar accessibility levels for all the different topics in the questionnaire.

APPENDIX 8-1:

FURTHER OPERATIONAL DETAILS OF THE DIFF AND OPTIM MEASURES

It is not the case that DIFF, as one of the components of OPTIM, would by design be related to any of the response styles under study. That is, the mere operationalization of DIFF and ARS, DRS, ERS and MRS will not artificially lead to correlations between DIFF and any of the response styles. This was verified by means of a simulation of 250 respondents responding to 112 seven-point items following a uniform response distribution (N replications = 100). The resulting data matrix corresponds to 100 times 250 by 112 random draws from a uniform distribution. For these simulated respondents DIFF, ARS, DARS, ERS and MRS scores were computed. The correlation between DIFF and each of the response styles was zero. This means that in the absence of systematic response tendencies and shared content, DIFF does not correlate with any of the four response styles. Note that the level of DIFF does impose a limit on the values that the response styles can take on. For example, if MRS is 1 (this means responding all items with a midpoint response) DIFF can only be zero, and if DIFF (this means using each response option with an equal frequency) is 1, MRS can only be .143. However, only trivial numbers of cases were situated near the boundaries of the bivariate space of DIFF and any response style. This was evaluated by computing the boundaries, i.e. the minimally and maximally possible values for DIFF given a level of a response style. Note that the relationship is most determining for MRS and ERS, since these response styles reflect proportions of certain options. The highest possible value of DIFF giving a specific MRS level, e.g. is given by the following formula, where n is the number of items and k is the number of response options: $([([n-MRS]/[k-1]+1)^{k-1} * (MRS+1)] – (n+1)) / (n/k + 1)^{k}$. In the actual data analysis, it was checked whether results were robust

against inclusion/omission of outliers. This proved to be the case. Hence, there are no relationships in the data that are merely due to the specific operationalization of DIFF and the response styles. Moreover, DIFF is only a component of OPT, since it is combined with TOT. The latter variable is measured independently of observed response frequencies.

OPT is not scaled in a way that is readily interpretable. This is not a problem: many scales in psychology are arbitrary and this needs in no way affect their reliability and/or validity (Blanton and Jaccard 2006). Objective metrics can be arbitrary if used as indicators of a latent construct rather than the objective physical reality they directly refer to. As argued by Rindskopf (2003, p. 368), "[s]ome researchers may object to transformations, as the interpretation in the transformed scale may not seem as natural. This may be so with many physical measurements, but in most social science research there is nothing sacred about the original measures, so no harm is done by transforming." For example, Implicit Association Tests yield an estimate of reaction time in milliseconds, but this does not mean such measurement results in attitude/association measurement with a rational zero point (Blanton and Jaccard 2006). Similarly, the indicators based on the product of DIFF and TOT in the current study refer to the single latent construct OPTIM, rather than to the interaction of two different constructs (DIFF and TOT). For this reason, it is important to use the square root of the product terms directly to measure the construct and to evaluate its internal consistency. Computing an interaction term based on two constructs DIFF and TOT would not be appropriate for the current purposes.

VALIDATION OF OPTIM

Some evidence in support of the OPTIM measure is provided. To this end, a MIMIC model was specified in which OPTIM, measured by its three indicators was regressed

on age, sex and education level. Using the ML estimator in Mplus 4.1, the model fitted the data quite well (χ^2 (6) =14.74, p=.0224; CFI = .987; TLI = .973; RMSEA=.050, RMSEA 90 Pct C.I. = 0.017 – 0.082; SRMR = .019). Standardized loadings were .85, .82 and .76. The demographics explained a small amount of the variance in OPTIM (R^2=.02) and only the effect of education on OPTIM was significantly positive, with a logistic regression weight B=.012 (s.e. =.005) and t-value = 2.351. In the panel used for this research, education has been found to be positively related to respondent motivation, expressed in the higher probability of participation (see Study 3 / Chapter 6). Hence, the positive relation of OPTIM to education lends support to the nomological validity of OPTIM.

CHAPTER 9: CONCLUSION

CHAPTER OUTLINE

In this concluding chapter, the previous chapters are recapitulated. Based on this overview, the theoretical and practical implications are discussed, focusing on three related issues: the impact of response styles, the meaning of response styles and remedies against response style bias.

RECAPITULATION

Questionnaires using closed-ended questions are indispensable for consumer research. Likert items are a commonly used type of such questions. Unfortunately, previous research has demonstrated that these measures may be biased due to response styles. In the conceptual section of the current dissertation, response styles were conceptualized as respondent-specific ways of mapping judgments onto response categories. Individuals may exhibit stylistic preferences for agree responses (Acquiescence Response Style or ARS), disagree responses (Disacquiescence Response Style or DRS), extreme responses (Extreme Response Style or ERS) and/or midcategory responses (Midpoint Response Style or MRS). It was illustrated in both the Confirmatory Factor Analysis and Item Response Theory frameworks how theses preferences may affect construct-indicator relations. Also, an overview was provided of how response styles may affect observed univariate response frequency distributions and multivariate relations.

Based on a review of the literature, a typology was proposed of how response styles may be measured, with a focus on two dimensions. As for the first dimension, the items or item sets used as the indicators of response styles can either serve only the specific purpose of measuring response styles, or they can be used simultaneously as indicators of content and as indicators of style. As for the second dimension, the influence of content on the item responses can be corrected for in different ways (since otherwise content provides an alternative explanation for response tendencies). First, when convenience samples of items are used, there is insufficient control. Second, items can be created that are free of content. Third, content can be manipulated in a controlled way similar to an experimental design. Finally, content can be reduced to random noise by using sets of items that are heterogeneous in

content. Based on a consideration of the advantages and limitations of the different possibilities, in the current dissertation, specific style indicators were used (corresponding to the second level of the first dimension) and the influence of content was corrected for by randomizing content over items (fourth level of the second dimension).

Empirical study 1 (Chapter 4) used data from over 3000 online respondents to a specifically designed set of Likert items to study the effects of item location and content. More specifically, the correlations between items were modeled as a function of the relation between the two items in terms of content and distance. Content refers to the items either measuring the same construct in the same direction, measuring the same construct in the opposite direction (reversed items) or measuring unrelated constructs. Distance refers to the number of other items that stand in between the two focal items. It was found that after controlling for content, items on average showed a positive correlation, which decreased slightly with an increase in inter-item distance. This phenomenon was attributed to the operation of ARS. An additional distance effect was found for content related items. For items that measured the same construct in the same direction, the strength of the correlation decreased as a function of item distance. For reversed items, the strength of the correlation increased (i.e. became more negative) as a function of item distance. This was interpreted as supporting a Unipolar Response model, according to which respondents interpret reversals as being more independent (i.e. measuring unrelated constructs) the closer they are to one another. An important implication of the findings in Study 1 is that the bias in reversed item responses cannot be equated to the operation of ARS (which would imply independence of content by definition) but is most probably a content driven context effect.

Study 2 (Chapter 5) examined the short term stability of response styles. Based on a literature review, nine alternative models were specified of how response style indicators based on subsequent parts of the same questionnaire can be related, corresponding to the combination of two dimensions with three levels each. The first dimension refers to the specification of a common factor, which can be congeneric, tau-equivalent or absent. The second dimension refers to the specification of an autoregressive effect, which can be time invariant, time variant or absent. These nine models were fitted to secondary data (Hui and Triandis 1985) and primary data. From the analyses, the presence of a common factor emerged consistently across data sets and response styles. The choice between a congeneric and a tau-equivalent common factor was less consistent, as was the strength of the autoregressive effect. For most data and styles, the latter was negligable however. It was concluded that response styles have a major stable component in the short term.

Study 3 (Chapter 6) extended the stability question to the long term. It was found that response styles are remarkably stable over two different questionnaires that were filled out by the same respondents with a one year time gap in between. Demographics explained only a small part of the variance in the stable component of the response styles (ranging from 2.3% in DRS through 9.5% in ERS).

Study 4 (Chapter 7) consisted of a comparison of three modes of data collection in terms of response style levels: paper and pencil, telephone and online. The major finding was that the telephone data showed a lower level of MRS and a higher level of ARS. It was shown how these differences led to predictable biases in a cross-mode Means And Covariance Structure analysis of a substantive construct (Trust in Frontline employees) and how measurement invariance tests might not be useful in addressing such cross-mode differences in response styles.

Study 5 (Chapter 8) investigated response styles as cognitive methods applied by respondents to reduce the burden imposed on them in the survey situation. Optimizing was defined as time-intensive differentiation of responses to items that are homogeneous in form but heterogeneous in content; the polar opposite of this is called satisficing. The relation of response styles to the optimizing-satisficing variable was studied by means of structural equation mixture modeling. Respondents could be classified in two major segments. One segment of respondents seemed to satisfice by increasing their levels of ERS and MRS. This suggested that these respondents simplify their task of selecting a response out of multiple response options (seven in particular) to a yes – neutral – no response. A second segment showed a positive relation between satisficing and MRS only. This presumably indicated that these respondents no longer chose sides once they minimized the amount of effort they invested in the respondent task. It was illustrated how these two segments cause heterogeneity in the meaning and predictive/convergent validity of observed item responses.

IMPLICATIONS

The theoretical and empirical developments in the previous chapters have provided insights on three related key issues for applied research and research concerning the four response styles under study (Acquiescence Response Style or ARS; Disacquiescence Response Style or DRS; Extreme Response Style or ERS; and Midpoint Response Style or MRS). First, further insights have been gathered concerning the potential impact of response styles on questionnaire data. Second, the meaning and conceptual status of response styles have been further crystallized. Third, tools have been provided to better avoid response style bias or cure it where necessary. These points are elaborated below. As is apparent from these topics, and

since in the response style literature practical measurement issues and theoretical meaning have been closely related (e.g., Rorer 1965; Welkenhuysen-Gybels, Billiet and Cambré 2003), the applied and theoretical implications are discussed together below.

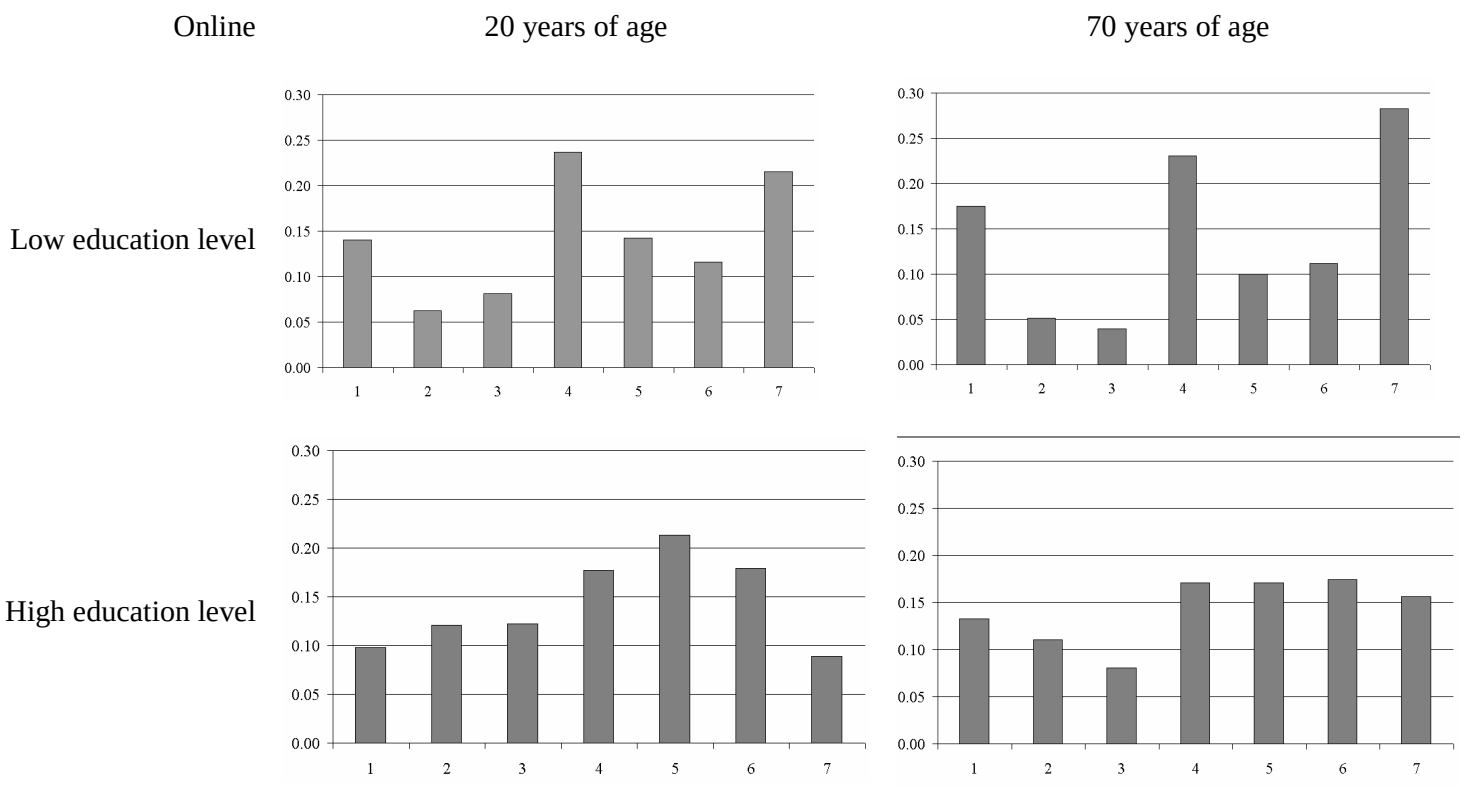
IMPACT OF RESPONSE STYLES

The same response category can have different meanings for different respondents. That is the essence of the response style problem as it has been conceptualized in the current dissertation. Response styles may be the cause that a given level of a latent construct of interest may lead to different levels of observed indicators. If such heterogeneity in the mapping function between construct and indicator were purely random (within and between subjects), the problem would merely increase the proportion of noise in questionnaire data. However, there are clear indications that there is a systematic component to the bias. After controlling for content, different demographic groups have different expected response frequency distributions. This is illustrated in Figure 9-1, which presents the expected frequency distributions for four demographic groups that differ to a substantial extent in terms of age (20 years versus 70 years of age) and education level (low education, corresponding to primary school only, versus high education, corresponding to 5 years of formal education after secondary school). The estimates were obtained from the online sample data in study 4 (Chapter 7) by regressing the percentage of category responses (e.g. the percentage of times a respondent chooses option 1 across the heterogeneous set of items) on demographics. The regression predicted values were then used to create the graphs. The average expected item score (and standard deviation) for the respective groups were 4.41 (1.95) for the young lowly educated; 4.19 (1.79) for the young highly educated; 4.52 (2.13) for the old lowly educated; 4.30 (1.97) for the old highly

educated. Figure 9-1 illustrates that younger, highly educated respondents have the lowest levels of ERS, as opposed to the older respondents with lower education levels, who have a strong preference for both the midpoint (high MRS) and the extreme response options (high ERS), while largely neglecting the options in between. Apart from the dramatic difference in observed distributions, the graphs clearly show that observed scores may often be normally distributed only among very specific groups of respondents, in particular young and highly educated people (i.e. the group including students). Clearly, measures that were validated in such samples might lead to surprises when used in other populations.

Figure 9-1:

Expected frequency distributions by age and education level



On the stimulus-side, study 4 (chapter 7) already showed how substantial differences arose between different modes of data collection. The telephone mode showed lower levels of MRS combined with higher levels of ARS. This led to predictable biases in the responses related to an unrelated measure. The bias would most probably have been confused with content if response styles had not been assessed.

In addition to the systematic differences in response styles between different demographic groups and modes of data collection, there is much heterogeneity between respondents that remains unexplained by demographics and modes of data collection. This relates to the next issue: the meaning of response styles.

MEANING OF RESPONSE STYLES

As discussed in Study 2 (Chapter 5) and Study 3 (Chapter 6), much of the debate in the response style literature has focused on the generality and stability of response styles. The high internal consistency of response styles across unrelated samples of items in the studies reported above provided convincing evidence of the generality of response styles. Further, in the short run response styles were found to be very stable (Study 2)[20].

Very remarkably, stability also held over a much longer period, a one year time lag in particular (Study 3). Demographics, though significant as antecedents of the styles, explained only a minor portion of the total variance in the stable component of the response styles. Future research might want to revive the study of personality

[20] For specific styles and contexts, there may be an autoregressive component to the style as well. The current data nor the literature have provided conclusive evidence on the meaning of this autoregressive component; this is discussed in the limitations and future research section. Nevertheless, it should be stressed that the major component of all four response styles was found to be stable over a single questionnaire.

correlates of response styles, this time taking care to avoid the shortfalls of the past. While long term stability is the major finding in Study 3, a substantial time specific component was observed too. For this aspect, it would be very interesting to investigate the impact of situational factors such as mood, time pressure and cognitive burden (such studies are being planned).

The complex interplay between response styles became more understandable by thinking of the response styles as components of satisficing strategies. Interestingly, two major segments of respondents were identified (study 5, chapter 8) based on two satisficing strategies. For one group, on average younger respondents, higher levels of satisficing led to an increasing concentration of responses at the midpoint of the scale. For a second group, to which more older respondents belonged, satisficing was related to a so-called trident response pattern, with a concentration at the midpoint and the extremes. The findings suggested that respondents may show similar response patterns when optimizing, but diverge dramatically once they decide to save time and cognitive effort while still responding to questions. In the latter study, but also in the other studies, ERS and MRS stand out as the apparently most consistent and influential styles. ARS and especially DRS may be less consistent and possibly less problematic. Future research might merit from shifting the focus accordingly. While ARS has proven to be most easy to grasp, to measure and to correct for, it has also been the easiest subject of harsh criticisms of the response style literature (Rorer 1965; Block 1971). ERS has rightfully received quite some recent attention (e.g., Arce-Ferrer 2006), but might deserve even more. MRS may have been underestimated as a source of systematic bias (but see Baumgartner and Steenkamp 2001).

THE CURE

In the literature, two major stages can be discerned where bias due to response styles may be tackled during a research study: the implementation occurs before data collection (design remedies) or after data collection (measurement/statistical remedies) (Baumgartner and Steenkamp 2001; Podsakoff et al. 2003).

DESIGN SOLUTIONS (EX ANTE)

The formulation and selection of items is important in preventing response style bias. One of the most hotly debated design options to counter response styles has been the use of reversals.

Reversals (as a thought experiment)

While study 1 has further established the problematic nature of using reversed items, their use is valid under specific conditions. In particular, if the reversals are located sufficiently far apart from their non-reversed counterparts, it seems that respondents do not consider them as relating to independent dimensions. Thus, it is important to consciously position reversals apart from non-reversals. Of course, this only confirms that it may be dangerous to expect respondents to interpret such items in the way that the researcher intends them to. Also, this indicates that factor structures should be considered in light of how the questionnaire was organized. It is easy to obtain neat and clear structures by providing respondents with blocked items, maybe even under a header that explicates what latent construct is being measured. However, in such situation it is not valid to apply classical test theory and the domain sampling model, since these models consider items in isolation and assume that their meaning and item response functions are independent of the context they are in. If a consistent context is deliberately created, criteria like Cronbach's alpha loose part of their meaning and

only confirm that a succesful manipulation has been applied in creating a homogeneous context for the items. When used in a different context, the reliability parameters would probably not apply. Also, in the case of reversed items, the manipulation might backfire, creating what was labeled self-generated non-validity in study 2 (Chapter 5).

Needless to say, the current research does not conclusively settle the issue of whether or not to use reversed items. In this context, it has been pointed out that some items seem to be irreversible (Ray 1979). A point in case is the Authoritarianism scale discussed by Peabody (1966). A crucial implication of such presumed irreversibility may deserve some further attention: if it takes measurement experts several decades to formulate reversals, it is questionable that respondents can meaningfully think of the connotation of a disagree response to such items in the span of a few minutes (or even seconds) while responding to the items in question. In these instances, it is not surprising that many respondents agree or show inconsistent double agreements or disagreements.

Therefore, it is not necessarily recommended that all scales should include reversed items. However, it may be recommendable for researchers to formulate a reversal for each item they include in a questionaire that uses an agree-disagree rating format. If this turns out to be impossible, it may imply that formulating a meaningful (disagreeing) response is impossible as well. In other words, coming up with reversals may be a useful thought experiment and criterion for evaluating items in a questionnaire. Items that do not pass the reversal test, should be rephrased or deleted. Test-reversals can be evaluated by measurement experts or a convenience sample of respondents in a pilot phase of testing. For example, it may be hard to think of a valid reversal to “No weakness or difficulty can hold us back if we have enough will

power" (cf. Peabody 1966). Consequently, the meaning of a disagree response to this item is not clear, and it makes little sense to ask respondents to indicate their level of agreement to such item. In a way, if a respondent disagrees with the statement, s/he is contradicting the meaning of the word 'enough'. This becomes obvious in a statement like "if we have enough to drink, we will not be thirsty" which is true by definition (almost by definition, strictly speaking).

All this is not to say that the problem of response styles resides in the item rather than the respondent. An interactional account seems in place, where response style bias is due to the combination of respondent and item (Paulhus 1991). Even authors that place most stress on the item effect explicitly or implicitly acknowledge this. For example, Peabody (1966), after arguing why specific items will lead to ARS, introduces cognitive sophistication as a moderator of this effect. Similarly, McClendon (1991b) simultaneously tests for item and respondent effects, and it seems that the interaction is by now accepted as a starting point for research (Baumgartner and Steenkamp 2001).

MEASUREMENT AND STATISTICAL CONTROLS (POST HOC)

The items commonly used in consumer research are not immune to response style bias, as evidenced by the results obtained in the above studies using representative samples of these items. Clearly, if prevention fails, a post hoc approach is called for. Such approach has two components: diagnosis and correction. Diagnosis of response styles refers to measuring levels of response styles. The results can be used to assess whether there is a problem, and – if so – to select specific respondents for analyses. Correction can be pursued by statistically controlling for response styles in analyses.

Diagnosis

The studies presented in the current volume offer some important guidelines for response style measurement. Where possible, it is recommended to include heterogeneous, representative samples of items in questionnaires. Since response styles are mainly stable over the time span of a single questionnaire, the specific location of these marker item sets in the questionnaire is largely inconsequential. However, to account for the (small) local component, spreading several sets of items throughout the questionnaire might be optimal. The current studies used at least three sets each consisting of 18 through 48 items as indicators of response styles. In applied settings, less marker items will probably be used, though a minimum of 20 items seems a reasonable requirement.

While the long term stability of response styles is a worrysome phenomenon, in that it suggests the presence of a consistent source of bias, it also opens doors in terms of measurement. In particular, response style variables could be created and used as stable background covariates in panel research. Still, the optimal approach seems to include response style indicators in every data collection. If this is too costly, a single (one time) measurement presents itself as the next-to-best solution. If not even this is possible, researchers should at least be aware of the demographic correlates of response styles. Hypotheses should be formulated and tested on how the measures can be expected to behave as a function of demographics mediated by response styles. For example, when comparing age cohorts, one should inspect response distributions for extreme responses in the different age groups, and take into account that this might be due to response styles.

To test for robustness against response style effects, analyses might be executed with and without respondents who show a three-peaked (high ERS and high MRS) or a one-peaked (high MRS) response pattern.

Correction

Ideally, analyses should statistically control for the effect of response styles. Two methods to do so are discussed briefly: response styles as covariates, and response styles as individually estimated mapping functions.

Response styles as covariates

Measures of response styles can be used as covariates in analyses. Ways to do so have among others been shown in the context of multiple regression analyses (Greenleaf 1992a; Baumgartner and Steenkamp 2001) and for the case where balanced scales are controlled for ARS using Structural Equation Modeling (SEM) (Billiet and McClendon 2000). In the current dissertation the following methodological requirements were proposed and applied. First, the use of response style specific marker items allows measuring several response styles and not only ARS-DRS. Splitting the total set of marker items into several subsets makes it possible to include the measures as a latent construct in a SEModel, with the related advantages it brings (Podsakoff et al. 2003). In multi-group settings especially, the relevance of this approach was demonstrated (study 4).

Towards individual measurement model parameters

To conclude, a potential route for future research is suggested. Response styles were conceptualized as individual difference variables that relate to the mapping function of constructs to measures. As shown in chapter 2, Item Response Theory models use threshold parameters that may closely correspond to ARS, DRS, ERS and MRS. For example, high ERS levels may indicate that the thresholds for the extreme options are

closer to the latent mean, resulting in a higher probability of checking the extreme options; ARS may be indicative of a general shift of the thresholds to the lower side of the latent score continuum, etc. Individual differences in mapping functions might be optimally accounted for if these threshold parameters could be modeled directly as latent variables with their own specific indicators.

The actual technical specification of such model is beyond the scope of the current dissertation. However, the concept might become feasible by extending algorithms like those used for computing polychoric correlations. Specifically, the likelihood of the joint multivariate distributions of the observed substantive variables could possibly also take into account the expected marginal frequency distributions based on the response style indicators. If considered at the group level (in multi-group analysis), this approach could anchor the measurement parameters, thus avoiding their indeterminateness. If considered at the individual level, respondent specific construct-indicator functions could be combined with sample level estimates of the structural relations of interest. Obviously, this approach might be complex. However, it would take into account the fact that the response categories have different meanings for different respondents.

Even so, regardless of the specific technical approach taken, the validity of questionnaire measurement would gain much from a more systematic integration of response style measures in data-analyses. It is hoped that the current dissertation has helped in enabling such approach.

LIMITATIONS AND FUTURE RESEARCH

In addition to the topics mentioned in the discussion above, some further limitations and opportunities for future research are worth noting.

GROWTH CURVES OF RESPONSE STYLES

As mentioned above, an autoregressive effect may be present in the response style effects (especially ERS and MRS) on subsequent sets of items. A fruitful avenue of future research might link this finding to the status of MRS as a satisficing strategy for nearly all respondents, and ERS as a satisficing strategy for some, an optimizing strategy for others. Possibly, respondent fatigue leads to specific curves of ERS and MRS over a questionnaire. These trajectories may be individual-specific. Studies using growth mixture modeling might provide interesting insights in this regard (such study is planned).

SCALE FORMAT

In the context of stimulus-side antecedents of response styles, an important limitation of the current set of studies is worth noting. In particular, systematic use was made of seven point scales. This choice was based on recommendations in the literature, after a review of which Cox (1980) noted: “If the number of response alternatives were to be established democratically, seven would probably be selected.” Nevertheless, the common advice has been to use 7 plus or minus 2 categories (referring to psychophysic research; Miller 1956) and five point scales are rather popular among survey researchers in marketing and management. It would be highly relevant to investigate to what extent the current findings (e.g. the trident satisficing strategy) generalize to other response scale formats. Also, higher numbers of response options may have been found to lead to higher reliability precisely because of the operation of response styles, a possibility already touched upon by Cronbach (1950). Also on this issue a study is planned.

The format issue is especially relevant in light of the close relation between response styles and some method effects. For example, the labeling of the response options

may lead to different levels of observed response styles by increasing the likelihood of selecting specific options. For example, a preference for the response option 'seven' might vary as a function of it being labeled or not, or it might be due to recency effects. Such effects could be seen as alternative explanations for some of the results reported in this dissertation. However, in the theoretical framework proposed here, it is more meaningful to think of these effects as antecedents and/or moderators of response styles, in that they alter the way respondents map underlying judgments to response scales (without being descriptive of the mapping function and its outcome itself).

CROSS-CULTURAL EXTENSIONS

The studies reported in the current dissertation are based on samples from a single country, Belgium. Research has shown that cross-cultural differences in response styles exist (Johnson et al. 2005), though the effect might be relatively small, especially across European countries (Baumgartner and Steenkamp 2001). It would be very interesting to extend the findings from the studies reported here to a cross-cultural context. First, replications are called for to establish the generality of the findings. Even more interesting would be the use of national culture as a moderator of the antecedents of response styles such as the mode of data collection. Similarly, it is possible that cultural dimensions or other cross-country differences affect the extent of satisficing and/or moderate the relation between satisficing and response styles.

APPENDIX A: DEFINITIONS AND ABBREVIATIONS

A-1 LIST OF ABBREVIATIONS

ADEM: ARS, DRS, ERS, MRS (see below)

AR: Autoregressive (see study 2)

ARS: Acquiescence Response Style

B : regression coefficient

BR: Bipolar Response (see study 1)

CFA: Confirmatory Factor Analysis

CFI: Comparative Fit Index

Diag: diagonal

$DIST\_REVERSE_{ij}$: Interaction of LN_DIST and REVERSE_ξ (study 1)

$DIST\_SAME_{ij}$: Interaction of LN_DIST and SAME_ξ (study 1)

df: degrees of freedom

DRS: Disacquiescence Response Style

ERS: Extreme Response Style

Est. : estimate

IRT: Item Response Theory (see conceptual background)

IV: Independent variable

λ (lambda) : factor loading

LN: natural logarithm (i.e. logarithm base e)

LN_DIST_{ij} : Natural logarithm of the distance between item i and item i' (study 1)

MACS: Means And Covariance Structure (see study 4)

MRS: Midpoint Response Style

N : sample size

NARS: Net Acquiescence Response Style

OPTIM : Optimizing (study 5)

p : probability

r : Pearson correlation coefficient

s : standard deviation

Stdd. : standardized

Unstdd.: unstandardized

var: variance

REVERSE_ξ_{ij} : dummy indicating that item i and j both measure construct ξ in the opposite direction (study 1)

RMSEA: Root Mean Square Error of Approximation

RS: Response style

s.e.: standard error

SAME_ξ_{ij} : dummy indicating that item i and j both measure construct ξ in the same direction (study 1)

S-B factor: Satorra-Bentler correction factor

SEM: Structural Equation Model(ing)

SEMM: Structural Equation Mixture Model(ing) (see study 5)

Sig.: significance level

t : t-value

TLI: Tucker-Lewis Index

UR: Unipolar Response (see study 1)

APPENDIX A-2: DEFINITIONS OF SOME KEY CONCEPTS

COMPONENTS OF A QUESTIONNAIRE

The current dissertation focuses on the items that are part of a questionnaire. An item refers to a closed question that can be answered by indicating one's position on a rating scale (i.e. an ordered set of numbers called response options or response categories). In this report, the word 'item' refers to both the question and the response options or categories. A multi-item measurement scale consists of several such items that are similar in content and are designed to measure the same construct (note the difference with 'rating scale'). Items can be reverse coded, which means they are formulated in such a way that their rating is negatively correlated to the score of the construct they measure. If half of a scales' items are reverse coded, the measurement scale is considered to be balanced.

LIKERT ITEMS

The focus of the studies reported in this dissertation is on response styles in Likert items. An example of two Likert items using a seven-point rating scale is given in Figure 1. Note that any number of points could be used, though the most common numbers are 5, 6, 7, 9 or 10. When initially proposing this format, Likert (1932) used five points, but later the number of seven has been recommended by experts (Cox 1980).

Figure A-1

Example of Likert items

	Strongly disagree	disagree	Slightly disagree	Neutral	Slightly agree	Agree	Strongly agree
I am a homebody	1	2	3	4	5	6	7
I eat more than I should	1	2	3	4	5	6	7

VARIABLES AND CONSTRUCTS

The ultimate aim of questionnaires and the items contained therein is the measurement of variables, some of which may be constructs. Hox (1997) defines the terms 'variable', 'concept' and 'construct' as follows: "A variable is a term or symbol to which values are assigned based on empirical observations, according to indisputable rules.[...] A concept is an abstraction formed by generalization from similar phenomena or similar attributes. [...] A construct is a concept that is systematically defined to be used in scientific theory." To make constructs subject to empirical testing, they need to be made measurable, a process referred to as operationalization. Most constructs are not directly observable. Variables referring to such constructs are called latent variables (as opposed to observed variables).

SETS OF ITEMS USED AS RESPONSE STYLE INDICATORS

While items that measure a construct are closely related in terms of content, in the studies reporterd here, use is also made of items that are not related to one another in terms of content. A 'set of items' as such does not imply any relation apart from the decision to put these items together, for example to use them to compute a response style indicator. A response style indicator refers to a variable indicating the level of a response style for a given respondent who filled out a given questionnaire. Multiple indicators can be combined into a response style measure. The relation between the indicators and the measure is then comparable to the relation between items and a scale: several indicators are meant to be as many operationalizations of one response style and can hence be summarized in one response style measure.

APPENDIX B: ITEMS

APPENDIX B – 1: ITEM SET 1

Item set 1 consists of 28 items measuring four related constructs, 20 items forming 10 reversal pairs, and 28 randomly sampled items from Bruner, James and Hensel (2001). The complete set was used for study 1 (Chapter 4). For study 4 (Chapter 7) and wave 1 of study 3 (Chapter 6), all reversal pairs and filler items were used, plus the first item of the three scales by Steenkamp and Gielens (2003). For study 4 (Chapter 7), as a substantive measure, the four Trust items were used (Sirdeshmukh, Singh and Sabol 2002). Bipolar items were adapted to Likert format. All items were subjected to a pilot test and rephrased if unclear or when leading to several missing values.

All items were rated on numbered seven point agreement scales, where option 1 was labeled ‘fully disagree’, 4 was labeled ‘neutral’, and 7 was labeled ‘fully agree’, as shown below.

	Fully disagree			neutral			Fully agree
Statement …….	1	2	3	4	5	6	7

RELATED CONSTRUCTS USED IN STUDY 1

Market mavenism, dispositional innovativeness & Consumer susceptibility to normative influence (Steenkamp and Gielens 2003)

Dispositional innovativeness:

Als ik een nieuw product in de rekken zie, ben ik afkerig om het te proberen	When I see a new product on the shelf, I'm reluctant to give it a try
Algemeen genomen ben ik bij de eersten om nieuwe producten te kopen wanneer ze op de markt komen.	In general, I am among the first to buy new products when they appear on the market
Als ik een merk goed vind, zal ik zelden veranderen van merk gewoon om iets nieuws te proberen.	If I like a brand, I rarely switch from it just to try something new
Ik ben heel voorzichting bij het proberen van nieuwe en andere producten.	I am very cautious in trying new and different products
Ik ben meestal bij de eersten om nieuwe merken uit te proberen.	I am usually among the first to try new brands
Ik koop zelden merken waarvan ik niet zeker ben hoe ze zullen presteren.	I rarely buy brands about which I am uncertain how they will perform
Ik hou ervan een risico te nemen bij het kopen van nieuwe producten.	I enjoy taking chances in buying new products
Ik koop niet graag een nieuw product vooraleer andere mensen dat doen.	I do not like bo buy a new product before other people do

Market Mavenism:

Ik leer mijn vrienden graag nieuwe merken en producten kennen.	I like introducing new brands and products to my friends
Ik praat niet tegen mijn vrienden over de producten die ik koop.	I don't talk to friends about the products that I buy
Mijn vrienden en buren komen vaak bij mij voor advies.	My friends and neighbors often come to me for advice
Mensen vragen zelden mijn mening over nieuwe producten.	People seldom ask me for my opinion about new products

Consumer susceptibility to normative influence:

Als ik wil zijn zoals iemand anders, probeer ik dikwijls dezelfde merken te kopen als deze persoon.	If I want to be like someone, I often try to buy the same brands that they buy
Het is belangrijk dat anderen de producten en de merken die ik koop leuk vinden.	It is important that other like the products and brands I buy
Ik koop zelden zelden iets heel modieus tot ik zeker weet dat mijn vrienden het mooi vinden.	I rarely purchase the latest fashion styles until I am sure my friends approve of them
Ik identificeer me vaak met andere mensen door dezelfde producten en merken te kopen als zij.	I often identify with other people by purchasing the same products and brands they purchase
Als ik producten koop, koop ik meestal de merken waarvan ik denk dat anderen ze goed zullen vinden.	When buying products, I generally purchase those brands that I think other will approve of
Ik weet graag welke merken en producten een goede indruk maken op anderen.	I like to know what brands and products make good impressions on others
Als andere mensen me een product kunnen zien gebruiken, koop ik vaak het merk dat ze verwachten dat ik koop.	If other people can see me using a product, I often purchase the brand they expect me to buy
Ik krijg het gevoel erbij te horen als ik dezelfde producten en merken koop als anderen.	I achieve a sense of belonging by purchasing the same products and brands that others purchase

Loyalty & trust in frontline employees (adapted from Sirdeshmukh, Singh and Sabol 2002)

Ik vind het personeel in deze winkel heel betrouwbaar	I feel that the employees of this store are very dependable
Ik vind het personeel in deze winkel heel competent	I feel that the employees of this store are very competent
Ik vind het personeel in deze winkel heel eerlijk	I feel that the employees of this store are of very high integrity
Ik vind het personeel in deze winkel heel responsief tegenover klanten	I feel that the employees of this store are very responsive to customers
De kans is groot dat ik in de toekomst nog in deze winkel kom	It is very likely that I will visit this store again
De kans is groot dat ik deze winkel aanraad aan mijn vrienden, buren en familie	It is very likely that I will recommend this store to my friends, neighbours and family
De kans is groot dat ik naar deze winkel kom de volgende keer dat ik iets van kleren nodig heb	It is very likely I will come to this store the next time I need any clothes
De kans is groot dat ik meer dan 50% van mijn budget voor kledij in deze winkel zal besteden	It is very likely I will spend more than 50% of my budget for clothing in this store

Reversal pairs

Arme mensen verdienen onze sympathie en steun	Poor people deserve our sympathy and support
Ik vind het tijdsverspilling om mee te voelen met arme mensen	I find it a waste of time to sympathize with poor people
Ik sta vaak in het middelpunt van de belangstelling	I am often in the center of attention
In een groep mensen ben ik zelden het middelpunt van de belangstelling	In a group of people I am seldom the center of attention
Ik bezit niet de juiste vaardigheden om een goede onderhandelaar te kunnen zijn	I don't possess the right set of skills to make a good negatiator
Ik ben goed in onderhandelen	I am good at negotiating
Het werk dat ik verricht is nutteloos	The work I do is useless
Het werk dat ik doe is waardevol	The work I do is valuable
Mijn familie is egoïstisch	My family is egotistical
Mijn familie is erg sociaal	My family is very social
Ik vind dat de meeste producten te duur verkocht worden	I think most products are being sold too expensively
In het algemeen ben ik tevreden met de prijs van de meeste producten	In general I am satisfied with the price of most products
Ik ben tevreden met mijn huidig inkomen	I am satisfied with my current income
Ik vind dat ik meer zou moeten verdienen	I think I should earn more
Ik geef vaak complimentjes aan anderen	I often give compliments to others
Ik vind het heel moeilijk om anderen een complimentje te geven	I find it very hard to give others a compliment
Ik sta zelden onder tijdsdruk	I rarely am under time pressure
Ik heb het gevoel voortdurend in tijdnood te zijn	I have the feeling of being in a constant need for time
Ik vind de meeste reclames geloofwaardig	I find most advertisements credible
Ik voel me vaak misleid door reclame	I often feel misled by advertisements

Filler items

Op een vrije avond hou ik ervan om een leuke film te zien	On a free night, I like watching a nice movie
Ik ben een gevoelig persoon	I am a sensitive person
Kinderen zouden veel discipline moeten hebben	Children should have a lot of discipline
Communicatie is heel belangrijk in een relatie	Communication is very important in a relationship
Ik probeer extreme standpunten te vermijden	I try to avoid taking extreme views
Ik hou ervan om dingen te verzamelen	I like collecting things
Ik ben heel nieuwsgierig naar hoe zaken in elkaar zitten	I am very curious about how things work
Kleren tonen een stukje van de persoon die ik ben	Clothes show part of the person I am
Een buitenshuis werkende vrouw met jonge kinderen is nog steeds een goede moeder	A woman working out of home with children is still a good mother
Ik hou ervan om snel te rijden met de auto	I like speeding when driving my car
Ik zou mijn familie bijna alles vergeven	I would forgive my family nearly anything
Ik knip graag bons uit de reclameblaadjes	I like to clip coupons from commercial publications
Ik ben er gerust in dat ik technologie-gerelateerde vaardigheden kan aanleren	I am confident that I can learn technology-related skills
Gevoelens zijn belangrijker dan feiten	Feelings are more important than facts
Ik neem graag de leiding over anderen	I like to take the lead over others
Ik beschouw mezelf als een merkentrouwe consument	I consider myself a brand loyal customers
Ik vind het heel belangrijk om het boodschappen doen goed te organiseren	I find it very important to organize my grocery shopping well
Ik koop geen producten die overdreven verpakt zijn	I don't buy products that have too much packaging
We ervaren een achteruitgang in de levenskwaliteit	We experience a decline in the quality of life
TV-kijken is mijn belangrijkste vorm van ontspanning	Television is my primary form of entertainment
Ik ben een dierenliefhebber	I am an animal-lover
Ik hecht veel belang aan de opinie van mijn vrienden	I attach a lot of importance to the opinion of my friends
Luchtvervuiling is een belangrijk wereldwijd probleem	Air pollution is an important worldwide problem
Ik doe mijn boodschappen in meer dan één supermarkt	I do my grocery shopping in more than one supermarket
Ik ben erg met mijn gezondheid begaan	I am very concerned with my health
Ik vind dat er een geweer aanwezig moet zijn in elk huis	I believe there should be a gun in every house
Menselijk contact bij het verlenen van diensten	Human contact when providing services makes the

maakt het proces prettig voor de consument	process more enjoyable for the consumer
Voor ik een product koop, zal ik steeds de prijs bekijken	Before I buy a product, I will always look at the price

APPENDIX B – 2: ITEM SET 2

Item set 2 consists of a heterogeneous sample of 112 items, taken from Robinson, Shaver and Wrightsman (1991) and Bruner, James and Hensel (2001). This item set was used in study 2 (Chapter 5), 3 (second wave; Chapter 6), and 5 (Chapter 8).

Bipolar items were adapted to Likert format. All items were subjected to a pilot test and rephrased if unclear or when leading to several missing values. The items are listed in alphabetical order (Dutch).

All items were rated on numbered seven point agreement scales, where option 1 was labeled ‘fully disagree’, 4 was labeled ‘neutral’, and 7 was labeled ‘fully agree’, as shown below.

	Fully disagree			neutral			Fully agree
Statement …….	1	2	3	4	5	6	7

Alle mensen moeten de volle vrijheid hebben propaganda te voeren, ook voor wat niet goed is voor hen zelf	Everyone should have the full liberty of propagandizing for what is not good for them
Alles is relatief en er zijn gewoon geen vaststaande regels om naar te leven	Everything is relative, and there just aren't any definite rules to live by
Als consument in de winkel het prijsetiket van een product veranderen, vind ik volstrekt ontoelaatbaar	Changing price tags in the store as a consumer is completely inadmissable
Als een actie een onschuldige persoon zou kunnen schaden, mag deze actie niet ondernomen worden	If an action could harm an innocent other, then it should not be done
Als er iets gebeurt, merk ik over het algemeen dat ik het belang ervan overschat	When something happened, I have generally found that I overestimated its importance
Als het erop aankomt, gaat er niemand veel om geven wat er met je gebeurt	No one is going to care much what happens to you, when you get right down to it
Als ik er niet in slaag te voldoen aan verwachtingen, voel ik me waardeloos	If I fail to live up to expectations, I feel unworthy
Als ik zoals iemand anders wil zijn, probeer ik vaak dezelfde merken te kopen als deze persoon	If I like to be like someone, I often try to buy the same brands that they buy
Alvorens een product te kopen, bekijk ik de prijs per stuk	Before buying a product, I check the unit price
Andere mensen wensen dat ze zo succesvol zouden zijn als ik	Others wish they were as successful as me
Ik ben goed in sport	I am good at sports
Soms denk ik dat ik niets waard ben	At times, I think I am no good at all
Bij het winkelen zoek ik zorgvuldig naar de beste waar voor mijn geld	When I’m shopping, I look carefully to find the best value for money
De afgelopen weken heb ik me voldaan gevoeld over iets dat ik bereikt had	During the past few weeks, I have felt pleased about having accomplished something
De dagelijkse inkopen doen is een sleur	Grocery shopping is a pain
De Franse taal is helemaal niet invloedrijk	The French language is not influential at all
De huidige politieke gebeurtenissen nemen een onvoorspelbare en vernietigende richting	Current political events have taken an unpredictable and destructive course
De kunstenaar en de professor zijn veel belangrijker voor de maatschappij dan de zakenman en de industrieel	The artist and the professor are much more important to society than the businessman and the manufacturer

De laatste tijd waren mensen vaak onvriendelijk tegen me	Recently, people often were unfriendly to me
De meeste mensen leiden een net en behoorlijk leven	Most people lead clean, decent lives
De meeste TV advertenties proberen te werken op de gevoelens van de kijkers	Most TV ads try to work on people's emotions
De meeste verkopers zijn eerlijk in het beschrijven van hun producten	Most salesmen are honest in describing their products
De prijzen van individuele producten verschillen misschien tussen supermarkten, maar over het algemeen zijn de prijzen overal ongeveer hetzelfde.	Prices of individual items may vary between grocery stores, but overall, there isn't much difference in the prices between grocery stores
De zaken die ik bezit, zijn niet zo erg belangrijk voor mij	The things I possess are not that important to me
Een buitenshuis werkende vrouw met jonge kinderen is nog steeds een goede moeder	A woman working out of home with children is still a good mother
Er is weinig dat ik kan doen om veel van de belangrijke dingen in mijn leven te veranderen	There is little I can do to change many of the important things in my life
Er wordt veel te veel nadruk gelegd op succes en vooruitkomen in onze maatschappij	There is far too much emphasis on success and getting ahead in our society
Financiële zekerheid is erg belangrijk voor me	Financial security is very important to me
Geloof in het bovennatuurlijke is een gevaarlijke zelfbegoocheling	Faith in the supernatural is a harmful self-delusion
Het is moeilijk voor mensen om controle te hebben over wat politici doen	It is difficult for people to have much control over the things politicians do in office
Het is slim om vriendelijk te zijn tegen belangrijke mensen, zelfs als je hen niet graag hebt	It is smart to be nice to important people even if you don't really like them
Hoe meer ik te weten kom over producten, hoe moeilijker het wordt om het beste te kiezen	The more I learn about products, the harder it seems to choose the best
Iedereen kan zijn levensstandaard verbeteren als hij probeert	Anyone can raise his standard of living if one tries
Ik begrijp mezelf	I understand myself
Ik ben er gerust in dat ik technologie-gerelateerde vaardigheden kan aanleren	I am confident that I can learn technology-related skills
Ik ben erg begaan met mijn gezondheid	I am very concerned with my health
Ik ben erg emotioneel	I am very emotional
Ik ben het type persoon dat gelooft dat vooruit plannen ervoor zorgt dat de zaken beter	I am the kind of person who believes that planning ahead makes things turn out better

aflopen	
Ik ben lid van een gelukkige familie	I am a member of a happy family
Ik ben nerveus	I am nervous
Ik ben te vermoeid om iets te doen	I am too tired to do anything
Ik ben totaal ontevreden met mijn leven in zijn geheel	I am completely dissatisfied with my life as a whole
Ik ben zelden op mijn gemak in grote groepen mensen	I am seldom at ease in a large group of people
Ik beschouw mezelf als een merkentrouwe consument	I consider myself a brand loyal customer
Ik besteed niet veel aandacht aan de materiële dingen die anderen bezitten	I don't pay much attention to the material objects other people own
Ik eet liever buitenshuis dan thuis	I prefer eating out to home-cooked meals
Ik heb er weinig vat op of mijn gewicht toeneemt, hetzelfde blijft of afneemt.	No matter what I intend to do, if I gain or lose weight, or stay the same in the near future, it is just going to happen
Ik heb favoriete merken, maar ik koop het merk dat een korting geeft	I have favorite brands, but if possible, I buy the brand that offers a cash rebate
Ik heb geen relatie waarin ik me begrepen voel	I don't have any specific relationship in which I feel understood
Ik heb graag dat een verkoper producten bovenhaalt om uit te kiezen	I like having a salesperson bring merchandise out for me to choose from
Ik heb het gevoel voortdurend in tijdnood te zijn	I have the feeling I am in a constant need for time
Ik heb niet veel gemeenschappelijks om over te praten met de mensen om me heen	I don't have much in common to talk about with those around me
Ik houd er echt van om in het middelpunt van de belangstelling te staan	I really like to be the center of attention
Ik kan mijn leven leiden op de manier die ik wil	I can live my life any way I want to
Ik kleed me vaak op een manier die tegen de stroom ingaat, zelfs al zijn anderen daardoor verontwaardigd	I often dress in an unconventional way, even if it offends people
Ik koop dingen graag impulsief	I like to purchase things on a whim
Ik koop geen producten die overdreven verpakt zijn	I don't buy products that have too much packaging
Ik leer graag dingen zelfs als ze mij nooit van nut zullen zijn	I like to purchase things on a whim
Ik lees nieuws en artikels die me informeren over de beste producten voor dagelijks gebruik	I read news features/articles which inform me about the best brands of grocery products

(zoals voeding, huishoudproducten, enz.)	
Ik neem graag risico's	I like to take chances
Ik negeer krantenadvertenties altijd	I always ignore newspaper ads
Ik roddel niet over andermans zaken	I don't gossip about other people's business
Ik vermijd sommige sporten en hobbies omwille van hun gevaarlijke aard	I avoid some sports and hobbies because of their dangerous nature
Ik vind dat een geordend en regelmatig leven bij mijn aard past	I find that a regular and ordered life suits my nature
Ik vind het heel belangrijk om het boodschappen doen goed te organiseren	I find it very important to organize my grocery shopping well
Ik vind het leuk om iets met mijn handen te maken	I enjoy making something with my hands
Ik voel dat ik mezelf volledig in de hand heb als ik voor een publiek spreek	I feel I am in complete possession of myself while speaking for an audience
Ik voel me boordevol energie	I feel full of pep
Ik voel me soms alsof ik op instorten sta	I sometimes feel that I am about to go to pieces
Ik voel me vaak misleid door reclame	I often feel misled by advertizing
Soms voel ik me compleet nutteloos	I certainly feel useless at times
Ik vraag me vaak af of ik de persoon aan het worden ben die ik wil zijn	I often wonder whether I'm becoming the kind of person I want to be
Ik weet veel over huidkanker	I know a lot about skin cancer
Ik wil zeker zijn voor ik iets koop	I want to be sure before I purchase something
Ik winkel omdat dingen kopen me gelukkig maakt	I shop because buying things makes me happy
Ik word graag betrokken in groepsdiscussies	I like to get involved in group discussions
Ik word niet zo vaak uitgenodigd door vrienden als ik zou willen	I don't get invited out by friends as often as I'd really like
Ik word soms kwaad	I get angry sometimes
Ik wou dat ooit iemand mijn biografie schreef	I wish somebody would someday write my biography
Ik zorg ervoor dat mijn garderobe de laatste mode volgt	I keep my wardrobe up-to-date with the latest fashions
Ik zou me beschaamd voelen als ik overdreven veel complimentjes kreeg over mijn aangename persoonlijkheid op mijn eerste afspraak	I would feel strongly embarrassed if I were being lavishly complimented on my pleasant personality by my companion on our first date
Ik zou zeggen dat mensen meestal behulpzaam proberen te zijn	I would say that most of the time people try to be helpful

In de voorbije maanden heb ik me verveeld gevoeld	In the last few months I have been feeling bored
In een groep mensen ben ik zelden het middelpunt van de belangstelling	In a group of people I am rarely the center of attention
In het algemeen vind ik dat ik erg gelukkig ben	In general I find I am very happy
In mijn ervaring zijn mensen behoorlijk koppig en onredelijk	In my experience, people are pretty stubborn and unreasonable
In onze maatschappij worden mensen meer beschouwd als dingen of objecten dan als menselijke wezens	In our society people are becoming things or objects rather than human beings
In tegenstelling tot wat sommigen zeggen, gaat het levenslot van de gemiddelde mens erop achteruit, niet vooruit	In spite of what some people say, the lot of the average person is getting worse, not better
Kortingsbonnen gebruiken maakt het winkelen aangenamer	Using coupons makes shopping more enjoyable
Luchtvervuiling is een belangrijk wereldwijd probleem	Air pollution is an important worldwide problem
Meer geld hebben zou mijn problemen oplossen	Having more money would solve my problems
Menselijk contact bij het verlenen van diensten maakt het proces prettig voor de consument	Human contact when providing services makes the process more enjoyable for the consumer
Mensen die hun leven in een schema passen, missen waarschijnlijk het meeste levensplezier	People who fit their lives to a schedule probably miss most of the joy of living
Mensen hebben de neiging te veel nadruk te leggen op gezag	People tend to place too much emphasis on respect for authority
Mensen stellen me vaak teleur	People often disappoint me
Mensen zouden aandacht moeten besteden aan nieuwe ideeën, zelfs als ze ingaan tegen onze huidige levensstijl	People ought to pay attention to new ideas even if they seem to go against our current way of life
Mijn familieleden geven me het soort steun dat ik nodig heb	Members of my family give me the kind of support that I need
Mijn leven wordt voornamelijk gecontroleerd door machtige anderen	My life is chiefly controlled by powerful others
Mijn vrienden zouden kunnen zeggen dat ik emotioneel ben	My friends might say I'm emotional
Onderwerping aan religieuze autoriteiten is gevaarlijk	Submission to religious authority is dangerous
Over het algemeen zijn vreemdelingen te vertrouwen	Strangers can generally be trusted
Overconsumptie door individuele huishoudens	Overconsumption by individual households has

heeft bijgedragen aan het energieprobleem	contributed to the country's energy problem
Schoolkinderen zouden veel discipline moeten hebben	School children should have plenty of discipline
Sommige mensen worden depressief geboren en blijven zo	Some people are born depressed and stay that way
Soms heb ik het gevoel dat andere mensen me gebruiken	Sometimes I have the feeling that other people are using me
TV reclame helpt me om te weten welke merken de eigenschappen hebben die ik zoek	TV advertising helps me to know which brands have the features I am looking for
TV-kijken is mijn belangrijkste vorm van ontspanning	Television is my primary form of entertainment
Van zodra ze iets verkopen, vergeten de meeste bedrijven de koper	As soon as they make a sale, most businesses forget about the buyer
Vergeleken met andere mensen denk ik dat ik eenzamer ben geweest dan gemiddeld	Compared to other people I think I have been lonelier than average
We ervaren een achteruitgang in de levenskwaliteit	We experience a decline in the quality of life
Winkelen is geen aangename bezigheid voor mij	Shopping is not a pleasant activity to me
Winkelmerken zijn van lage kwaliteit	Store brands are of poor quality

REFERENCES

A

Ajzen, I. (2001), "Nature and operation of attitudes," *Annual Review of Psychology*, 52, 27-58.

Alwin, Duane F., and Jon A. Krosnick (1991), "The reliability of survey attitude measurement: The influence of question and respondent attributes," *Sociological Methods and Research*, 20(1), 139-181.

Anderson, James C., and David W. Gerbing (1988), "Structural equation modeling in practice: A review and recommended two-step approach," *Psychological Bulletin*, 103(3), 411-423.

Andrews, Frank M. (1984), "Construct validity and error components of survey measures: A structural modeling approach," *Public Opinion Quarterly*, 48, 409-442.

Anseel, Frederik, Filip Lievens, and Katrien Vermeulen (2006), "A meta-analysis of response rates in I/O psychology, management and marketing survey research, 1995-2000," Research note presented at the SIOP Conference 2006.

Arbuckle, James L. (1994-2003). *AMOS 5.0.1*. Smallwater Corporation.

Arce-Ferrer, Alvara J. (2006), "An investigation into the factors influencing extreme-response style," *Educational and Psychological Measurement*, 66(3), 374-392.

Arthur, Artur Z., and Anna Freemantle (1966), "Extreme response bias and associative commonality," *Psychonomic Science*, 6(8), 399-399.

Ayidiya, Stephen, and McKee J. McClendon (1990), "Response effects in mail surveys," *Public Opinion Quarterly*, 54, 229-247.

B

Babakus, Emin, and Carl E. Ferguson Jr. (1988), "On choosing the appropriate measure of association when analyzing rating scale data," *Journal of the Academy of Marketing Science*, 16(1), 95-102.

Babakus, Emin, Carl E. Ferguson, Jr. and Karl G. Jöreskog (1987), "The sensitivity of confirmatory maximum likelihood factor analysis," *Journal of Marketing Research*, 24, 222-228.

Bachman, Jerald G., and Patrick M. O'Malley (1984), "Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles," *Public Opinion Quarterly*, 48, 491-509.

Bandalos, Deborah (2006). The use of monte carlo studies in structural equation modeling research. In Gregory R. Hancock and Ralph O. Mueller (Eds.), *Structural Equation Modeling: A second course*, Information Age Publishing.

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, 361-370.

Bauer, Daniel J., and Patrick J. Curran (2004), "The integration of continuous and discrete latent variable models: Potential problems and promising opportunities," *Psychological Methods*, 9(1), 3-29.

Baumgartner, Hans and Jan-Benedict E.M. Steenkamp (2001), "Response styles in marketing research: A cross-national investigation," *Journal of Marketing Research*, 38 (May), 143-156.

Becker, Gilbert (2000), "How important is transient error in estimating reliability? Going beyond simulation studies," *Psychological Methods*, 5(3), 370-379.

Bentler, P.M. (1969). “Semantic space is (approximately) bipolar,” *Journal of Psychology*, 71, 33-40.

Bentler, P.M. (1990), “Comparative fit indices in structural models,” *Psychological Bulletin*, 107, 238-246.

Bentler, P.M., Douglas N. Jackson, and Samuel Messick (1971), “Identification of content and style: A two-dimensional interpretation of acquiescence,” *Psychological Bulletin*, 76(3), 186-204.

Billiet, Jaak B., and McKee J. McClendon (2000), “Modeling acquiescence in measurement models for two balanced sets of items,” *Structural Equation Modeling*, 7(4), 608-628.

Bishop, George F. (1987), “Experiments with the middle response alternative in survey questions,” *Public Opinion Quarterly*, 51, 220-232.

Blanton, Hart, and James Jaccard (2006), “Arbitrary metrics in psychology,“ *American Psychologist*, 61(1), 27-41.

Block, Jack (1971), “On further conjectures regarding aquiescence,” *Psychological Bulletin*, 76(3), 205-210.

Bohrnstedt, G. W. (1983), *Handbook of Survey Research*, Academic Press.

Bollen, Kenneth A., and Kenney H. Barb (1981), “Pearson's r and coarsely categorized measures,” *American Sociological Review*, 46(2), 232-239.

Browne, M.W., and Cudeck R. (1993), “Alternative ways of assessing model fit,” In Bollen, K.A. and Long, J.S. (Eds.), *Testing structural equation models*, Newbury Park CA: Sage, p. 136-162.

Bruner, Gordon C., Karen E. James, and Paul J. Hensel (2001), *Marketing Scales Handbook, A Compilation of Multi-Item Measures, Volume III*. American Marketing Association, Chicago, Illinois USA.

Budd, Richard J. (1987), "Response bias and the theory of reasoned action," *Social Cognition*, 5(1), 95-107.

C

Campbell, D.T., and D. Fiske (1959), "Convergent and discriminant validation by the multitrait-multimehtod matrix," *Psychological Bulletin*, 56, 81-105.

Carroll, James M., Michelle S. M. Yik, James A. Russell, and Lisa Feldmann Barrett (1999), "On the psychometric principles of affect," *Review of General Psychology*, 3(1), 14-22.

Chan, Wai, and Daniel W.-L. Chan (2004), "Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: a simulation study," *Psychological Methods*, 9(3), 369-385.

Charles, Eric P. (2005), "The correction for attenuation due to measurement error: clarifying concepts and creating confidence sets," *Psychological Methods*, 10(2), 206-226.

Cheung, Gordon W., and Roger B. Rensvold (2000), "Assessing extreme and acquiescence response sets in cross-cultural research using Structural Equation Modeling," *Journal of Cross-cultural Psychology*, 31(2), 187-212.

Cheung, Gordon W., and Roger B. Rensvold (2002), "Evaluating goodness-of-fit indices for testing measurement invariance," *Structural Equation Modeling*, 9(2), 233-255.

Churchill, Gilbert A. Jr. (1979), "A paradigm for developing better measures of marketing constructs," *Journal of Marketing Research* 16(Feb), 64-73.

Clancy, Kevin J., Robert A. Wachsler (1968), "Positional effects in shared cost surveys," *Public Opinion Quarterly*, 35, 258-265.

Coenders, Germà, and Willem E. Saris (2000), "Testing nested additive, multiplicative, and general multitrait-multimethod models," Structural Equation Modeling, 7(2), 219-250.

Converse, Jean M. (1984), "Strong arguments and weak evidence: The open/closed questioning controversy of the 1940s," *Public Opinion Quarterly*, 48, 267-282.

Cordery, John L., and Peter P. Sevastos (1993), "Responses to the original and revised Job Diagnostic Survey: Is education a factor in responses to negatively worded items?", *Journal of Applied Psychology*, 78(1), 141-143.

Couch, Arthur, and Kenneth Keniston (1960), "Yeasayers and naysayers: Agreeing response set as a personality variable," *Journal of Abnormal and Social Psychology*, 60(2), 151-174.

Cox, Eli P. III (1980), "The optimal number of response alternatives for a scale: A review," *Journal of Marketing Research*, 17, 407-422.

Cronbach, Lee J. (1946), "Response sets and test validity," *Educational and Psychological Measurement*, 6, 475-494.

Cronbach, Lee J. (1950), "Further evidence on response sets and test design," *Educational and Psychological Measurement*, 10, 3-31.

Curran, P. J., West, S. G, & Finch, J. F. (1996), "The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis," *Psychological Methods*, 1, 16-29.

Curran, Patrick J., and Kenneth A. Bollen (2001), "The best of both worlds: Combining autoregressive and latent curve models," In Collins, Linda M., and Aline G. Sayer, *New methods for the analysis of change*, APA, Washington D.C.

D

De Cannière, Marie Hélène (2006), *Invsetigating the moderating influence of customer characteristics on the relationship between behavioural loyalty and its antecedents*, Doctoral dissertation, Faculty of Economics and Business Administration, Ghent University.

de Leeuw, Edith D. (2005), "To mix or not to mix: Data collection modes in surveys," *Journal of Official Statistics*, 21(2), 2005. pp. 233-255

DeMaio, T.J., and J.M. Rothgeb (1996), "Cognitive interviewing techniques: In the lab and in the field," In N. Schwarz and S. Sudman (Eds.), Answering Questions: Methodology for Determining Cognitive and Communciative Processes in Survey Research, p. 177-195. San Francisco: Jossey-Bass.

Deutskens, Elisabeth C. (2006), *From paper-and-pencil to screen-and-keyboard: Studies on the effectiveness of Internet-based marketing research,* Doctoral dissertation, Maastricht University.

Deutskens, Elisabeth C., Ko de Ruyter, and Martin G.M. Wetzels (2006), "An assessment of equivalence between online and mail surveys in service research," *Journal of Service Research,* Forthcoming.

Deutskens, Elisabeth C., Ko de Ruyter, Martin G.M. Wetzels, and Paul Oosterveld (2004), "Response rate and response quality of Internet-based surveys: An experimental study," *Marketing Letters, 15(1), 21-36.*

Dodds, William B., Kent B. Monroe, and Dhruv Grewal (1991), "Effects of price, brand, and store information on buyers' product evaluations," *Journal of Marketing Research*, 28 (August), 307-19.

Drolet, Aimee, and Donald G. Morrison (2001), "Do we really need multiple-item measures in service research?," *Journal of Service Research*, 3(3), 196-204.

E

Elliot, Lois Lawrence (1961), "Effects of item construction and respondent aptitude on response acquiescence," *Educational and Psychological Measurement*, 21(2), 405-415.

Enders, Craig K. (2006). Analyzing structural equation models with missing data. In Gregory R. Hancock and Ralph O. Mueller (Eds.), *Structural Equation Modeling: A second course,* Information Age Publishing, USA.

F

Feldman, Jack M., and John G. Lynch, Jr. (1988), "Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior," *Journal of Applied Psychology*, 73(3), 421-435.

Feldman, Jack M. (1992), "Constructive processes as a source of context effects in survey research: Explorations in self-generated validity," In Norbert Schwarz and Seymour Sudman (Eds.), *Context effects in social and psychological research,* Springer-Verlag, 49-61.

Ferrando, P.J. (2002), "An IRT-based two-wave model for studying short-term stability in personality measurement," *Applied Psychological Measurement*, 26, 286-301.

Ferrando, Pere J., and Urbano Lorenzo-Seva (2005), "IRT-related factor anlaytic procedures for testing the equivalence of paper-and-pencil and Internet-administered questionnaires," *Psychological Methods*, 10 (2), 193-205.

Finney, Sara J., and Christine DiStefano (2006), "Nonnormal and categorical data in structural equation modeling," In Gregory R. Hancock and Ralph O. Mueller (Eds.), *Structural Equation Modeling: A second course,* Information Age Publishing, USA.

Flora, David B., and Patrick J. Curran (2004), "An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data," *Psychological Methods*, 9(4), 466-491.

Forehand, Garlie A. (1962), "Relationships among response sets and cognitive behaviors," *Educational and Psychological Measurement*, 2, 287-302.

Fornell, Claes, and David F. Larcker (1981), "Evaluating Structural Equation Models with unobservable variables," *Journal of Marketing Research*, 18, 39-50.

Frederiksen, Norman, and Samuel Messick (1959), "Response set as a measure of personality," *Educational and Psychological Measurement*, 19(2), 137-157.

G

Gage, N.L., George S. Leavitt, and George C. Stone (1957), "The psychological meaning of acquiescence set for authoritarianism," *Journal of Abnormal and Social Psychology*, 55, 98-103.

Gagné, Phill (2006), "Mean and covariance structure mixture models," In Gregory R. Hancock and Ralph O. Mueller (Eds.), *Structural Equation Modeling: A second course*, Information Age Publishing, USA. In Gregory R. Hancock and Ralph O. Mueller (Eds.), *Structural Equation Modeling: A second course*, Information Age Publishing, USA.

Ganzach, Yoav (1997), "Misleading interaction and curvilinear terms," *Psychological Methods*, 2(3), 235-247.

Green, Donald Philip, Susan Lee Goldman, and Peter Salovey (1993), "Measurement error masks bipolarity in affect ratings," *Journal of Personality and Social Psychology*, 64(6), 1029-1041.

Green, Paul E., and Vithala R. Rao (1970), "Rating scales and information recovery: How many scales and response categories to use?" *Journal of Marketing*, 34(July), 33-39.

Green, Samuel B. (1991), "How many subjects does it take to do a regression analysis?" *Multivariate Behavioral Research*, 26(3), 499-510.

Green, Samuel B., and Scott L. Hershberger (2000), "Correlated errors in true score models and their effect on coefficient alpha," *Structural Equation Modeling*, 7(2), 251-270.

Greene, William H. (2003). *Econometric Analysis*. Fifth Edition. Prentice Hall.

Greenleaf, Eric A. (1992a), "Improving rating scale measures by detecting and correcting bias components in some response styles," *Journal of Marketing Research*, 29(2), 176-188.

Greenleaf, Eric A. (1992b), "Measuring extreme response style, "*Public Opinion Quarterly*, 56(3), 328-350.

Griffis, Stanley E., Thomas J. Goldsby, and Martha Cooper (2003), "Web-based and mail surveys: A comparison of response, data, and cost," *Journal of Business Logistics*, 24(2), 237-258.

Grouzet, Frederick M.E., Nancy Otis, and Luc G. Pelletier (2005), "Longitudinal cross-gender factorial invariance of the academic motivation scale," *Journal of Educational Psychology*, 97(2), 170-183.

Gunter, Barrie, David Nicholas, Paul Huntington and Peter Williams (2002), "Online versus offline research: Implications for evaluating digital media", *Aslib Proceedings*, 54(4), 229-239.

H

Hamilton, David L. (1968), “Personality attributes associated with extreme response style,” *Psychological Bulletin*, 69, 3, 192-203.

Herche, Joel, and Brian Engelland (1996), “Reversed-Polarity Items and Scale Unidimensionality,” *Journal of the Academy of Marketing Science*, 24(4), 366-374.

Herzog, A. Regula, and Jerald G. Bachman (1981), “Effects of questionnaire length on response quality,” *Public Opinion Quarterly*, 45, 549-559.

Hippler, Hans-J., and Norbert Schwarz (1986), “Not forbidding isn't allowing: The cognitive bias of the forbid-allow asymmetry,” *Public Opinion Quarterly*, 50, 87-96.

Horran, Patrick M., Christine DiStefano, and Robert W. Motl (2003), “Wording effects in Self-Esteem scales: Methodological artifact or response style?” Structural Equation Modeling, 10 (3), 435-455.

Hox, Joop (1997), “From theoretical concept to survey question,” In L. Lyberg, P. Biemer, M.Collins, E. De Leeuw, C. Dippo, N. Schwarz, D. Trewin (Eds), *Survey Measurement and Process Quality*, Wiley, p. 47-69.

Hu, Li-tze and Bentler, P.M. (1999), “Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives,” *Structural Equation Modeling*, 6, 1, 1-55.

Hui, C. Harry, and Harry C. Triandis (1985), “The instability of response sets,” *Public Opinion Quarterly*, 49, 253-260.

Husek, T.R. (1961), “Acquiescence as a response set and as a personality characteristic,” *Educational and Psychological Measurement*, 2, 295-307.

J

Jackson, Douglas N., and Leonard Pacine (1961), “Response styles and academic achievement,” *Educational and Psychological Measurement*, 21(4), 1015-1028.

Jackson, Douglas N. (1967), "Acquiescence response styles: Problems of identification and control," In Irwin A. Berg (Ed.), *Response set in personality assessment*, Aldine Publishing Company, Chicago, p. 71-114.

Jackson, Douglas N., and Samuel Messick (1958), "Content and style in personality assessment," *Psychological Bulletin*, 55(4), 243-252.

Jedidi, Kamel, Harsharanjeet S. Jagpal, Wayne S. DeSarbo (1997), "Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity," *Marketing Science*, 16(1), 39-59.

Jobe, J.B., and D.J. Mingay (1989), "Cognitive research improves questionnaires," *American Journal of Public Health*, 79, 1053-1055.

Johnson, Eric J. (2001), "Digitizing consumer research," *Journal of Consumer Research*, 28(Sept), 331-336.

Johnson, Timothy, Patrick Kulesa, Young Ik Cho, and Sharon Shavitt (2005), "The Relation between Culture and Response Styles," *Journal of Cross-cultural Psychology*, 36(2), 264-277.

Jordan, Lawrence A., Alfred C. Marcus, and Leo G. Reeder (1980), "Response styles in telephone and household interviewing: A field experiment," *Public Opinion Quarterly*, 44(2), 210-222.

Jöreskog, Karl G. (1971), "Simultaneous factor analysis in several populations," *Psychometrika*, 36 (Dec.), 409-426.

K

Kalton, G., J. Roberts, and D. Holt (1980), "The effects of offering a middle response option with opinion questions," *The Statistician*, 29(1), 65-78.

Kiesler, Sara, and Lee S. Sproul (1986), "Response effects in the electronic survey," *Public Opinion Quarterly*, 50, 402-143.

Knauper, Barbel (1999), "The impact of age and education on response order effects in attitude measurement," *Public Opinion Quarterly*, 63(3), 347-370.

Knowles, Eric S. (1988), "Item context effects on personality scales: Measuring changes the measure," *Journal of Personality and Social Psychology*, 55(2), 312-320.

Knowles, Eric S., Michelle C. Coker, Deborah A. Cook, Steven R. Diercks, Mary E. Irwin, Edward J. Lundeen, John W. Neville, and Mark E. Sibicky (1992). Order effects within personality measures. In Norbert Schwarz and Seymour Sudman (Eds.), *Context effects in social and psychological research,* Springer-Verlag.

Knowles, Eric S., and Kobi T. Nathan (1997), "Acquiescent responding in self-reports: Cognitive style or social concern?" *Journal of Research in Personality*, 31, 293-301.

Kraemer, Helena Chmura (1980), "Robustness of the distribution theory of the product moment correlation coefficient," *Journal of Educational Statistics*, 5(2), 115-128.

Kraut, Allen I., Alan D. Wolfson, and Alan Rothenberg (1975), "Some effects of position on opinion survey items," *Journal of Applied Psychology*, 60(6), 774-776.

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

Krosnick, Jon A. (1999), "Survey research," Annual Review of Psychology, 50, 537-567.

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

L

Leite, Walter L., and S. Natasha Beretvas (2005), "Validation of scores on the Marlowe-Crowne Social Desirability Scale and the Balanced Inventory of Desirable Responding," *Educational and Psychological Measurement*, 65 (1), 140-154.

Likert, Rensis (1932), "A technique for the measurement of attitudes," *Archives of Psychology*, 140, 44-53.

Lindell, Michael K., and David J. Whitney (2001), "Accounting for Common Method Variance in Cross-sectional research designs," *Journal of Applied Psychology*, 86 (1), 114-121.

Little, Todd D. (1997), "Mean and Covariance Structures (MACS) analyses of cross-cultural data: Practical and theoretical issues," *Multivariate Behavioral Research*, 32(1), 53-76.

Little, Todd D. (2000), "On the Comparability of Constructs in Cross-Cultural Research: A Critique of Cheung and Rensvold," *Journal of Cross-Cultural Psychology*, 31(2), 213-219.

Little, Todd D., William A. Cunningham, and Golan Shahar (2002), "To parcel or not to parcel: Exploring the question, weighing the merits," *Structural Equation Modeling*, 9(2), 151-173.

Lo, Y., N.R. Mendell, and D.B. Rubin (2001), "Testing the number of components in a normal mixture," *Biometrika*, 88, 767-778.

Lubke, Gitta H., and Bengt Muthén (2005), "Investigating population heterogeneity with factor mixture models," *Psychological Methods*, 10(1), 21-39.

M

Marín, Gerardo, Raymond J. Gamba, and Barbara V. Marìn (1992), "Extreme response styles and acquiescence among Hispanics," *Journal of Cross-Cultural Psychology*, 23(December), 498-509.

Marsh, Herbert W. (1993), "Stability of individual differences in multiwave panel studies: Comparison of simplex models and one-factor models," *Journal of Educational Measurement*, 30(2), 157-183.

Marsh, Herbert W. (1996), "Positive and negative global self-esteem: A substantively meaningful distinction or artifactors?" *Journal of Personality and Social Psychology*, 70(4), 810-819.

Marsh, Herbert W., and Allen Parducci (1977), "Natural anchoring at the neutral point of category rating scales," *Journal of Experimental Social Psychology*, 14, 193-204.

Marsh, Herbert W., John R. Balla, and Roderick McDonald (1988), "Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size," *Psychological Bulletin*, 103(3), 391-410.

Marsh, Herbert W., Kit-Tai Hau, and Zhonglin Wen (2004), "In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings," *Structural Equation Modeling*, 11(3), 320-341.

Matell, Michael S., and Jacob Jacoby (1972), "Is there an optimal number of alternative for likert-scale items?" *Journal of Applied Psychology*, 56(6), 506-509.

Maydeu-Olivares, Albert (2005), "Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data," Multivariate Behavioral Research, 40(2), 261-279.

McClendon, McKee J. (1991a), "Acquiescence and response-order effects in interview surveys," *Sociological Methods and Research*, 20(1), 60-103.

McClendon, McKee J. (1991b), "Acquiescence: Tests of the cognitive limitations and question ambiguity hypotheses," *Journal of Official Statistics*, 7(2), 153-166.

McFarland, Lynn A., Ann Marie Ryan, and Aleksander Ellis (2002), "Item placement on a personality measure: Effects on faking behavior and test measurement Properties," *Journal of Personality Assessment*, 78(2), 348-369.

McGee, Richard K. (1967), "Response set in relation to personality: An orientation," In Irwin A. Berg (Ed.), *Response set in personality assessment,* Aldine Publishing Company, Chicago, p. 1-31.

McLachlan, G., and D. Peel (2000), *Finite mixture models*, John Wiley, New York.

McPherson, Jason, and Philip Mohr (2005), "The role of item extremity in the emergence of keying-related factors: An exploration with the Life Orientation Test," *Psychological Methods*, 10(1), 120-131.

Meade, Adam W., Gary J. Lautenschlager (2004), "A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance," *Organizational research methods,* 7(4), 361-388.

Meredith, W. (1993), "Measurement invariance, factor analysis and factorial invariance," *Psychometrika*, 58, 525-543.

Mick, David Glen (1996), "Are studies of dark side variables confounded by socially desirable responding? The case of materialism," *Journal of Consumer Research*, 23(Sept), 106-119.

Miller, George A. (1956), "The magical number seven, plus or minus two: some limits on our capacity for processing information," *Psychological Review*, 63, 81-97.

Mirowsky, John, and Catherine E. Ross (1991), "Eliminating defense and agreement bias from measures of the sense of control: A 2x2 index," *Social Psychology Quarterly*, 54, 2, 127-145.

Mittal, Vikas, and Wagner A. Kamakura (2001), "Satisfaction, Repurchase Intent, and Repurchase Behavior: Investigating the Moderating Effect of Customer Characteristics," *Journal of Marketing Research*, 38 (Feb), 131-142.

Motl, Robert W., and Christine DiStefano (2002), "Longitudinal invariance of self-esteem and method effects associated with negatively worded items," *Structural Equation Modeling*, 9(4), 562-578.

Moxey, Linda M., and Anthony J. Sanford (1992), "Context effects and the communicative functions of quantifiers: Implications for their use in attitude research," In Norbert Schwarz and Seymour Sudman (Eds.), *Context effects in social and psychological research*, Springer-Verlag, p. 279-296.

Muthén, Linda K., and Bengt O. Muthén (2004). *Mplus, Statistical analysis with latent variables: user guide*. Third Edition, Los Angeles.

Muthén, L.K., and B. Muthén (2006). *Mplus User's Guide*. Fourth Edition. Los Angeles, CA: Muthén & Muthén.

N

Narayan, Sowmya, and Jon A. Krosnick (1996), "Education moderates some response effects in attitude measurement," *Public Opinion Quarterly*, 60, 58-88.

Nylund, Karen L., Tihomir Asparouhov, and Bengt O. Muthén (2006), "Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study," Paper downloaded from www.statmodel.com on June 12, 2006.

O

Olsson, Ulf (1979)," Maximum Likelihood Estimation of the Polychoric Correlation Coefficient," *Psychometrika*, 44 (December), 443-60.

O'Neill, Harry W. (1967), "Response Style Influence in Public Opinion Surveys," *Public Opinion Quarterly*, 31(1), 137-157.

Osgood, Charles E. (1941), "Ease of individual jugdment-process in relation to polarization of attitudes in the culture," Journal of Social Psychology, 14, 403-418.

Ostrom, Thomas M., Andrew L. Betz, and John J. Skowronski (1992), "Cognitive representation of bipolar survey items," In Norbert Schwarz and Seymour Sudman (Eds.), *Context effects in social and psychological research*, Springer-Verlag.

P

Parasuraman, A. (2000), "Technology Readiness Index (TRI): A multiple-Item scale to measure readiness to embrace new technologies," Journal of Service Research, 2(4), 2-4.

Parasuraman, A., Valarie A. Zeithaml, and Arvind Malhotra (2005), "A multiple-item scale for assessing electronic service quality," Journal of Service Research, 7 (3), 213-233.

Parducci, Allen (1974), "Contextual effect: A range-frequency analysis," In E.C. Carterette, and M.P. Friedman (Eds.), *Handbook of perception, volume 2*, Academic Press, New York, p.128-141.

Paulhus, Delroy L. (1991). Measurement and control of response bias. In John P. Robinson, Phillip R. Shaver, and Lawrence S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes*, Academic Press.

Peabody, Dean (1962), "Two components in bipolar scales: Direction and extremeness," *Psychological Review*, 62(2), 65-73.

Peabody, Dean (1966), "Authoritarianism scales and response bias," *Psychological Bulletin*, 1, 11-23.

Pearson, Karl and Egon Pearson (1922), "On Polychoric Coefficients of Correlation," *Biometrika*, 14 (July), 127-56.

Ployhart, Robert E., and Frederick L. Oswald (2004). "Applications of Means and Covariance Structure Analysis: Integrating Correlational and Experimental Approaches," *Organizational Research Methods*, 7(1), 27-65.

Ployhart, Robert E., Jeff A. Weekley, Brian C. Holtz, and Cary Kemp (2003), "Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable?" *Personnel Psychology*, 56(Autumn), 733-752.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff (2003), "Common method biases in behavioral research: A critical review of the literature and recommended remedies," *Journal of Applied Psychology*, 88(5), 879-903.

Q

Quilty, Lena C., Jonathan M. Oakman, and Evan Risko (2006), "Correlates of the Rosenberg Self-Esteem Scale method effects," *Structural Equation Modeling* 13(1), 99-117.

R

Raju, Nambury S., Larry J. Laffitte, and Barbara M. Byrne (2002), "Measurement Equivalence: A Comparison of Methods Based on Confirmatory Factor Analysis and Item Response Theory," *Journal of Applied Psychology*, 87(3), 517-529.

Ray, John J. (1979), "Is the acquiescent response style problem not so mythical after all? Some results from a successful balanced F scale," *Journal of Personality Assessment*, 43(6), 638-643.

Ray, John J. (1983), "Reviving the problem of acquiescent response bias," *Journal of Social Psychology*, 121, 81-96.

Reise, Steven P., Keith F. Widaman, and Robin H. Pugh (1993), "Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance," *Psychological Bulletin*, 114(3), 552-566.

Rindskopf, David (2003), "Mixture or homogeneous? Comment on Bauer and Curran (2003)," *Psychological Methods*, 8(3), 364-368.

Robinson, John P., Phillip R. Shaver, Lawrence S. Wrightsman (1991), *Measures of Personality and Social Psychological Attitudesof Social Psychological Attitudes: Volume 1*, Academic Press.

Rorer, Leonard G. (1965), "The great response-style myth," *Psychological Bulletin*, 63(3), 129-156.

Rossi, Peter E., Zvi Gilula, and Greg M. Allenby (2001), "Overcoming scale usage heterogeneity: A Bayesian hierarchical approach," *Journal of the American Statistical Association*, 96 (453), 20-31.

Rossiter, J.R. (2002), 'The C-OAR-SE procedure for scale development in marketing," *International Journal of Research in Marketing*, 19(4), 305-335.

Russell, James A., and James M. Carroll (1999), "On the bipolarity of positive and negative affect," *Psychological Bulletin*, 125, 1, 3-30.

S

Sackett, Paul R., and Hyuckseung Yang (2000), "Correction for range restriction: an expanded typology," *Journal of Applied Psychology*, 85(1), 112-118.

Samejima, F. (1969), "Estimation of latent ability using a response pattern of graded scores," *Psychometrika Monographs*, 34 (Suppl. 17).

Saris, Willem E., and Chris Aalberts (2003), "Different explanations for correlated disturbance terms in MTMM studies," Structural Equation Modeling, 10(2), 193-213.

Saris, Willem E., Albert Satorra, and Germà Coenders (2004), "A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design," Sociological Methodology, 34, p. 311-347.

Satorra, Albert, and Peter M. Bentler (2001), "A scaled difference chi-square test statistic for moment structure analysis," *Psychometrika*, 66, 507-514.

Schaeffer, Nora Cate, and Stanley Presser (2003), "The science of asking questions," *Annual Review of Sociology*, 29, 65-88.

Schafer, J.L. (1999), *NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2*, Software for Windows 95/98/NT, available from http://www.stat.psu.edu/~jls/misoftwa.html.

Schafer, Joseph L., and John W. Graham (2002), "Missing data: Our view of the state of the art," *Psychological Methods*, 7(2), 147-177.

Schimmack, Ulrich, Ulf Böckenholt, and Rainer Reisenzein (2002). "Response styles in affect ratings: Making a mountain out of a molehill," *Journal of Personality Assessment*, 78(3), 461-483.

Schmitt, N., and Stults, D.M. (1985), "Factors defined by negatively keyed items: The result of careless respondents?" *Applied Psychological Measurement*, 9, 367-373.

Schriesheim, C.A., and Hill, K.D. (1981), "Controlling acquiescence response bias by item reversals: The effect on questionnaire validity," *Educational and Psychological Measurement*, 41, 1101-1114.

Schriesheim, C.A., Eisenbach, R.J. and Hill, K.D. (1991), "The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An

experimental investigation," *Educational and Psychological Measurement*, 51, 67-78.

Schuman, Howard (1992), "Context effects: State of the past/state of the art," In Norbert Schwarz and Seymour Sudman (Eds.), *Context effects in social and psychological research*, Springer-Verlag, p. 5-20.

Schuman, Howard, and Stanley Presser (1981). *Questions and answers in attitude surveys: Experiments on question form, wording and context*. Academic Press.

Schwarz, Norbert (1997), "Moods and attitude judgments: A comment on Fishbein and Middlestadt," *Journal of Consumer Psychology*, 6(1), 93-98.

Shaeffer, Eric M., Jon A. Krosnick, Gary E. Langer, Daniel M. Merkle (2005), "Comparing the quality of data obtained by minimally balanced and fully balanced attitude questions," *Public Opinion Quarterly*, 69(3), 417-428.

Shulman, Art (1973), "A comparison of two scales on extremity response bias," Public Opinion Quarterly, 37 (Autumn), 407-412.

Shiv, B., and Alexander Fedorikhin (1999), "Heart and mind in conflict: the interplay of affect and cognition in consumer decision making," *Journal of Consumer Research*, 26(3), 278-292.

Simsek, Zeki, and John F. Veiga (2001), "A primer on Internet organizational surveys," *Organizational Research Methods*, 4(July), 218-235.

Sirdeshmukh, Deepak, Jagdip Singh, and Barry Sabol (2002). "Consumer trust, value, and loyalty in relational exchanges," *Journal of Marketing*, 66(1), 15-37.

Spector, P.E., P.T. Van Katwyck, M.T. Brannick, and P.Y. Chen (1997), "When two factors don't reflect two constructs: How item characteristics can produce artifactual factors," *Journal of Management*, 23, 659-677.

Srinivasan, V., and Amiya K. Basu (1989), "The metric quality of ordered categorical data," *Marketing Science* 8(3), 205-230.

Steenkamp, Jan-Benedict E.M., and Katrijn Gielens (2003), "Consumer and market drivers of the trial probability of new consumer packaged goods," *Journal of Consumer Research*, 30(4), 368-384.

Steenkamp, Jan-Benedict E.M., and Hans Baumgartner (1998), "Assessing measurement invariance in cross-national consumer research". *Journal of Consumer Research*, 25, 78-90.

Steenkamp, Jan-Benedict E.M., Frenkel ter Hofstede, and Michel Wedel (1999), "A cross-national investigation into the individual and national cultural antecedents of conumser innovativeness," Journal of Marketing, 63 (2), 55-69.

Steiger, J.H. (1990), "Structural model evaluation and modification: An interval estimation approach," *Multivariate Behavioral Research*, 25, 173-180.

Stein, Gertrude (1928), *Useful Knowledge*, Payson and Clarke, New York.

Stening, B.W., and J.E. Everett (1984), "Response styles in a cross-cultural managerial study," *Journal of Social Psychology*, 122, 151-156.

Szymanski, David M., and Richard T. Hise (2000) "e-Satisfaction: An Initial Examination", *Journal of Retailing*, 76(3), 309-322.

T

Tabachnick, Barbara G., and Linda S. Fidell (1996), *Using Multivariate Statistics, 3rd Edition*, Harper Collins.

Thompson, Lori Foster, Eric A. Surface, Don L. Martin, and Michael G. Sanders (2003), "From paper to pixels: Moving personnel surveys to the Web," *Personnel Psychology*, 56 (Spring), 197-227.

Thompson, Marilyn S., and Samuel B. Green (2006), "Evaluating between-group differences in latent variable means," In Gregory R. Hancock and Ralph O. Mueller (Eds.), *Structural Equation Modeling: A second course*, Information Age Publishing.

Tomas, José M., and Amparo Oliver (1999), "Rosenberg's Self-Esteem Scale: Two Factors or Method Effects?" *Structural Equation Modeling*, 6(1), 84-98.

Tourangeau, Roger (1984), "Cognitive science and survey methods." In T. Jabine, M. Straf, J. Tanur, and Roger Tourangeau (Eds.), *Cognitive aspects of survey design: Building a bridge between disciplines*, National Academy Press, Washington D.C., p. 73-100

Tourangeau, Roger (1992), "Context effects on responses to attitude questions: attitudes as memory structures," In Schwarz, Norbert, and Seymour Sudman (Eds.). *Context effects in social and psychological research*. Springer-Verlag.

Tourangeau, Roger, Eleanor Singer, and Stanley Presser (2003), "Context effects in attitude surveys: Effects on remote items and impact on predictive validity," *Sociological Methods and Research*, 31(4), 486-513.

Tourangeau, Roger, Kenneth A. Rasinski, Norman Bradburn, and Roy D'Andrade (1989), "Carryover effects in attitude surveys," *Public Opinion Quarterly*, 53, 495-524.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski (2000), *The psychology of survey response*. Cambridge University Press.

Traub, Ross E. (1994). *Reliability for the Social Sciences: Theory and Applications*. Sage Publications.

Truell, Allen D. (2003), "Use of Internet tools for survey research," *Information Technology, Learning, and Performance Journal*, 21(1), 31-37.

Tuerlinckx, Francis, and Paul De Boeck (2001), "The effect of ignoring item interactions on the estimated discrimination parameters in Item Response Theory," *Psychological Methods*, 6(2), 181-195.

V

Van der Kloot, Willem A., Pieter M. Kroonenberg, and Dini Bakker (1985), "Implicit theories of personality: Further evidence of extreme response style," *Multivariate Behavioral Research*, 20, 369-387.

Van Herk, Hester, Ype H. Poortinga, and Theo M. M. Verhallen (2004), "Response Styles in Rating Scales: Evidence of Method Bias in Data From Six EU Countries," *Journal of Cross-cultural Psychology*, 35(3), 346-360.

Vandenberg, Robert J., and Charles E. Lance (2000), "A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research," *Organizational Research Methods*, 3(1), 4-70.

Venkatesh, Shankar, Amy K. Smith, and Arvind Rangaswamy (2003), "Customer satisfaction and loyalty in online and offline environments", *International Journal of Research in Marketing*, 20, 153- 175.

W

Warr, P., Barter, J., & Brownbridge, G. (1983). On the independence of positive and negative affect. *Journal of Personality and Social Psychology*, 44, 644-651.

Watson, David (1988). The vicissitudes of mood measurement: effects of varying descriptors, time frames, and response formats on measures of positive and negative affect. *Journal of Personality and Social Psychology*, 55(1), 128-141.

Watson, Dorothy (1992), "Correcting for acquiescent response bias in the absence of a balanced scale: An application to class consciousness", *Sociological Methods and Research*, 21(1), 52-88.

Wedel, Michel, and Wayne S. DeSarbo (1995), "A mixture likelihood approach for generalized linear models," *Journal of Classification*, 12, 21-55.

Wedell, Douglas H., and Allen Parducci (1988), "The category effect in social judgment: experimental ratings of happiness," *Journal of Personality and Social Psychology*, 55, 3, 341-356.

Wegener, Duane T., John Downing, Jon A. Krosnick, and Richard E. Petty (1995) "Measures and manipulations of strength-related properties of attitudes: Current practices and future directions," In Petty, R.E., and J.A. Krosnick, *Attitude strength: Antecedents and consequences*, Mahwah, NJ: Erlbaum.

Weijters, Bert, Niels Schillewaert, and Maggie Geuens (2004), "Measurement bias due to response styles: A structural equation model assessing the effect of modes of data collection," paper presented at the Marketing Science Conference 2004, Rotterdam. Available online at www.vlerick.com as Vlerick Leuven Gent Management School working paper 2004/20.

Welkenhuysen-Gybels, Jerry, Jaak Billiet, and Bart Cambré (2003), "Adjustment for acquiescence in the assessment of the construct equivalence of likert-type score items," *Journal of Cross-cultural Psychology*, 34(6), 702-722.

Wilson, T.D., and S. Hodges (1992), "Attitudes as temporary constructions," In L. Martin, and A. Tesser (Eds.), *The construction of social judgments*, Springer-Verlag, New York, p. 37-66.

Wong, Nancy, Aric Rindfleisch, and James E. Burroughs (2003), "Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale," *Journal of Consumer Research*, 3(1), 72-91.

Y

Yu, C.Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Doctoral dissertation, University of California, Los Angeles.

Z

Zeithaml, Valarie A., Leonard L. Berry, and A. Parasuraman (1996), "The behavioral consequences of service quality," *Journal of Marketing*, 60 (April), 31-46.

Zimmerman, Donald W., Bruno D. Zumbo, and Richard H. Williams (2003), "Bias in estimation and hypothesis testing of correlation," *Psicologica*, 24, 133-158.