The impact of situational and dispositional variables on response styles with respect to attitude measures

Elke F.K. Cabooter

2010

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

Supervisors:
Prof. dr. Maggie Geuens, Prof. dr. Bert Weijters, Prof. dr. Iris Vermeir

Doctoral Exam Committee

Prof. dr. Maggie Geuens
(Universiteit Gent)

Prof. dr. Bert Weijters
(Vlerick Leuven Gent Management School)

Prof. dr. Iris Vermeir
(Hogeschool Gent & Universiteit Gent)

Prof. dr. Hester van Herk
(Vrije Universiteit Amsterdam)

Prof. dr. Alain De Beuckelaer
(Radboud Universiteit Nijmegen, Universiteit Gent)

Prof. dr. Mario Pandelaere
(Universiteit Gent)

Dean Prof. dr. Marc Declercq
(Universiteit Gent)

Prof. dr. Patrick Van Kenhove
(Universiteit Gent)

Acknowledgements

Dankwoord

Hier heb ik ontzettend naar uitgekeken: het neerpennen van mijn dankwoord. Ik herinner me de dag waarop alles begon als was het gisteren: een bureau met twee mannelijke collega's (waarvan slechts 1 fysiek aanwezig was) en daarop, blinkend en wel, het 294 pagina's tellende doctoraat van Bert, klaar om verslonden te worden. Daar, op dat moment, startte mijn rit op de rollercoaster van het academisch leven. Met enkele dieptepunten, maar minstens evenveel stijle hoogtepunten. Nu loopt die rit op z'n eind.

Net zoals de rit op een heuse rollercoaster op zich slechts 'half the fun' is (gillen in groep schijnt zelfs therapeutisch te werken), werk je ook aan een doctoraat liever niet alleen. Drie schitterende promotoren had ik. Al het goede komt in drievoud, stelt een spreekwoord (denk maar aan de 3 biggetjes die de wolf te slim af waren, de 3 wijzen die de pasgeboren Jezus kwamen eren, de heilige drievuldigheid,…), wie ben ik om dat te weerleggen. Bij deze wil ik hen dan ook uitvoerig danken.

Eerst en vooral: prof. Dr. Maggie Geuens. Niet in het minst voor de kans die ik van haar kreeg om een doctoraat te schrijven. Ik had de universiteit immers al verlaten en werkte al enkele jaren voor het marktonderzoekbureau Censydiam. Het werk kon absoluut boeien, maar toch bleef de droom om te doctoreren. Dankzij jou, Maggie, komt die droom nu uit. Je was daarnaast ook steeds beschikbaar om grote en kleine problemen van de baan te helpen, alsook om me advies te geven tijdens de diverse fasen van mijn doctoraat. Je geloofde volop in mijn capaciteiten en je steunde me steeds onvoorwaardelijk, ook toen het eens wat minder ging. Je bleek een steun en toeverlaat, ondanks een drukke agenda. Bedankt.

Voorts wil ik de schijnwerpers ook even op mijn copromotoren, prof. Dr. Bert Weijters en prof. Dr. Iris Vermeir, richten. Bert, bedankt voor jouw deskundig advies, voor de vele hulp en bovenal voor het geduld dat je wist op te brengen bij het aanleren van nieuwe meettechnieken. Je ontwarde m'n knopen en wist me ongelooflijk veel te leren. Iris, bedankt voor de talloze inzichten, het uitvoerig nalezen van mijn artikels en de oprechte steun tijdens donkere dagen. Je bleek een bron van aanmoedigingen en ik kon bij jou ook steeds terecht voor een goede babbel.
Bedankt alle drie. Bedankt voor de adviezen, de ontelbare verijkende discussies en voor de ondersteuning die jullie boden tijdens het gehele doctoraatsproject. Ik heb de voorbije jaren veel van jullie opgestoken, en, even schrikken wellicht, jullie hebben mij mee gevormd tot de onderzoeker die ik nu ben. Ik zie jullie dan ook als mijn professionele mentors en hoop nog vele jaren met jullie te mogen samenwerken. Al het goede komt in drievoud. Driemaal bedankt.

Ook de leden van mijn doctoraatsjury zou ik willen bedanken. Dit voor hun constructieve feedback en diverse waardevolle suggesties. Prof. Dr. Mario Pandelaere, Prof. Dr. Hester van Herk en Prof. Dr. Alain De Beuckelaer, jullie opmerkingen stelden mij in staat om dit doctoraat verder te optimaliseren.

Vervolgens wil ik de mede-auteurs van de verschillende artikelen bedanken: Prof. Dr. Niels Schillewaert, Prof. Dr. Luk Warlop en Leen Adams. Hartelijk dank voor jullie hulp, suggesties en opmerkingen. Het bleek prettig samenwerken.

Dit doctoraat had ook zijn levenslicht niet gezien zonder de ondersteuning van enkele instituten. Tijdens de eerste jaren voorzag de Universiteit Gent mij van een financiële

ondersteuning via het BOF (Bijzonder Onderzoeksbursaal). Het laatste jaar zag ik mij verzekerd van de noodzakelijke financiële middelen via de vakgroep Marketing. Daarvoor o.m. mijn uitdrukkelijke dank aan de vakgroepvoorzitter, Prof. Dr. Patrick Van Kenhove.

Ook mijn collega's wil ik bedanken, niet in het minst voor de vele leuke momenten samen. Er heerst zowat altijd een aangename sfeer op de vakgroep marketing en dat maakt het (academische) leven zoveel lichter. Speciale dank gaat uit naar mijn initiële bureaugenoten, Dries en Koen, mijn huidige bureaugenoten Anneleen, Griet en Tina, en naar Hendrik, Bart en Karin.

Dries en Koen, samen hebben we mijn eerste jaar doorgemaakt. We hadden plezier, maar wisten ook hard te werken. De eerste wetenschappelijke artikels kregen zo stilaan vorm. Door de expansie van de vakgroep zitten we niet langer samen op bureau, toch zijn jullie nog steeds "mijn mannen" en dat zal nog lang zo blijven.

Anneleen, ik ben aan de eindstreep; voor jou is die niet veraf. Australië is in zicht. We deelden lach en traan. Ik kan nu wel met zekerheid zeggen dat jij een vriendin bent voor het leven. Niet enkel op het werk, maar ook erbuiten, kan ik zeven op zeven, vierentwintig op vierentwintig, bij je terecht. Weet, dat dat ook voor jou geldt. Ik wil je bedanken voor de vele werk-gerelateerde discussies die we hadden en die mijn doctoraat significant beter maakten.

Doctor Griet, het was leuk met jou samen op de bureau te zitten en ook ernaast ben je een goeie vriendin. Een doctoraat schrijven is niet altijd even gemakkelijk, maar weten dat je niet de enige bent die nog aan de slag is op zondagavond, terwijl iedereen voor de TV zit of iets leuks doet, doet wonderen voor de moraal. Ook buiten het werk beleefden we veel leuke momenten. Ik hoop dat dat zo blijft als straks onze wegen op werkgebied elkaar scheiden. Iets zegt me, dat dat zo zal zijn.

Tina, je kwam net op de unief werken in de laatste maanden van mijn doctoraat en … nou … in die periode was ik wellicht niet steeds de meest aangename collega, maar daar komt straks verandering in. Ook al wogen die laatste loodjes toch echt wel het zwaarst, jouw goedlachse persoonlijkheid maakte ze aangenamer.

Hendrik, we waren collega's bij Censydiam tot jij besloot te doctoreren. Je liet me verweesd (nou ja) achter. Jij was het echter, die mij de tip gaf in te gaan op een aanbod van Maggie. Toen je nog geen jaar later het Brusselse verliet om voortaan in de universitaire hoofdstad, van België, Gent, te werken, waren de collega's van weleer opnieuw verenigd. Altijd was je bereid mij te helpen, bedankt.

Bart, van thesisbegeleider over collega tot vriend, een heel traject hebben we samen afgelegd en hopelijk staat er ons nog een erg lang traject samen te wachten. Heel wat treinen werden gemist, vele uren vlogen pijlsnel voorbij. We konden altijd bij elkaar terecht voor een broodnodige opkikker of simpelweg een babbel die onze gedachten even deed afglijden van wat die dag op het programma stond.

Karin, ook jij verdient een extra dankwoord. Op het tweede verdiep liep ik de deur iets vaker plat dan op het derde, maar van onze korte babbels heb ik altijd genoten. Op administratief vlak, zorgde je ervoor dat alles vlot en haast als vanzelfsprekend verliep. Je verlichtte m'n werk zo aanzienlijk.

Bedankt Isabel, Prof. Dr. Dirk, Leen, Anita, Kristof C., Tine F., Jonathan, Tine DB., Sofie, Jonas, Phillipe, Maarten, Katrien, Christophe, Willem, Evelien, Michel, Raminta, Elisa, Genevieve, Nele en alle anderen die zijn gekomen en, letterlijk dan, gegaan.

Vrienden, of je ze nu alle dagen ziet of maar heel af en toe eens hoort, ze zorgen ervoor dat je zaken waardevol acht om ze te doen. Ik neem aan dat jullie niet altijd volledig goed begrepen

waar ik nu wel of niet de voorbije jaren mee bezig ben geweest. Toch vroegen jullie me steeds ‘en, Elke, hoe gaat het met je doctoraat?’ en luisterden jullie geïnteresseerd naar wat ik te vertellen had. Ik zou jullie hier één voor één allemaal persoonlijk kunnen bedanken, maar dat zou van mijn doctoraat een 1000 pagina’s dik boek maken. Toch een woordje van dank voor Emilie voor de hulp bij het maken van die appendix en Jelle voor het programmeren van enkele studies.

Papa, mama en grote broer, ik wil jullie graag danken omdat jullie er steeds voor me waren en omdat ik altijd op jullie kan rekenen. Ik ben bij momenten, wanneer het wat minder ging, niet altijd de liefhebbende dochter/zus geweest, toch hielpen jullie mij er steeds weer bovenop. Bedankt voor de kansen, de steun en de liefde.

The future belongs to those who believe in the beauty of their dreams.

Table of Contents

Samenvatting

De kwaliteit van de antwoorden op vragenlijsten blijft een belangrijke uitdaging in marketing onderzoek. Antwoordstijlen vormen een belangrijke bedreiging voor deze kwaliteit omdat ze data vertekenen en daardoor de correcte interpretatie van resultaten bemoeilijken. Ondanks hun nadelige gevolgen en de beschikbaarheid van diverse correctiemethoden; wordt er vaak niet gecontroleerd voor antwoordstijlen. Dit is ondermeer te wijten aan de complexiteit van de correctiemethoden en de gebrekkige kennis betreffende de oorzaken van de antwoordstijlen.

Twee mogelijke antecedenten van de antwoordstijlen die in deze dissertatie worden onderzocht zijn: situationele variabelen enerzijds en persoonlijkheidskenmerken anderzijds. Een beter begrip van deze antecedenten laat toe om een systematische meetfout te minimaliseren door een aangepast onderzoeksopzet te kiezen. Daarom werden vier empirische studies uitgevoerd. De drie eerste studies concentreren zich op de situationele antecedenten, een vierde studie focust op de persoonlijkheid als antecedent.

Een eerste studie onderzoekt het effect van de schaalkarakteristieken, zoals het aantal antwoordcategorieën en het labelen van deze antwoordcategorieën, op de antwoordstijlen. Deze studie is belangrijk, aangezien ze resulteert in concrete richtlijnen voor onderzoekers bij de keuze van de schaalformats. Uit de studie blijkt dat een 5-punten schaal met enkel gelabelde uitersten, het beste presteert in het minimaliseren van de antwoordstijlen in lineaire relaties.

Een tweede studie gaat dieper in op de dimensies van de schaalformat zelf (in plaats van op de schaalkarakteristieken). Schaalformats verschillen hoofdzakelijk op twee dimensies,

namelijk polariteit (unipolair of bipolair) en de aard van de ankerpunten (enkel positieve ankers of negatieve en positieve ankers). Deze studie onderzoekt bijgevolg welke schaalformat het best presteert in het minimaliseren van de antwoordstijlen. Dit laat ons toe om aanbevelingen te formuleren betreffende de optimale schaalformat. De unipolaire schaal met positieve ankerpunten resulteert in betere validiteit dan de bipolaire schalen en de unipolaire schaal met positieve en negatieve ankerpunten. Bijgevolg toont deze studie de superioriteit van Likert schalen empirisch aan.

Studie 3 onderzoekt het effect van de cognitieve belasting op net acquiescence antwoordstijl (of met andere woorden, de tendens om meer positieve dan negatieve antwoorden te geven). Deze studie is van belang; aangezien cognitieve belasting aanwezig is in veel onderzoeksituaties; bijvoorbeeld wanneer men een vragenlijst invult, terwijl men televisie kijkt.
In twee studies blijkt dat respondenten onder een hoge cognitieve belasting positiever antwoorden op een serie van heterogene items dan respondenten die de cognitieve belasting als matig tot laag ervaren. Daarenboven geven deze studies het belang aan van subjectieve belasting (in tegenstelling tot objectieve belasting).

De laatste studie onderzoekt het verband tussen Self-regulatory focus en de tendens om onevenredig veel de extreme opties of de middelpunt optie te kiezen. De resultaten tonen aan dat de promotie georiënteerde respondenten vaker de extreme optie aanvinken; terwijl de preventie georiënteerde respondenten vaker het middelpunt aanduiden. Dit artikel benadrukt daarenboven het belang van het gebruik van testen in het meten van de persoonlijkheid en het vermijden van het hanteren van schaaltechnieken voor het meten van diezelfde persoonlijkheid.

Hoewel een deel van de variantie in antwoordstijlen onverklaard blijft, draagt deze dissertatie bij tot een beter inzicht in deze antwoordstijlen. Verschillende stappen kunnen ondernomen worden om de impact van de antwoordstijlen te minimaliseren. Zo is het op basis van de resultaten uit de eerste twee studies beter om een 5-punten Likert schaal te gebruiken. Wanneer er echter geen omgekeerde items aanwezig zijn, kan ook een 7-punten Likert schaal gebruikt worden. Studie 3 toont aan dat men de datacollectie het best uitvoert in een testruimte waar gecontroleerd kan worden op allerlei invloeden. De laatste studie duidt dan weer op het belang van een goed gerandomiseerde steekproef. Al deze aanbevelingen maken het voor de onderzoeker eenvoudiger om te controleren op de aanwezigheid van de antwoordstijlen.

Introduction, Situation of dissertation

"The origin of the response styles remain the mystery it has always been (Yates, Lee & Bush, 1997, p. 88)"

Chapter outline

The topic of the current dissertation, the impact of situational and dispositional variables on response styles with respect to attitude measures, is introduced. Response styles are explained and its importance is indicated. An outline of the dissertation is given.

Measurement error decomposed

Answering questions entails moving through a set of four subsequent processes (Tourangeau, Rips & Rasinski 1984). These processes are (1) comprehension – interpreting the question and assigning meaning to the question, (2) retrieval – recalling relevant information, (3) judgment – combining or adding the items that have been retrieved and (4) selecting and reporting a response – mapping the judgment onto the response category. Researchers often assume that respondents carefully go through each of the four response process phases to reflect their true opinion on questionnaire items. Unfortunately, this is often not the case. For instance, respondents often truncate some of the response processes or carry them out sloppily. Consequently, the observed response is not always a reflection of one's true opinion, due to the presence of measurement errors (Paulhus, 1991; Baumgartner & Steenkamp, 2006).

Measurement errors can be split into two components: random error and systematic error (De Pelsmacker & Van Kenhove, 2006). Random errors are statistical fluctuations (in either direction) in the measured data due to inherently unpredictable fluctuations in the measurement device. For instance, when a respondent mistakenly gives a strong positive answer on an item or when an interviewer registers some answers wrongly. The effect of this error type is typically small and generally accounted for by using multi-item scales or by averaging over a large number of observations (Churchill, 1979). Important is that random error does not have any consistent effect across the entire sample. Instead, it pushes observed scores up or down randomly. This means that if we could see all of the random errors in a distribution they would sum to 0 - there would be as many negative errors as positive ones. The important property of random error is that it adds variability to

the data but does not affect average performance of the group. Because of this, random error is sometimes considered noise.

Systematic errors, on the other hand, are predictable and typically constant or proportional to the true value. For example, when there is a mistake in the calculation of a variable. So unlike random error, systematic errors tend to be consistent and are therefore considered to bias measurement data. They can affect estimates of the means of observed variables in a given sample, across samples or over time (through biasing the intercept or slope) (Podsakoff et al., 2003). For instance, disturbing traffic noises in the vicinity of a test room can affect the answers of all respondents in the room. Not only the means of observed variables, but also the estimates of relationships can be affected through systematic error (Greenleaf, 1992a; Baumgartner & Steenkamp, 2006). Fortunately, if the cause of the systematic error can be identified, it usually can be eliminated.

Systematic error can be further divided into (1) content related systematic error, response sets, and (2) non-content related error, response styles (Rorer, 1965). Response sets (e.g. social desirable responding) are defined as the reflection of an exaggerated but honestly held self-view – an unconscious tendency to claim positive attributes and deny negative ones (self-deceptive enhancement) or the conscious desire to project a favorable self-image (impression management) (Paulhus, 1991). Response styles, on the other hand, refer to a tendency to select some response category a disproportional amount of the time independently of the item content (Paulhus, 1991). So, in contrast with response sets, response styles occur irrespective of item content. In addition, response styles are not limited to specific content domains, such as alcohol abuse, drug usage and other socially

sensitive variables (Mick, 1996). This dissertation focuses on systematic non-content related error or response styles only.

General Objective

Although the problem of response styles is well known in the literature and several methods have been suggested to correct for these biases (Greenleaf, 1992b; Baumgartner & Steenkamp, 2001; Podsakoff et al., 2003; De Jong, Steenkamp & Fox, 2008), at present these correction methods are hardly used because they are rather complex (see Augustin & Singh, 2005 for a notable exception). So, despite their biasing effects, response styles usually are not corrected for. However, instead of putting all efforts in correcting for response styles, it may be more fruitful to avoid response styles. This dissertation follows the latter perspective. Therefore, the main objective is to investigate when and for which type of respondents, response styles are most likely to occur and which research methods and research settings help to limit the presence of response styles.

Indeed, since it is unrealistic to assume that measurements can be completely free of error, a researcher may want to minimize measurement bias as much as possible. Therefore, gaining insights in when and for which type of respondents response styles are most problematic, is an important step toward the avoidance of systematic measurement bias in the future.

In line with this, Baumgartner & Steenkamp (2001) argue that response styles can be caused by either situational or dispositional variables (or a combination of both). Situational determinants explain response styles through task characteristics or situational

influences, whereas dispositional variables link stylistic responding to characteristics of the respondent (Baumgartner & Steenkamp 2001). This dissertation taps into both of these factors. We investigate, for example task characteristics such as scale format used in surveys and the cognitive load imposed upon respondents. As for dispositional variables, we look at the effect on response styles of respondents' self regulatory focus, a motivational factor that has gained a lot of attention in recent academic research (Pham & Higgins, 2005).

According to Belk (1975), situational influences can be split up into five subcategories: physical shape of the situation (e.g., noise, location), social shape of the situation (e.g., interviewer, influence of others), temporal perspective, task definition (e.g., buying for a friend or for yourself) and former state (e.g., mental fatigue). Although there are several subcategories, we mainly focus in this dissertation on the physical shape and the temporal perspective of the situation (e.g., time pressure, dual tasking).

Outline of the dissertation

We first focus on the concept and source of response styles (Chapter I). Chapter II till V present our empirical studies (see Figure 1). Whereas chapters' IIa/IIb and III study situational explanations of response styles, chapter IV focuses on dispositional explanations.

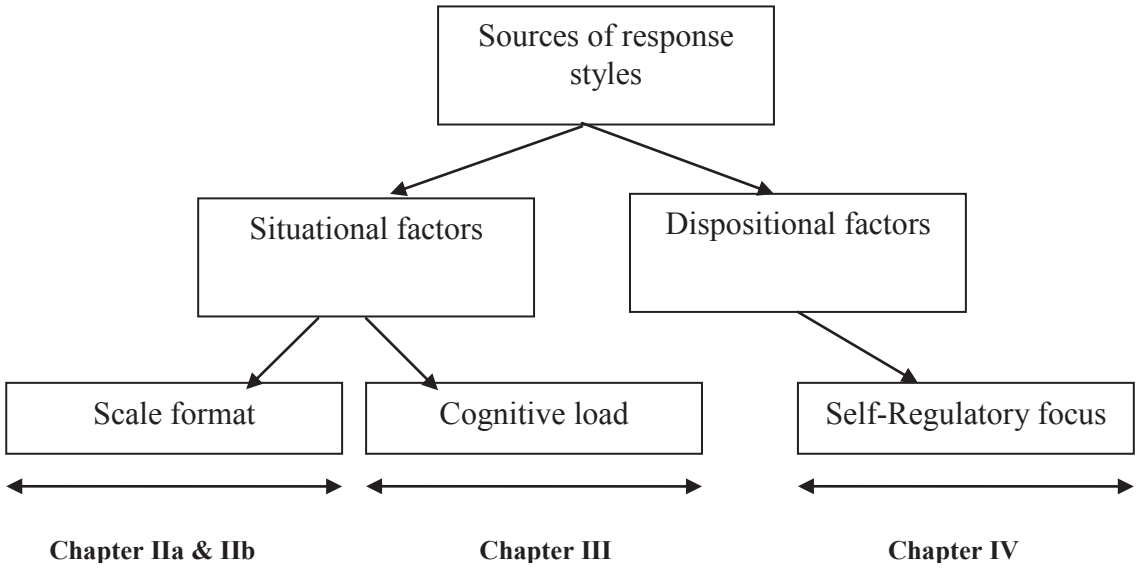


FIGURE 1: OVERVIEW OF FACTORS EXPLAINING RESPONSE STYLES

- Chapter IIa: *The effect of rating scale format on response styles: The number of response categories and response category labels.* This study investigates how scale formats, broken down into two major components: the number of response categories offered, including the choice for an odd or even number of categories, and the labeling of response categories, affect response styles and misresponse (responding positively or negatively on both an item and its reversal). This study is important as it offers concrete guidelines on which number of response categories and what type of labeling to use to minimize response styles. Results from two studies show that a 5-point endpoint labeled scale performs best in minimizing response styles in linear relations.

- Chapter IIb: *Who said that looks do not matter? The effects of scale format on response styles.* This study focuses on the rating scale format, whereas chapter IIa focuses more on the format of the response options. Scale formats basically differ on

two major dimensions, namely Polarity (unipolar versus bipolar) and Anchoring (only positive numbers or negative and positive numbers). Consequently, this study investigates which scale format performs best in minimizing response styles. The study will allow us to formulate recommendations on the choice of an optimal scale format. Results show superiority of the unipolar scale format with positive anchors and provide as such empirical validation of the Likert scale. However, the choice of a scale format should also be based on the researchers knowledge about the relevant distribution of the sample and upon the match between question interpretation and researcher interest.

- Chapter III: *The effect of cognitive load on yeah-saying and nay-saying.* This study investigates the effect of cognitive load on response styles. This article is important as cognitive load is often present in many situations. In two studies we show that respondents under high cognitive load respond more positively on a set of heterogeneous items than respondents who perceived the cognitive load as moderate or low. In addition, the importance of subjective load is stressed above objective load.

- Chapter IV: *How Self-Regulatory Focus Shapes Item Responses Regardless of Content*. This study investigates the link between Self-regulatory focus and the Extreme and Midpoint response styles. Since the individual antecedents of response styles have proven to be elusive, this study is the first to provide empirical evidence for a central link between personality and response styles. More specifically, results show that promotion focused respondents show higher levels of ERS, whereas prevention focused respondents show higher levels of MRS. In addition, this article stresses the importance of the avoidance of rating scales to measure personality traits in the quest for antecedents of response styles and shows that the alternative is the use of tests.

Finally, Chapter V concludes with a general discussion and future research avenues. All chapters can be read in isolation. This implies that some information will be repeated, although we will try to restrict replications to a minimum.

Note: what this dissertation is not about

This current dissertation does not focus on how to eliminate response styles from data as several post-data elimination techniques already exist (Greenleaf, 1992b; Baumgartner & Steenkamp, 2001; Podsakoff et al., 2003; Wong, Rindfleisch & Burroughs, 2003; De Jong et al., 2008; Van Rosmalen, van Herk & Groenen, 2010).

Chapter I

Definition, sources, measures and consequences of response styles

I.1 Definition

A response style can be defined as a person's tendency to respond systematically to questionnaire items on some basis other than what the items were specifically designed to measure (Paulhus, 1991). Of all response styles identified in literature, the following are identified as most important (Baumgartner & Steenkamp, 2001):

- Acquiescence response style (ARS), or the tendency to more frequently select the positive scale options regardless of content (e.g. Paulhus, 1991; Winkler, Kanouse & Ware, 1982; Baumgartner & Steenkamp, 2001). For example, response options 5, 6 and 7, multiplied by their weight, on a 7-point scale where 1 means 'strongly disagree' and 7 'strongly agree'and where the items of the scale are positively worded (see formulas on page 17)[1].
- Disacquiescence response styles (DARS, being the tendency to select negative scale options regardless of content) (Stening & Everett, 1984). For example, response options 1, 2 and 3, multiplied by their weight, on a 7-point scale going from 1 'strongly disagree' to 7 'strongly agree' and where the items of the scale are

[1] In many countries and for the majority of the scales in the Handbook of Marketing scales (Baerden & Netemeyer, 1999) and the Marketing scales handbook (Bruner, James & Hensel, 2001) the position of the item 'strongly agree' is at the right end of the scale. However, in some countries (e.g., Germany) and in some studies (Tourangeau, Couper & Conrad, 2004), the position of the label 'strongly agree' is on the left end side of the scale. In these cases, ARS would be measured by multiplying 1, 2 and 3 by their respective weight and DARS by multiplying 7, 6, 5 by their respective weight.

positively worded. Several researchers have not made a distinction between ARS and DARS, considering those response styles as opposites (for instance Cronbach, 1942; Harzing, 2006).

- Net acquiescence response style (NARS, being the tendency to show greater ARS than DARS).
- Extreme response style (ERS, or the tendency to select the extreme scale options). For example, options 1 and 7 on a 7-point scale. Bachman and O'Malley (1984) indicated that respondents that answer extreme positive also answer extreme negative. As a result, ERS is conceptually not different from to response range (the tendency to use a wide or narrow range of response intervals about the individual's mean response), but in practice often correlates highly with this response style (Bachman & O'Malley, 1984; Greenleaf, 1992b). Therefore, we will only focus on ERS in this dissertation.
- Midpoint response style (MRS, or the tendency to make disproportionate use of the midpoint of a scale). For example, option 4 on a 7-point scale.

I.2 Sources of response styles

Several authors have argued that response styles are stable individual characteristics (e.g., Billiet & Davidov, 2008; Hamilton, 1968; Messick, 1968; Weijters, Geuens & Schillewaert, 2010). However, most of the variance of the response styles remains unexplained, and it is still unclear how response styles are exactly related to situational determinants or dispositional determinants (e.g., demographics, culture and personality traits). In addition, most of the research on antecedents of response styles has been criticized on at least two important grounds (Baumgartner & Steenkamp, 2001; Hamilton,

1968). First, former research often did not explain the underlying mechanism of their findings or provide a clear theoretical rationale. Second, previous research mainly measured response styles on the basis of the same items that measured theoretically relevant constructs. However, if style and content are confounded then conclusions about the contamination of the scale scores and correlations between scales will be exaggerated. Therefore, several authors (e.g., Baumgartner and Steenkamp, 2001; De Beuckelaer, Weijters and Rutten, 2009) have recommended the use of a random set of items to measure response styles when the aim of the study is to establish relationships between response styles and antecedents (see also I.3. 'measurement of response styles').

I.2.1 Situational sources: scale content and involvement

ARS has been found to be more prevalent for items that are ambiguous, vague, or neutral in desirability (Peabody, 1966; Messick, 1968), for items with an extreme and enthusiastic tone (Couch & Keniston, 1960), and for issues where the respondent is uncertain (i.e., when one lacks knowledge about the item content) (Paulhus, 1991). So, item form, such as the direction and tone of its phrasing, may stimulate the respondent to answer in a habitual stylistic way. Hui and Triandis (1989) suggest that difficulty in mapping subjective scale values onto the available number of scale options, for instance with scale formats with less scale points, lead to higher levels of ERS. However, Grimm and Church (1999) could not find support for Hui and Triandis (1989) findings. Albaum and Murphy (1988) encourage the use of a two-stage scale format, where a respondent first has to indicate whether he/she agrees or disagrees with an item and in a second stage has to decide upon the intensity of the agreement/disagreement, as a one-stage versus a two-stage scale format led to higher levels of ERS. Arce-Ferrer (2006), however, did not find any differences between the two rating scale formats.

Earlier research on response styles in relation to scale formats has mainly focused on the content of questionnaire items. In this dissertation, we will instead focus on the format of the scale (e.g., number of response options, even or odd response options, fully- or extreme labeled response options, polarity of the scale, anchoring of the scale).

Finally, Hui & Triandis (1989) found that involvement could encourage ERS. When respondents are involved with the subject of the questionnaire, they may want to give more outspoken opinions and may consequently select more frequently the extreme options. However, Hui and Triandis (1989) did not disentangle content and style when measuring ERS. As a result, the link between involvement and extreme responding may be a reflection of the respondents' true extreme response as well.

I.2.2 Dispositional sources

I.2.2.1 Demographics

Gender, age and education seem to have an impact on response styles (Greenleaf, 1992b; Krosnick & Fabrigar, 2003). However, earlier findings between demographics and response styles are not that straightforward. Regarding education, earlier research found that higher educated people responded less in terms of ERS (Greenleaf, 1992b; Marín, Gamba & Marín, 1992; Weijters, Geuens & Schillewaert, 2010), less in terms of ARS (Narayan & Krosnick, 1996; Weijters et al., 2010; Winkler, Kanouse & Ware, 1982) and less in terms of MRS (Weijters et al., 2010).

With respect to gender, Hamilton (1968), Eid & Rauber (2000) and De jong et al. (2008), found that females answer more in terms of ERS than males. However, other studies (Light, Zax & Gardiner, 1965; Bachman & O'Malley, 1984; Marín et al., 1992; Grimm &

Church, 1999) show no significant difference between females and males in terms of ERS. Similarly, no significant ARS differences in gender were found in the studies of Marín et al. (1992), Grimm & Church (1999) and Johnson et al. (2005), whereas Greenleaf (1992a) and Ross & Mirowsky (1984) showed that females have lower levels of ARS and Weijters et al. (2010) showed the opposite.

Several experiments show age differences in ARS and ERS, but also here conflicting results exist. Concerning ARS, Winkler et al. (1982) found less ARS among older respondents whereas Ross & Mirowsky (1984); Greenleaf (1992a) and Weijters et al. (2010) found more ARS among older respondents and Johnson et al. (2005) found no significant age differences in ARS. Also concerning ERS, no effect (De Jong et al., 2008; Johnson et al., 2005), more ERS when older (Greenleaf, 1992b; Hamilton, 1968; Ross & Mirowsky, (1984); Weijters et al., 2010), and less ERS when older (Light et al., 1965) has been reported. In terms of MRS, Weijters et al. (2010) found a positive relation between age and MRS.

The relationship between age, gender, education and response styles will be further elaborated upon in chapter 4b. Although some effects are quite robust, the explained variance is modest (R squares below 10%) (Weijters et al., 2010). This indicates that there is a substantial component of stable response style variance that is not explained by demographic variables.

I.2.2.2 Culture

The link between culture and response styles has been investigated by several authors (e.g. Baumgartner & Steenkamp, 2001; Chun, Campbell & Yoo, 1974; Lee & Green, 1991). The response styles ARS, ERS and MRS are not only an important threat to the validity of domestic survey-based research, but also for cross-cultural research (Cheung & Rensvold, 2000; De Jong et al., 2008).

Earlier research found differences in ARS and ERS across countries, for instance for ERS more ERS have been found with respondents in the U.S. compared to Korea (Chun & Campbell, 1974; Lee & Green, 1991; Riordan & Vandenberg, 1994), Japan (Stening & Everett, 1984; Zax & Takahashi, 1967; Chen, Lee & Stevenson, 1995), and Taiwan (Chen, Lee & Stevenson, 1995). However, Stening and Everett (1984) found that Indonesian and Malaysian noncollege graduates displayed more extreme scoring than American respondents. Between Northern and Southern European countries, van Herk, Poortinga and Verhallen (2004), and Harzing (2006) found the highest levels of ERS for Greek respondents. Spanish and Italian respondents also had consistently higher scores than British, German and French respondents. Higher levels of ERS where also found for French compared to Australian respondents (Clarke III, 2000) and for Australian children compared to Chinese, Nigerian, Nepalese and Philippine respondents (Watkins & Cheung, 1995).

In cross-cultural and cultural research, also ARS differences have been found (Cunningham, Cunningham & Green, 1977; Grimm & Church, 1999; Bachman & O'Malley, 1984; Hui & Triandis, 1989; Ross & Mirowsky, 1984; Marín, Gamba & Marín, 1992; Baumgartner & Steenkamp, 1998; van Herk et al., 2004; Harzing, 2006). ARS

differences have been explored between Afro-American and European American respondents (Bachman & O'Malley, 1984), and in groups of Hispanic and non-Hispanic Americans (Hui & Triandis, 1989; Marín, Gamba & Marín, 1992). In these studies, the European American respondents tended to display less often ARS. Ross and Mirowksy (1984) found that Mexicans from Mexico showed more ARS compared to Mexicans from the US and Americans. Watkins and Cheung (1995) reported less acquiescence for children in Australia compared to children in China, Nepal, and the Philippines. Grimm and Church (1999) found that Philippine students respond more in terms of ARS than American students. Between European countries, Baumgartner and Steenkamp (2001), van Herk et al. (2004) and Harzing (2006), Greek respondents displayed more ARS than respondents from other EU countries. In addition, ARS was lowest for UK, Germany and France.

However, most of these findings were not theoretically founded. More recent research uses Hofstede's dimensions (i.e., individualism, uncertainty avoidance, power distance and masculinity) or the Globe dimensions of House et al. (2004) (i.e. in-group collectivisism, institutional collectivism, uncertainty avoidance, power distance and extraversion) to base their findings, but conflicting results appeared. To give an example: in a study of 26 countries, De Jong et al. (2008) found that individualistic countries, masculine countries and countries with high uncertainty avoidance respond more extreme than collectivistic countries, feminine countries and countries with low uncertainty avoidance. Across 19 countries Johnson et al. (2005) only found ERS differences for the dimensions power distance and masculity. Both dimensions are positively associated with ERS. De Jong et al. (2008) attribute these different findings to differences in ERS measures.

Compared to Johnson et al. (2005) and De Jong et al. (2008), Harzing (2006) did not found differences in terms of ERS on the Hofstede dimensions except for extraversion. Harzing (2006) demonstrated that extraverted countries showed higher levels of ERS (more specifically positive ERS). In addition, she found that the Globe's dimension of uncertainty avoidance is positively linked with ERS, however not the Hofstede dimension of uncertainty avoidance.

Concerning ARS, Hofstede's four dimensions are negatively associated with ARS (Johnson et al., 2005). The 26 countries-study of Harzing (2006), however, showed that ARS was negatively associated with Hofstede's dimensions power distance and individualism, but not the Globe dimensions. Harzing (2006) further demonstrated that extraverted countries showed higher levels ARS. In addition, she found that the Globe's dimension of uncertainty avoidance is positively linked with ARS.

Harzing (2006) also showed that MRS was negatively linked with the both Hofstede dimensions power distance and individualism.

In terms of language, Harzing (2006) demonstrated that questionnaires in native languages lead to higher levels of ERS and lower levels of MRS compared to questionnaires in non-native language, which was in this study English.

I.2.2.3 Personality variables

Couch and Keniston (1960) characterized ARS respondents as impulsive and emotional extraverts. Other research has shown that measures of acquiescence are negative related to verbal ability, logical consistency of attitudes and social taste; and positively to uncertainty

(Cronbach, 1942; Schuman & Presser, 1996; Messick, 1991). DARS respondents have opposite characteristics than ARS respondents; they have an introvert personality and want to keep control by avoiding external stimuli (Couch & Keniston, 1960).

Respondents with high need for certainty, a high level of anxiety, a high level of rigidity, a high intolerance of ambiguity, and a high level of decisiveness respond more in terms of ERS (Hamilton, 1968; Naemi, Beal & Payne, 2009). ERS is also associated with people who have ill-developed cognitive structures, e.g. in reaction to new brands (Shulman, 1973). ERS has also been found to be related with extraversion (Austin et al., 2006).

MRS has three possible causes according to Shuman and Presser (1981): evasiveness (one's desire to not reveal their true opinion), indecision (uncertainty about one's position), or indifference (disinterest in an issue).

So, there are few studies that suggest that the tendency to endorse response styles is a manifestation of personality attributes (e.g., Couch & Keniston, 1960). However, most of these studies were not entirely consistent or have been subject to much methodological criticism (Hamilton, 1968). For instance, none of the studies used heterogeneous items whereas the use of the latter is of primary importance in detecting response styles (see next paragraph). In addition, many of the studies were largely exploratory in nature and have not based their findings on theory.

I.3 Measurement of response styles

Items are often created to measure a construct that underlies the researchers or company's interest such as, for example, brand or company satisfaction, brand likeability, intention behavior, attitude towards the ad, attitude towards the brand, etcetera. However, using such homogeneous items to capture the link between, for instance, a personality trait and response styles is not recommended. By using those items, it is almost impossible to determine to what extent response styles represent stylistic tendencies of an individual or a subject's meaningful response to an item. On the other hand, the use of heterogeneous items, in other words items that are minimally correlated, (see Table 1) avoid this confounding between content and style. As a consequence, it can be expected that relations or similarities in individual's responses are mainly due to pure behavioral tendencies, i.e. response styles. This method is also advocated by Baumgartner & Steenkamp (2001), Greenleaf (1992b) and De Beuckelaer, Weijters & Rutten (2009).

TABLE 1:

EXAMPLES OF HETEROGENEOUS ITEMS

Strangers can be trusted
I like to watch a good movie
The French language is still influential
I often daydream
A woman working out of home with children is still a good mother

To create operational measures of the response styles ARS, DARS, NARS, MRS and ERS, we used the formulas of Baumgartner and Steenkamp (2001) whenever possible[2].

$ARS_{weighted}$= (#agreements(option 5*1, option 6*2, option 7*3))/ #statements, (1)
ARS= ((#agreements(option 5, option 6, option 7))/ #statements, (2)
$DARS_{weighted}$ = (# disagreements (option 1*3 , option 2*2, option 3*1))/ #statements, (3)
DARS = (# disagreements (option 1, option 2, option 3))/ #statements, (4)
NARS = (ARS) - (DARS), (5)
ERS = ((# negative extremes (option 1)) + (# positive extremes (option 7)))/ #statements, (6)
MRS = (# midpoints (option 4))/ #statements, (7)
The formulas above are adapted to a 7-point likert scale; # stands for frequency of response option

ERS has often been measured by the spread or standard deviation (Response Range or RR) of an individual's ratings across a set of heterogeneous items (Hui & Triandis, 1985; Greenleaf, 1992b). As mentioned in I.1., RR and ERS are not identical although they are highly correlated, loading on the same component (Baumgartner & Steenkamp, 2001; Diamantopoulos, Reynolds & Simintiras, 2006).

The formula of $ARS_{weighted}$ makes it more refined than ARS, because it distinguishes between cases in which someone strongly agrees, somewhat agrees, or agrees with an item. However, Baumgartner and Steenkamp (2001) compared these different measures of ARS and found that all measures substantially loaded on a single factor.

For measuring ARS, an alternative measure can be used (Baumgartner & Steenkamp, 2001; Winkler et al., 1982). This method assumes that a scale is perfectly balanced (i.e., the scale contains an equal amount of positively as well as negatively worded items). An

[2] In Chapter IIa, different measures, that are measures based on log odd, were used as the traditional measures are not scale invariant. For the traditional measures, different weights are given to each response option, however when comparing results from different scale formats (e.g., 4 versus 6 point likert scales) weights cannot be used because of the difference in number of response options. Although we used different measures, our new measures correlate highly with the traditional measures.

example of a balanced item is “I find most advertisements credible.” versus “I often feel misled by advertisements”. If a person agrees with both items simultaneously than a researcher can assume that the respondent answers stylistically rather than substantively. Although this type of scale has a built-in control for ARS, the presence of balanced scales is rare in marketing. Developing such a scale is difficult as for some items there is no logical opposite (Schuman & Presser, 1996). In addition, balanced scales only control for ARS but not for ERS or MRS. Therefore, in this dissertation, we measure response styles based on a set of heterogeneous items.

I.4 Consequences of response styles

Response styles can seriously bias research conclusions because they affect reliability[3] (Cronbach, 1946; Greenleaf, 1992a), affect validity and contaminate respondents’ answers to substantive questions (Cronbach, 1942; Alwin & Krosnick, 1991). ARS and DARS affect the central tendency of a measure (and hence the intercept in a linear relation), whereas MRS and ERS directly affect the spread in observed data (and hence the slope in a linear relation) (Baumgartner & Steenkamp, 2001; Greenleaf, 1992b).

Next to distorted mean scores, also estimates of a relationship between observed variables can be misleading when systematic measurement error is ignored. The presence of response bias in one observed variable makes it likely that this bias will correlate with bias in other observed variables. As a result, the true relationship between variables will be distorted. As an example of this problem, Baumgartner and Steenkamp (2001) found high

[3] ARS and DARS lead to inflated internal consistency estimates, if scales are unidirectional, as well as misresponse (MR, see chapter IIa). When there are reversed items present then ARS and DARS deflate internal consistency. Concerning ERS and MRS, in literature, it is not clear what the effect is on reliability.

correlations between the variables of health consciousness, consumer ethnocentrism, quality consciousness and environmental consciousness. However, when the authors corrected for stylistic responding, these correlations substantially reduced (with an average reduction of 0.23).

Other implications for researchers can be found in the clustering of data (Greenleaf, 1992a). When a clustering is performed on collected data, there could be a problem with the quality of the results. Data will as such be misinterpreted and misclassifications will be made (Greenleaf, 1992a). Consider, for example, a study that identifies respondents in the highest and lowest deciles on a measure of attitude towards a new innovative product. Through the presence of response bias, many of the respondents in the highest or lowest deciles may be classified as extreme because of their response styles, but may actually belong in more moderate segments and vice-versa.

Response styles can also affect the result of regression analyses (e.g. Chun et al., 1974; Lorr & Wunderlich, 1980; Heide & Grønhaug, 1992; Baumgartner & Steenkamp, 2001) or factor analyses, as it can lead to factors composed exclusively of negatively worded items (Lorr & Wunderlich, 1980). Moreover, as response styles are not limited to content, they can occur in every existing item scale (Bachman & O'Malley, 1984). Furthermore, it has been proven that these stylistic consistencies endure to a certain extent over time (e.g., Hamilton, 1968; Messick, 1968; Hui & Triandis, 1985; Billiet & Davidov, 2008; Weijters, Geuens & Schillewaert, 2010).

In sum, response styles seriously affect data and lead to wrong conclusions. In addition, comparability between two or more countries or between different groups cannot be

guaranteed as the corresponding measurement parameters are not equivalent, but biased by different response styles (e.g. Bachman and O'Malley, 1984). So, when there is model non-invariance, that is for instance when some indicators (i.e., weights, intercepts, variances) of a regression between an independent and dependent variable for two groups differ, one can interpret this as manifestations of response styles (Cheung & Rensvold, 2000).

Therefore research in antecedents of response styles is of importance since these antecedents can be used ad hoc to minimize the effect of response styles.

Chapter IIa

The effect of rating scale format on response styles: The number of response categories and response category labels

IIa.1 Chapter outline

Questionnaires using Likert-type rating scales are an important source of data in marketing research. Researchers use different rating scale formats with varying number of response categories and varying label formats (e.g., seven point rating scales labeled at the endpoints, fully labeled five point scales…), but have few guidelines when selecting a specific format. Drawing from the response style literature, we formulate hypotheses on the effect of the labeling of response categories and the number of response categories on net acquiescence response style, extreme response style and misresponse to reversed items. We test the hypotheses in an online survey (N=1,207) with eight experimental conditions and a follow-up study with two experimental conditions (N = 226). We find evidence of strong effects of scale format on response distributions and misresponse to reversed items and formulate recommendations on scale format choice.

IIa.2 Introduction

A lot of what we know about consumers is based on questionnaire data. When creating questionnaires, researchers face several design-related choices. One such choice concerns the format of rating scales used to administer Likert items (e.g., a five point rating scale where 1 = 'strongly disagree' and 5 = 'strongly agree'). The choice for a particular rating scale format can be broken down into two major components: the number of response categories to be offered, including the choice for an odd or even number of categories, and the labeling of response categories. A lot of variation exists in the Likert formats used to administer marketing scales. Commonly used formats include those with 5, 6 or 7 categories, either fully labeled (i.e., all response categories are explicitly labeled) or labeled at the extremes (e.g., labeling the first category with 'strongly disagree' and the last category with 'strongly agree') (Bearden & Netemeyer, 1999; Bruner, James & Hensel, 2001). Table 1 provides an overview of formats that are regularly used in marketing research, based on an analysis of the scale formats used in the marketing scale inventory by Bruner et al. (2001) and research published in the International Journal of Research in Marketing, Journal of Consumer Research, and Journal of Marketing Research between 2004 and 2009.

TABLE 1:
OVERVIEW OF SCALE FORMATS USED IN QUESTIONNAIRES

	Bruner, James and Hensel 2001 (N = 603)		IJRM (N = 82)		JCR (N = 381)		JMR (N = 201)	
Number of response categories	Extremes labeled	All labeled	Extremes labeled	All labeled	Extremes labeled	All labeled	Extremes labeled	All labeled
< 4	0.5%	1.2%	6.1%	.0%	2%	0.2%	2.5%	.0%
4	0.8%	.8%	3.7%	.0%	1.1%	0.6%	1.5%	2%
5	30.0%	2.7%	19.5%	2.4%	12.2%	0.2%	13.9%	.0%
6	2.0%	.0%	2.4%	.0%	1%	.0%	1.5%	.0%
7	55.2%	.2%	43.9%	.0%	43%	.0%	42.8%	.8%
> 7	6.6%	0.0%	22.0%	.0%	39.8%	.0%	35.9%	.0%

Self-report measurement quality remains an ongoing concern (e.g., Rossiter, 2002; Sharma & Weathers, 2003; Strizhakova, Coulter & Price, 2008), but the choice for a specific format appears to receive relatively little attention in marketing research. Yet, response scale format might affect the quality of questionnaire data. Greenleaf (1992a, p. 187) suggested that response category labels and the number of response categories may influence the level of response bias and called for further research on the matter. Specific evidence of response bias due to scale format remains scarce in the marketing literature though (but see Weathers, Sharma & Niedrich (2005) for a notable exception). An important reason for this gap is that most response style research has focused on only a single response scale format. For example, Arce-Ferrer (2006) used 7-point Likert scales with endpoint labels; Baumgartner and Steenkamp (2001) and De Jong et al. (2008) used 5-point fully labeled Likert scales; Greenleaf (1992a) used 6-interval Likert scales with endpoint labels. As a consequence it is not clear how response styles differ across the response scale formats used in these studies. This issue is of importance as there is no complete standardization in terms of response scale formats across studies in marketing research (although two formats are dominant, the 5 and 7-point likert scale; cf. Table 1) and cross-study comparability and generalizability is at stake.

To address this issue, the current study compares some of the most commonly used response scale formats in terms of three key response biases: net acquiescence response style (NARS), extreme response style (ERS), and misresponse to reversed items (MR)[4]. We focus on NARS, ERS and MR because they bias observed means, variances and

[4] In the current article, we do not include Midpoint Response Style (e.g., Weijters, Schillewaert & Geuens, 2008) because we study the effect of including (or omitting) a midpoint.

internal consistencies of scales, three parameters that are generally of interest in marketing research.

IIa.3 Conceptual background

IIa.3.1 Response styles

The central tendency of rating scale measures is directly influenced by a directional bias called Net Acquiescence Response Style (NARS; Greenleaf, 1992a; Baumgartner & Steenkamp, 2001; Rossi, Gilula & Allenby, 2001). This response style concerns the extent to which respondents tend to show greater acquiescence (tendency to agree) rather than disacquiescence (tendency to disagree) with items, irrespective of content. Extreme response style (ERS) is defined as the tendency to disproportionately use the extreme response categories in a rating scale (Greenleaf, 1992a,b; Baumgartner & Steenkamp, 2001). ERS affects the spread in observed data (Baumgartner & Steenkamp, 2001; Greenleaf, 1992a; Rossi, Gilula & Allenby, 2001).

To counter the effect of NARS, the use of balanced scales has been suggested (Paulhus, 1991)[5]. A balanced scale contains reversed items, i.e. items that are coded in the opposite direction of their non-reversed counterparts (e.g., 'I feel sad' would be a reversed item measuring happiness). Unfortunately, respondents often show a particular bias when responding to such items, in that they often respond in the same direction to two items that are opposite in meaning, i.e. agree to an item and its reversal or disagree to an item and its

[5] Contrary to NARS, ERS cannot be corrected for in advance (i.e., during scale construction). However, techniques have been developed to correct for response styles statistically, e.g., the procedures by Baumgartner and Steenkamp (2001) or Greenleaf (1992a), and the new improved technique to convert for ERS by De Jong et al. (2008).

reversal. This bias is labeled misresponse to reversed items (MR). A growing body of evidence indicates that MR cannot be equated with NARS (Wong, Rindfleisch & Burroughs, 2003; Swain, Weathers & Niedrich, 2008; Weijters, Geuens & Schillewaert, 2009).

IIa.3.2 Response styles and scale format

Exploratory research suggests that scale format influences response styles. For example, Hui and Triandis (1989) illustrate how different formats yield response distributions that are substantially different in shape irrespective of content. Though intriguing in many respects, previous studies on the relation between response styles and response formats are limited for one or several of the following reasons.

First, some studies use secondary data in which content and format are confounded to an unknown extent (e.g., Andrews, 1984; Alwin & Krosnick, 1991). Further, we are not aware of studies that have related different formats to a broad set of response styles that capture biases in terms of central tendency (NARS), spread (ERS), and internal consistency (MR). Finally, student samples may be inappropriate for studying response styles, as young adults of high education typically show lower levels of several response styles (Narayan & Krosnick, 1996; Greenleaf, 1992a; Marín, Gamba & Marín, 1992; Knauper, 1999; Mirowsky & Ross, 1991).

In summary, evidence on the relation between scale formats and response styles is far from conclusive. Nevertheless, there are good theoretical reasons to expect such a relation. Most response style research has focused on differences between individuals or groups of individuals (e.g., Baumgartner & Steenkamp, 2001; De Jong et al., 2008; Greenleaf, 1992a,

b; Rossi et al., 2001). There is consensus, however, that response styles are a function not only of individual characteristics but also of the stimuli, i.e. the questionnaire items and format (Baumgartner & Steenkamp, 2001; Paulhus, 1991). In previous work, researchers have made conjectures about such effects (e.g., Greenleaf, 1992a) and Arce-Ferrer (2006) recently provided evidence that the perceived meaning of response categories play a key role in response styles.

IIa.3.3 Hypothesis development

According to Tourangeau, Rips and Rasinski (2000), respondents perform a set of cognitive processes when answering questionnaire items: (1) comprehension (they attend to the question and interpret it), (2) retrieval (they generate a retrieval strategy and then retrieve relevant beliefs from memory), (3) judgment (they integrate the beliefs into a conclusive judgment), and (4) response (they map the judgment onto the available response categories and answer the question). Response style bias can occur as a result of problems during one or more of these processes (Krosnick, 1991; Swain et al., 2008). In the current study we focus on the response process because the translation of a judgment into an answer clearly depends on the response categories provided, i.e., the format of the scale (Tourangeau et al., 2000).

We construct our hypotheses around two main mechanisms through which formats affect response styles. First, different response scale formats imply differences in the perceived meaning and salience of response categories, thus changing the chance of them being selected (Arce-Ferrer, 2006; Schaeffer & Presser, 2003). Second, response scale formats vary in the extent to which they force ambivalent and indifferent or truly neutral

respondents to choose sides when responding; this has an effect on response distributions (Nowlis, Khan & Dhar, 2002).

We study the labeling of response categories and the number of response categories offered. As for labeling, we center our attention on the two most common approaches (cf. Table 1): labeling all response categories versus labeling the endpoints only (Hippler & Schwarz, 1987, p. 111). As for the number of response categories, we include the two most popular formats, i.e. 5- and 7-point scales (cf. Table 1). To assess the impact of a midpoint we also include 4 and 6-point scales in our study. Accordingly, and in line with recent methodological research in this area (Lozano, Garcia-Cueto & Muñiz, 2008), we limit the current study to scale formats using 4 through 7-points[6]. For conceptual and analytical reasons, we classify the different numbers of response categories along two orthogonal dimensions, 'midpoint inclusion' and 'gradations of (dis)agreement' as follows: 4-point scale = no midpoint, 2 gradations of (dis)agreement; 5-point scale = midpoint, 2 gradations of (dis)agreement; 6-point scale = no midpoint, 3 gradations of (dis)agreement; 7-point scale = midpoint, 3 gradations of (dis)agreement. In what follows, we formulate hypotheses concerning the effect of the scale format characteristics on NARS, ERS and MR.

IIa.3.3.1 Labeling of response categories (all or endpoints only)

Using endpoint labels without intermediary labels makes it easier to construct a rating scale as only two labels have to be formulated. Also, this format seems intuitively more in line with an interval scale assumption. On the other hand, formats with all categories

[6] We note that binary response formats may also be common, especially in (psychological) research using Item Response Theory. However, the focus of the current article is on Likert scales.

labeled facilitate interpretation both by respondent and researcher (Wildt & Mazis, 1978). A fully labeled format is also associated with higher reliability (Alwin & Krosnick, 1991; Krosnick, 1991; Weng, 2004). However, this increase in reliability may be partially due to response style bias (Greenleaf, 1992a).

When all response options are verbally labeled, the intermediate options are more salient. Respondents use the meaning of the labels that are provided to them when mapping judgments to response scales (Rohrmann, 2003; Wegner, Faulbaum & Maag, 1982; Wildt & Mazis, 1978). Salient options will attract more responses due to their increased accessibility (Posavac, Sanbonmatsu & Fazio, 1997; Posavac, Herzenstein & Sanbonmatsu, 2003) and consequently, respondents tend to be attracted to labeled points (Krosnick & Fabrigar, 1997).

Labels denoting (dis)agreement make the valence of a negative/positive response more explicit. As respondents have a desire to show agreeableness[7] (Schuman & Presser, 1996; McClendon, 1991), the clarity and salience of full labeling is likely to reinforce the felt pressure to agree. As a result, the response distribution may shift to the positive side as a result of full labeling.

H1: Labeling all response categories leads to higher levels of NARS.

In line with this, when the intermediate options become more salient through full labeling, we expect a shift towards those intermediate categories at the expense of the extreme categories (Simonson, 1989). In contrast, using verbal labels only for the endpoints attracts

[7] This sense to agree can differ across countries (see chapter I, cross cultural differences). This sense to agree is especially prevalent in Greece and for Hispanic and Afro-American respondents

respondents to the endpoint categories (Krosnick & Fabrigar, 1997). Hence, we hypothesize:

H2: Labeling all response categories leads to lower levels of ERS.

When all response categories are verbally labeled, the meaning of each response category to the respondent is less ambiguous than in situations where only end labels are provided (Lazovik & Gibson, 1984). For the latter, respondents need to figure out the meaning of the intermediate response categories to determine the option that comes closest to expressing their opinion. In doing so, respondents can attach different meanings to the same response option (Arce-Ferrer, 2006; Schaeffer & Presser, 2003; Schwarz et al., 1991). For instance, in a four point scale with end labels fully disagree/fully agree, the second option in row can get the meanings 'slightly disagree' or 'disagree' or even 'agree'. With labels for the end points only, selecting the right response option will be more challenging when respondents need to make up the right meaning for each response category (De leeuw, 1992; Krosnick, 1991). Since reversed items are in general more difficult to answer (Steenkamp & Burgess, 2002; Swain et al., 2008), this extra amount of cognitive difficulty at the response phase will increase the level of MR. Conversely, a fully labeled version enhances interpretation and facilitates response (Rohrmann, 2003); hence it will be clearer to respondents that two same direction responses to reversed items are inconsistent.

H3: Labeling all response categories leads to lower levels of MR.

IIa.3.3.2 Midpoint

The issue of whether or not to offer a midpoint has been disputed for decades (e.g., Converse & Presser, 1986; Garland, 1991; Moser & Kalton, 1972; O'Muircheartaigh et al., 2000). The major argument in favor of offering a midpoint simply states that respondents

with a truly neutral stance need to have the possibility to choose the middle option and should not be forced to choose a polar alternative (Schuman & Presser, 1996). Offering a midpoint allows respondents to indicate neutrality or ambivalence and makes people more comfortable when selecting a response option (Nunnally, 1967). Opponents argue that the midpoint is an easy way out for respondents, leaving them the possibility to avoid thinking about the issue (Converse & Presser, 1986). Following this line of reasoning, omitting the midpoint would increase data quality (Klopfer & Madden, 1980).

The midpoint attracts truly neutral/indifferent respondents (being neither positive nor negative) on the one hand, and ambivalent respondents (being both positive and negative) on the other hand (Nowlis et al., 2002). Both types of respondents will be forced to choose an option when no midpoint is offered (Schuman & Presser, 1996). Since neutral or indifferent respondents do not hold strong positive or negative evaluations, they are unlikely to experience task related distress when they are forced to choose. When no midpoint is offered, Nowlis et al. (2002) and Presser & Shuman (1980) provided evidence that neutral respondents will randomly shift their response in either direction to the closest category. For these respondents the omission of a midpoint will thus leave the distribution unaffected (Parducci, 1965; Schuman & Presser, 1980; Nowlis et al., 2002).[8]

[8] This random selection of positive and negative response options by truly neutral respondents is mainly attributed to the low activation of either positive or negative evaluations among neutral respondents. However, dependent on the construct of interest the expected random effect can disappear. For instance, if the subject of the survey is primarily answered in a positive (negative) direction, than the omission of the midpoint will probably lead neutral respondents more to the positive (negative) response options of the scale. Since, our study uses different subjects (i.e., heterogeneous items) the chance that all subjects answer in a positive (negative) direction is small.

Ambivalent respondents, on the other hand, do hold strong beliefs at both ends of the scale. For them the midpoint response is the result of their inability or unwillingness to make the required trade-offs to choose sides (Nowlis et al., 2002). According to Nowlis et al. (2002), respondents who are forced to choose sides will make use of heuristics in order to reduce the conflict. Consequently, ambivalent respondents will focus on the most important attribute of the evaluation object. This means that the direction of the distribution can be either positive or negative or remain unaffected.

However, both Velez & Ashworth (2007), and O'Muircheartaigh (2000) found a disproportional movement of negative answers to the midpoint when it was provided. This phenomenon can be explained by the negative affect induced by the task which is demonstrated by Dhar (1997). When the midpoint is omitted, the frustration for being forced to choose may bring along task-related negative affect. It is noted that these negative affective reactions to conflicting situations often produce negativity dominance, meaning that when thoughts are conflicting, negative thoughts tend to become more salient and dominant (Dhar, 1997; Rozin & Royzman, 2001; Schimmack & Colcombe, 2002). So unless evaluation objects have a dominant attribute that is positively or negatively evaluated and that can be easily used for heuristic processing, ambivalent respondents will tend to react negatively in the absence of a midpoint. Hence we hypothesize that when no midpoint is offered, ambivalent respondents (and approximately half of the indifferent respondents) will tend to express disagreement, whereas they would have selected the midpoint if it had been offered. As a consequence, we expect a higher level of NARS when a midpoint has been added because of the disproportional decrease in negative answers compared to positive answers. We also expect a decrease in ERS, because ambivalent

respondents who would have selected the extreme alternatives when the midpoint is omitted (Nowlis et al., 2002) will opt for the midpoint if it is provided.

H4: NARS increases when adding a midpoint.

H5: ERS decreases when adding a midpoint.

In case of an even numbered format, truly neutral respondents will randomly shift between positive and negative response options. They will probably do so for nonreversed items as well as reversed items related to the same topic. Consequently, there is more chance that these respondents will contribute to a higher level of MR. As stated earlier, ambivalent respondents experience negative affect in the absence of a midpoint and – consequently – tend to respond negatively. If this happens in response to both a nonreversed item and a reversed item related to the same topic, MR will result. Hence, we hypothesize:

H6: MR decreases when adding a midpoint.

Note that we expect ambivalent respondents to disagree to both an item and its reversal; we will refer to this as negative MR.

IIa.3.3.3 Gradations of (dis)agreement

Previous research has provided recommendations on the optimal number of response categories drawing from a diversity of theories. From an information theory perspective, it has been suggested that a scale range must be refined enough to allow for maximal information transmission (Cox, 1980; Garner, 1960). In this tradition, Green & Rao (1970) dismissed the use of two to three response categories, favoring the use of six or seven-point scales instead, as these formats perform well in recovering continuous latent variables.

Subject-centered research has demonstrated that respondents may not optimally use some response formats for reasons that are mainly cognitive and/or motivational in nature (Krosnick, 1991; Hippler & Schwarz, 1987; Weathers et al., 2005). Studies in the subject-centered tradition with a focus on cognitive limitations have tried to identify the optimal number of response categories based on reliability measures, often finding higher reliability with an increasing number of response alternatives (e.g., Chang, 1994; Matell & Jacoby, 1971; Preston & Colman, 2000). However, the increase in reliability might be merely due to response styles (Cronbach, 1950; Greenleaf, 1992a; Peabody, 1962; Parducci & Wedell, 1986).

From a motivational perspective, respondents want to meet expectations set by the survey situation and provide information to the researcher. The availability of extra response categories allows respondents to differentiate their responses within the range of responses that express agreement or disagreement (Krosnick, 1991). By doing so, respondents can qualify the strength of their opinion (Ghiselli, 1939). Respondents will consequently bring more variation in their answers but the valence of the answer will not change. In other words, negative answers will vary in their level of being negative but will not become positive, and positive items will vary in their level of being positive but will not become negative (Marsh & Parducci, 1978). As a result, we do not expect that an increasing number of gradations will lead to a difference in NARS or in MR as such. However, due to the higher variation in the intermediate response range, we do expect a decrease in the level of ERS (Hui & Triandis, 1989).

H7: ERS decreases when more gradations of (dis)agreement are offered.

IIa.3.3.3.1 Labeling and midpoint

When the midpoint is present, full labeling is likely to affect both NARS and ERS. The hypothesized impact of the midpoint on NARS varies according to whether respondents interpret the midpoint for what it stands, that is neutrality. When the midpoint is offered and all response options are labeled the midpoint option will be more salient which leads to a higher attraction of respondents towards the midpoint (Schaeffer & Barker, 1995). In a fully labeled scale also the intermediate options become more salient and attract respondents to those response options. These effects will reinforce the decrease in ERS. Hence, we hypothesize:

H8: Full labeling of the response categories strengthens the positive effect of offering a midpoint on NARS.

H9: Full labeling of the response categories strengthens the negative effect of offering a midpoint on ERS.

As stated earlier, when the midpoint is offered, MR will decrease since the midpoint will attract respondents who otherwise might have misreponded (Velez & Ashworth, 2007). When the scale is fully labeled, it will become more readily apparent that one is responding inconsistently to a reversed item (Rohrmann, 2003). Consequently, we hypothesize:

H10: Full labeling of the response categories strengthens the negative effect of inclusion of a midpoint on MR.

IIa.3.3.3.2 Gradations and midpoint

When a midpoint category is present, an increase in the number of gradations is likely to affect its perceived width. The provision of more intermediate categories around the midpoint reduces the size of the middle category as it stimulates respondents to express

their attitude even if their attitude is only slightly positive or negative (Weems & Onwuegbuzie, 2001; Matell & Jacoby, 1972; Presser & Shuman, 1980). Some indifferent respondents – who would normally choose the middle position – now opt for one of the nearby categories. These respondents will be randomly distributed across the negative and positive sides, leaving the level of NARS unaffected (Parducci, 1965). As discussed, adding more gradations and adding a midpoint both reduce ERS. The reason is that non-extreme options (i.e., the extra intermediate categories and the midpoint) attract respondents that might otherwise have responded extremely. Adding a midpoint will reduce the level of ERS less with the addition of intermediate options, since the amount of ERS is low in a format with more gradations, because of the salience of the extra intermediate options. As both effects draw from the same pool of otherwise extreme respondents, we expect an interaction effect:

H11: The presence of a midpoint mitigates the negative effect of adding more gradations of (dis)agreement on ERS.

The reduction in perceived width of the middle response category in scales with more gradations will probably lead to more MR. Since more respondents do make a choice, they can make processing errors and respond wrongly to a reversed item. As a result, we expect that including a midpoint does lead to a decrease of MR but this decrease will be lower when there are more response options. Hence, we hypothesize:

H12: Offering a midpoint diminishes the negative effect of adding more gradations of (dis)agreement on MR.

IIa.3.3.3.3 Gradations and labeling

As discussed, in a fully labeled scale the salience of the intermediate options results in lower levels of ERS and higher levels of NARS. For NARS we do not expect an interaction effect of labeling and gradations, since adding extra response categories does not change the valence of the answers (Marsh & Parducci, 1978). On the other hand, adding more gradations will lead to a decrease in ERS. However, this effect is likely to be different according to the degree of labeling. In a fully labeled scale we expect the decrease of ERS, due to the addition of extra response options, to be weaker when compared to an endpoint only setting. The reason is that in a fully labeled scale some of the respondents already shifted their responses towards the more salient intermediate response categories.

H13: Fully labeling scales weakens the negative effect of adding more gradations of (dis)agreement on ERS.

We do not expect that adding extra gradations has an unconditional direct effect on MR. However, we do expect such effect for scales with endpoint labels. A fully labeled scale makes all response options salient and clear for the respondent, which facilitates responding (Rohrmann, 2003). In case of an endpoint only format, we expect an increase in MR when more gradations of (dis)agreement are offered. When extra response options are added in an endpoint only setting, respondents need to put more cognitive effort in both attaching meanings to the extra response options and keeping these meanings in mind. The resulting cognitive resources limitation is likely to result in MR (Swain et al., 2008).

H14: In formats with labels for the endpoints only, adding more gradations of (dis)agreement leads to an increase in MR

IIa.4 Methodology

IIa.4.1 Empirical Study 1

IIa.4.1.1 Design

To test our hypotheses, we conducted an online survey, orthogonally manipulating the rating scale format characteristics labeling of the response categories (either only the extreme response categories were labeled or all response categories were labeled) and number of response categories (4 to 7). The 7 response category labels were the Dutch back-translated local equivalents of 'strongly disagree' ('Helemaal niet akkoord'), 'disagree' ('Niet akkoord'), 'slightly disagree' ('Eerder niet akkoord'), 'neutral' ('neutraal'), 'slightly agree' ('Eerder akkoord'), 'agree' ('Akkoord'), and 'strongly agree' ('Helemaal akkoord'). In the fully labeled conditions with only 4 or 6 categories, the midpoint category was dropped. In the fully labeled conditions with 4 and 5 categories, we also dropped the categories 'slightly agree' and 'slightly disagree'. The respondents were randomly assigned to the conditions. This resulted in the following cell counts. All labeled: 4-point (N=137), 5-point (N=153), 6-point (N=143), 7-point (N=150). Extreme categories labeled: 4-point (N=175), 5-point (N=156), 6-point (N=154), 7-point (N=139).

IIa.4.1.2 Sample

The sample was randomly drawn from all men in the panel of an Internet marketing research company in a European country, representative for local Internet users. Only men were invited to participate because of reasons not related to this study but to the questionnaire of which the current items were part. 1,207 people responded (response rate = 27%). Age ranged from 15 to 65 years with a median of 49. 42.2 % of respondents did not have any formal education after secondary school, 57.8% did.

IIa.4.1.3 Instrument

The questionnaire consisted of two parts, one designed to measure MR and an intention measure to be used for illustrative purposes, and the other part to measure NARS and ERS. The first set of questions consisted of multi-item measures for three constructs, containing both reversed and non-reversed items. A specific brand in the GPS product category was used as the study topic. We included the following three reversed item pairs to calculate the level of misresponse (Bearden & Netemeyer, 1999): (a) "Compared to other products, this product is important to me" and "I am not interested in this product"; (b) "I love this brand" and "I find this a very bad brand"; (c) "This brand is really something for me" and "In no case would I use this brand". Each item pair was used to compute an indicator of MR. Specifically, the MR score for a reversed item pair was 1 for a respondent who responded positive or negative to both items (before reverse coding the item responses), 0 otherwise (Swain et al., 2008). This operation resulted in three MR indicators, labeled a, b and c. The intention items included to illustrate the impact of response bias were "I would like to try this product," and "Next time I make a purchase in this product category, I will consider the product that was shown".

The second part of the questionnaire consisted of items that were included with the specific aim of measuring NARS and ERS. In particular, we randomly sampled 21 items from as many unrelated marketing scales in Bearden and Netemeyer (1999) and Bruner et al. (2001). Thus we made sure that the contents of these items had no substantial true correlations. This was confirmed by the low inter-item correlations, ranging from .03 to .10 across conditions. As the items were randomly sampled from existing marketing scales, they were highly heterogeneous, and 21 items could be reasonably assumed to be sufficient

to validly measure NARS and ERS (Greenleaf, 1992b; Weijters et al., 2008; De Beuckelaer, Weijters & Rutten, 2009).

To create measures of NARS and ERS we used log odds. The traditional measures of Baumgartner and Steenkamp (2001) could not be used since these measures are not format invariant, whereas our measures need to be format invariant[9]. The odds is the ratio of the probability that the event of interest occurs to the probability that it does not, often estimated by the ratio of the number of times that the event of interest occurs to the number of times that it does not (Bland & Altman, 2000). An important advantage of using odds based measures of NARS or ERS is that it facilitates interpretation and that it does not require an assumption of interval measurement level of the rating scales (which is a requirement when using means or measures that capture the deviation from the midpoint, for example). Sample odds ratios are limited at the lower end as they cannot take on negative values, but not at the upper end, resulting in a skewed distribution. The log odds ratio, however, can take any value and has an approximately normal distribution centered round zero (Bland & Altman, 2000). NARS was computed as the log odds of the number of agreements plus one over the number of disagreements plus one (the ones were added to avoid zero values):

NARS = ln ((# agreements+1) / (# disagreements+1)), with # signifying the frequency of (dis)agreements (1)

[9] For instance, in a 7-point scale the response option 7 is multiplied by 3 and the response option 6 is multiplied by 2. However, in contrast with a 7-point scale format, in a 5-point scale format the same extreme answer, response option 5, is only multiplied by 2 and the response option 4 is multiplied by 1. Therefore comparisons between scale formats cannot be made using the formulas of Baumgartner & Steenkamp (2001).

where ln indicates the natural logarithm and # (dis)agreements stands for a count of the items to which a positive (negative) response was given. Similarly, ERS was computed as the log odds of the number of extreme responses plus one over the number of non-extreme responses. Extreme responses were defined as responses in the most positive and the most negative categories.

ERS = ln ((# extreme responses + 1) / (# non-extreme responses + 1)); with # signifying the frequency of (non)extreme responses (2)

NARS and ERS had a range from -3.09 (which corresponds to ln(1/22) for respondents who did not engage in NARS or ERS) through 3.09 (which corresponds to ln(22) for respondents who answered all items positively or extremely). An NARS (ERS) value of zero indicates that a respondent gave as many positive (extreme) responses as negative (non-extreme) responses. The correlation between NARS and ERS was -.08 (p = .004).

To assess concurrent validity, we estimate the correlation between our proposed NARS measure and the traditional NARS measure based on the mean of the items, as well as the correlation between our proposed ERS measure and the traditional ERS measure based on the standard deviation of the items (Greenleaf, 1992b; Baumgartner & Steenkamp, 2001). Because the traditional measures are scale format specific, we average the correlations of the new and traditional measures across the 8 experimental conditions. For NARS the correlation is .74, for ERS the correlation is .78. Hence, the shared variance (i.e., r^2; Fornell & Larcker, 1981) exceeds 50% in both cases, providing evidence in support of concurrent validity of the proposed measures.

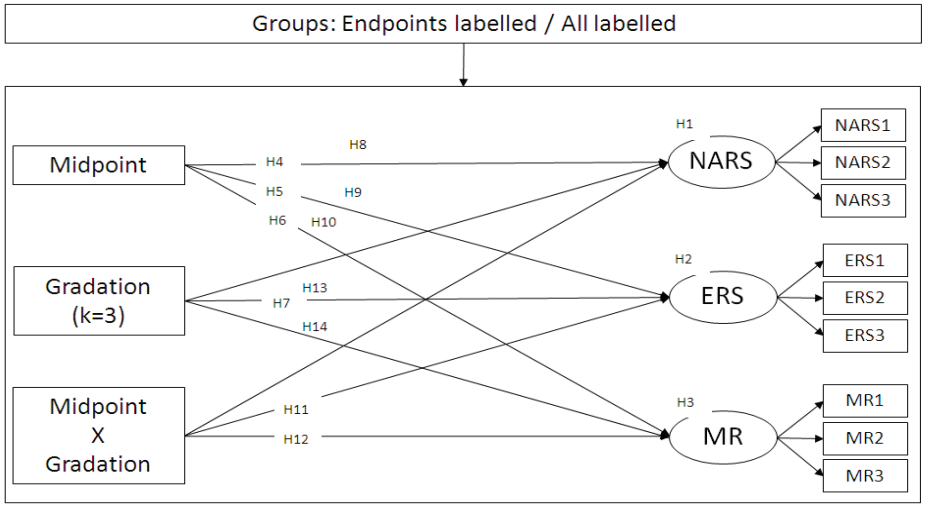
IIa.4.1.4 Findings

Figure 1 shows the model we test. In line with Weijters et al. (2008), we create three indicators for NARS and three indicators for ERS by splitting the items in three groups (item 1, 4, 7... for group 1; item 2, 5, 8... for group 2, etc.). As a result, we can model NARS and ERS as two latent factors with three scale level indicators each, thus accounting for unique variance in the response style indicators due to content specificity and random error. MR is modeled as a latent factor with three binary indicators: each indicator is based on one reversed item pair and takes on a value of 0 if no MR occurs for this item pair and a value of 1 if MR does occur for this item pair[10].

[10] We verified that using a summated score for MR gave parallel results and led to the same substantive conclusions.

FIGURE 1:

RESPONSE STYLES AS A FUNCTION OF SCALE FORMAT CHARACTERISTICS (STUDY 1)



NARS = NET ACQUIESCENCE RESPONSE STYLE; ERS = EXTREME RESPONSE STYLE; MR = MISRESPONSE TO REVERSED ITEMS. RESIDUAL TERMS AT THE CONSTRUCT AND INDICATOR LEVEL ARE NOT SHOWN FOR READABILITY

We code the experimental variables as follows. The labeling manipulation is used as the grouping variable (group one contains the conditions where only the extremes are labeled, group two contains the conditions where the response categories are fully labeled). The manipulations related to the number of scale points (gradations and midpoint) are coded by means of effect coded variables. For gradations, we create a variable that takes on a value of -1 for conditions with 2 gradations of (dis)agreement and a value of 1 for conditions with 3 gradations of (dis)agreement. For midpoint, we create a variable that takes on a value of -1 if no midpoint is present and a value of 1 if a midpoint is present. We also include a contrast variable to account for the gradation by midpoint interaction, coding the seven-point condition as 1, the other formats as -1/3. Hence, this variable captures the

effect (not explained by the main effects) of simultaneously having 3 gradations and a midpoint (resulting in a seven-point scale). The coding scheme is summarized in Table 2.

TABLE 2:

CODING OF THE EXPERIMENTAL CONDITIONS (STUDY 1)

Experimental condition		Coding			
Labeling	Number of categories	Labeling	Adding gradations	Midpoint	Seven-point[11]
Endpoints labeled	4	Group 1	-1	-1	-1/3
	5	Group 1	-1	1	-1/3
	6	Group 1	1	-1	-1/3
	7	Group 1	1	1	1
All categories labeled	4	Group 2	-1	-1	-1/3
	5	Group 2	-1	1	-1/3
	6	Group 2	1	-1	-1/3
	7	Group 2	1	1	1

We specify NARS, ERS and MR as latent factors with three indicators each. The NARS, ERS and MR factors are regressed on the experimental variables. The regression weights capture the effects of increasing the number of gradations to 3 and of including a midpoint, or both, relative to the grand mean and while controlling for the other experimental manipulations.

Group differences in the NARS, ERS and MR intercepts reflect the effect of labeling. We assess the labeling effects by means of Wald chi² tests (testing the hypothesis of a null effect). For the hypothesis tests, we use alpha=0.05 as the threshold for statistical significance, but we do report exact p-values for completeness. We estimate the model with the WLSMV estimator in Mplus 5.1 (Muthén & Muthén, 2007). As respondents were

[11] The seven-point scale format was used as a contrast group against which we test the other groups. This makes it easier to interpret the results. However, we verified that the substantial conclusions hold when using Ancova's with contrast effects. The results remained.

randomly assigned to groups, the measurement parameters (factor loadings, indicator residuals and indicator intercepts) were set to equality across groups (extremes labeled versus all labeled).

The model fits the data acceptably well ($chi^2(57) = 107.71$, $p = .0001$; CFI = .952; TLI = .953; RMSEA = .038). All indicators have substantial and highly significant standardized factor loadings (.589, .577, .573 for NARS; .806 .835, .831 for ERS; .428, .855, .842 for MR[12]; all $p<.001$), indicating that the multiple indicators for the response styles indeed tap into a common underlying dimension. In other words, convergent validity of the multiple indicators per response style is supported. The variance explained (R^2) by the experimental variables is 11.3% for NARS, 15.3% for ERS, and 45.2% for MR. The observed proportions of MR are shown in Table 3. The model estimates are shown in Table 4.

[12] As pointed out by the Area Editor, it is interesting to see that the loading of indicator a on MR is smaller than the other two. Indicator a is about the product, while b and c are about the brand. Further research could find out whether brands have a higher effect on MR than products. In other words, whether the context of the item has an effect on the level of MR.

TABLE 3:

MR (% OF MISRESPONDERS TO REVERSED ITEMS) BY RESPONSE FORMAT (STUDY 1)

		Indicator			
Labeling	Number of categories	a	b	c	Average
All labeled	4	52.6%	65.7%	67.2%	61.8%
	5	11.1%	7.8%	12.4%	10.5%
	6	46.2%	60.8%	62.9%	56.6%
	7	22.0%	6.0%	16.7%	14.9%
Endpoints labeled	4	50.3%	61.1%	60.6%	57.3%
	5	27.7%	19.4%	21.3%	22.8%
	6	57.1%	68.8%	67.5%	64.5%
	7	38.1%	37.4%	39.6%	38.4%
Average		38.1%	40.9%	43.4%	40.8%

By means of Wald chi² group difference tests, we test for group differences in regression weights (i.e., moderating effects of labeling). Therefore, we set the regression weights to equality across groups (i.e., this in order to reveal whether the estimates are equal for the extremes labeled group and the fully labeled group; cf. Table 4). So, when there are differences across groups than we can determine that there are significant main effects of labeling, significant two-way interactions and perhaps a signficiant three-way interaction. The invariant weights were set to equality. From table 4, one can see that for 6 out of the 9 regressions the effect of labeling is non-significant (i.e., group differences are invariant). As such, results indicate that the three-way interactions of labeling, gradations and midpoint were not significant for NARS ($chi^2(1) = 0.02$, $p = 0.893$), ERS ($chi^2(1) = .99$, $p = .320$), and MR ($chi^2(1) = .02$, $p = .881$).

The same is true for the two-way interactions of labeling with gradations on NARS ($chi^2(1) = .04$, $p = .834$), the two-way interaction of labeling with midpoint on NARS ($chi^2(1) = .33$, $p = .567$) (thus, no evidence is found in support of H8), and the two-way interaction of labeling with midpoint on ERS ($chi^2(1) = 1.25$, $p = .263$) (thus, no evidence is found in support of H9). For an overview of the hypotheses see Table 5.

TABLE 4:

MODEL ESTIMATES OF FORMAT EFFECTS ON NARS, ERS, MR (STUDY 1)

			Extremes labeled				All labeled				
	DV	IV	Estimate	SE	t	p	Estimate	SE	t	p	
B	NARS	MIDPOINT	0.078	0.018	4.32	0.000	0.078	0.018	4.32	0.000	H4
		Gradations (K=3)	0.017	0.019	0.91	0.183	0.017	0.019	0.91	0.183	
		Interaction (7-point scale)	-0.012	0.038	-0.31	0.378	-0.012	0.038	-0.31	0.378	
	ERS	MIDPOINT	-0.117	0.029	-4.09	0.000	-0.117	0.029	-4.09	0.000	H5
		Gradations (K=3)	-0.242	0.044	-5.47	0.000	-0.097	0.038	-2.54	0.006	H7
		Interaction (7-point scale)	0.133	0.062	2.16	0.016	0.133	0.062	2.16	0.016	H11
	MR	MIDPOINT	-0.703	0.081	-8.71	0.000	-1.132	0.093	-12.11	0.000	H6
		Gradations (K=3)	0.134	0.074	1.82	0.035	-0.090	0.077	-1.16	0.123	
		Interaction (7-point scale)	0.295	0.125	2.37	0.009	0.295	0.125	2.37	0.009	H12
Intercepts	NARS		0.000				0.168	0.027	6.32	0.000	H1
	ERS		0.000				-0.436	0.045	-9.67	0.000	H2
	MR		0.000				-0.490	0.084	-5.81	0.000	H3

Labeling

The group differences in the intercepts of NARS, ERS and MR represent the effect of labeling. The intercepts of group one (extremes labeled) are zero as to the model specification, so the t-test of the intercepts in group two (all labeled) provide a test of the labeling effect. The intercept estimates are shown in the lower rows of Table 4. Labeling has a significant effect on all three dependent variables and leads to higher NARS (H1), lower ERS (H2) and lower MR (H3).

Midpoint

Inclusion of a midpoint leads to a significant increase in NARS (H4) and a significant decrease in ERS (H5) (cf. Table 4). The decrease in ERS is smaller when the inclusion of the midpoint is combined with an increase of the number of gradations from 2 to 3 (H11). Adding the midpoint leads to lower MR (H6). As expected, we found more negative MR for the three indicators (respectively 40% for indicator a, 35% for indicator b, 42% for indicator c) than positive (1% for indicator a, 3% for indicator b, 1% for indicator c) (respectively $t(1205)=-26.963$, $p<0.001$; $t(1205)=-20.866$, $p<0.001$; $t(1205)=-27.533$, $p<0.001$). In line with H10, the reduction in MR due to the inclusion of a midpoint is significantly stronger in the fully labeled conditions (the parameter estimates are significantly different across groups: Wald $chi^2(1) = 13.31$, $p = .0003$). Also, the decrease in MR due to inclusion of the midpoint is weaker when the number of gradations is three (H12).

Gradations

Increasing the number of gradations from 2 to 3 does not affect NARS, but results in a significant decrease in ERS (H7), and this effect is stronger in the extremes labeled conditions (H13) (the parameter estimates are significantly different across groups: Wald $chi^2(1) = 6.12$, $p = .013$). Increasing the number of gradations increases MR, but only so in the extremes

labeled conditions (H14): the effect is non-significant in the fully labeled condition (the parameter estimates are significantly different across groups; Wald $chi^2(1) = 4.39$, $p = .036$).

TABLE 5:

SUMMARY OF HYPOTHESIS TESTS (STUDY 1)

Hypothesis	**Test**	**Decision**
H1: Labeling all response categories leads to higher levels of NARS.	$t = 6.32$, $p < .001$	Support
H2: Labeling all response categories leads to lower levels of ERS.	$t = -9.67$, $p < .001$	Support
H3: Labeling all response categories leads to lower levels of MR.	$t = -5.81$, $p < .001$	Support
H4: NARS increases when adding a midpoint.	$t = 4.32$, $p < .001$	Support
H5: ERS decreases when adding a midpoint.	$t = -4.09$, $p < .001$	Support
H6: MR decreases when adding a midpoint.	$t = -8.71$, $p < .001$ for group 1, $t = -12.11$, $p < .001$ for group 2	Support
H7: ERS decreases when more gradations of (dis)agreement are offered	$t = -5.47$, $p < .001$ for group 1, $t = -2.54$, $p < .001$ for group 2	Support
H8: Full labeling of the response categories strengthens the positive effect of offering a midpoint on NARS.	$chi^2(1) = .33$, $p = .567$	No support
H9: Full labeling of the response categories strengthens the negative effect of offering a midpoint on ERS.	$chi^2(1) = 1.25$, $p = .263$	No support
H10: Full labeling of the response categories strengthens the negative effect of inclusion of a midpoint on MR.	$chi^2(1) = 13.31$, $p < .001$	Support
H11: The presence of a midpoint mitigates the negative effect of adding more gradations of (dis)agreement on ERS.	$t = 2.16$, $p = .016$	Support

H12: Offering a midpoint diminishes the negative effect of adding more gradations of (dis)agreement on MR.	$t = 2.37$, $p = .009$	Support
H13: Fully labeling scales weakens the negative effect of adding more gradations of (dis)agreement on ERS.	$chi^2(1) = 6.12$, $p = .013$	Support
H14: In formats with labels for the endpoints only, adding more gradations of (dis)agreement leads to an increase in MR.	$chi^2(1) = 4.39$, $p = .036$	Support

IIa.4.1.5 Impact of format on intention measures

If an analyst would want to report trial and purchase intentions of a product, s/he might use the percentage of respondents agreeing with intention items as a simple and efficient statistic. To make the impact of the format manipulation and the resulting differences in response distributions tangible, Table 6 presents the percentage of respondents agreeing with two intention items. As shown in Table 7, the distributions were significantly affected by labeling and inclusion of a midpoint, but not the addition of a gradation of (dis)agreement. The two-way interaction effects are all significant. However, the significant interactions between labeling x gradations and midpoint x gradations are driven by the strong significant main effect of labeling and midpoint on purchase intentions. That is, a fully labeled scale format leads to an increase in purchase intention both for 2 gradation scales as for 3 gradation scales. The opposite is true when a midpoint is offered. For the interaction between labeling x midpoint, results show that when a midpoint is offered, the reduction in purchase intention is higher when only the endpoints are labeled compared to a fully labeled scale. Depending on the scale format used, estimates of the percentage of responders agreeing with the intention items varied between 22.6% and 60.6%. This finding succinctly demonstrates the danger of interpreting item scores in an absolute way. The results in Table 6 also illustrate the relevance of the effects observed in the main study: conditions associated with higher NARS indeed result in higher proportions of respondents expressing a positive intention.

TABLE 6:

% AGREEING TO INTENTION ITEMS BY RESPONSE FORMAT CONDITION (STUDY 1)

% agree	k	Item 1	Item 2
Extremes only	4	50.3%	48.6%
	5	24.5%	22.6%
	6	44.2%	46.1%
	7	27.3%	23.0%
All options labeled	4	60.6%	57.7%
	5	38.6%	37.9%
	6	51.0%	49.7%
	7	42.7%	48.0%

Item 1 = "I would like to try this product." - Item 2 = "Next time I make a purchase in this product category, I will consider the product that was shown." - k = number of response categories

TABLE 7:

CHI² TEST FOR INTENTION MEASURES BY EXPERIMENTAL CONDITION (STUDY 1)

		Total effect	Main effects			Interaction effects		
		Conditions	Labeling	Midpoint	Gradations	Labeling x Midpoint	Labeling x Gradations	Gradations x Midpoint
	df	7	1	1	1	1	1	1
Item 1	chi²	61.727	13.902	39.55	0.381	57.100	14.448	40.359
	p	0.000	0.000	0.000	0.537	0.000	0.002	0.000
Item 2	chi²	69.311	18.537	36.889	0.035	64.079	18.815	43.783
	p	0.000	0.000	0.000	0.852	0.000	0.000	0.000
Average	chi²	65.52	16.22	38.22	0.21	60.59	16.63	42.07
	p	0.000	0.000	0.000	0.648	0.000	0.001	0.000

Item 1 = “I would like to try this product.” - Item 2 = “Next time I make a purchase in this product category, I will consider the product that was shown.” - k = number of response categories.

IIa.4.1.6 Discussion Study 1

This first study demonstrates that the scale format components labeling and the number of response categories affect NARS, ERS and MR. The main conclusion therefore is that empirical results based on different scale formats may not be comparable. Also, interpreting levels of agreement with Likert items in an absolute sense (e.g., 'the majority of respondents agree') is necessarily a tentative exercise at best.

Current practice is validated to some extent by our findings, in that formats with an even number of categories are hardly used in practice and also perform poorly in terms of MR in the current study.

Yet, the default format in marketing scales, i.e. the 7 point scale with labels at the extremes, does not necessarily provide the best data quality. The problem associated with this scale format is the higher level of MR compared to the 5 point scale with labels at the extremes.

Researchers evaluating the results of Study 1 may look for better alternatives than the default 7 point scale with labels at the endpoints by reasoning as follows. The results indicate that a five point scale with labels at the extremes results in better data quality, as it leads to lower MR. Labeling all response options would further decrease MR but may be difficult in cross-cultural research. Our results show that labeling also results in higher NARS, but – in absence of a criterion measure – it is not clear to what extent this is problematic. To address the latter issue, i.e. whether or not all response categories should be labeled, we set up an additional study.

IIa.4.2 Empirical Study 2

We set up Study 2 to investigate labeling effects more closely for five point scale formats. Note that labeling all response categories is more common for this number of response categories than for formats with any other number of categories (see Table 1).

IIa.4.2.1 Design and sample

To further cross-validate and extend our findings, we conducted an additional online survey among a sample of British respondents. For this study, we focused on five point scales only and manipulated the labeling of the response categories at two levels (only the extreme response categories were labeled or all response categories were labeled). The response category labels were 'strongly disagree', 'slightly disagree', 'neutral', 'slightly agree' and 'strongly agree'. Respondents were randomly assigned to the two conditions (N = 113 for the all labeled condition; N = 113 for the extremes labeled condition). The sample was randomly drawn from all UK residents in the panel of an Internet marketing research company. Age ranged from 18 through 85, with a median of 55 years (SD = 14.5). In our sample, 32.7% of respondents were female and 65.5% had attended college or university.

IIa.4.2.2 Instrument

The questionnaire was inspired by Greenleaf's (1992a) work and contained questions related to 10 diverse but common behaviors. Intentions related to all behaviors were measured on a %-scale and the question "How likely is it that you will do the following activities at least once during the next 2 weeks? Please indicate a number from 0% to 100%. 0% means 'definitely not' (i.e., there is no chance I will do this the next two weeks) and 100% means 'definitely will' (i.e., it is certain that I will do this activity in the next two weeks). Numbers in between indicate how likely it is you will do the activity (e.g., 50% means there is a

fifty/fifty chance that I will do this activity in the next two weeks).” This question is concrete and specific, and uses a format that has an objective meaning (probabilities). For these reasons, we assume that the data obtained with this measure do not share substantial method bias with attitudinal Likert scales (Greenleaf, 1992a; Rindfleisch et al., 2008).

Later in the questionnaire, the attitude towards each behavior was probed with a 5-point Likert item and the following question: “Please indicate to what extent you (dis)agree with the following statements. In general, I like to….” With the following behaviors listed subsequently: go shopping; go to a restaurant; invite friends at my place; attend a concert; go for a walk; go to the gym; play computer game(s); communicate online with friends (chat, e-mail, Facebook); go to the cinema; go to a bar to have a drink with friends. The average inter-item correlation across behaviors was .21 for the intention items and .18 for the attitude items, indicating that the activities were heterogeneous.

IIa.4.2.3 Findings and discussion: The effect of labeling on attitude-intention models

We relate intentions measured on a %-scale to attitudes measured on 5-point Likert scales that either have all categories labeled or only the extremes labeled. This allows us to study how labeling affects model estimates in simple regression models of a type that is quite common in marketing research. The findings from Study 1 provide some hypotheses on how model estimates may be biased.

Consider a simple linear regression where intention on a %-scale is regressed on attitude on a 5-point scale. As the intention scale is the same across conditions, differences in model estimates can be attributed to the attitude measurement effects. We expect that attitude measures in the fully labeled condition show higher NARS. This could translate in higher

observed means and/or lower intercept terms (Greenleaf, 1992a). The reason for the latter is that the attitude responses will be inflated relative to the intention scores; a negative shift in intercept compensates for this. Attitude measures in the endpoints labeled condition are expected to show higher ERS and we therefore expect higher variances in this condition. A key question that relates to this but that was not yet addressed in Study 1, is which of the two formats shows highest criterion validity. Higher criterion validity would show up in higher explained variance and a higher regression weight, since the regression is univariate and true components[13] can be assumed to be identical.

We study several behaviors' attitude-intention pairs. In the questionnaire, ten were included. A preliminary analysis shows that for one behavior, 'go to a restaurant', the intention score is significantly different across conditions ($t(224)=-2.139$, $p = .034$). As this suggests that the two random samples coincidentally differ in terms of this behavior, we omit this attitude-intention pair for further analysis, leaving us with 9 pairs. In the model of interest, every intention item is regressed on its related attitude item. The attitude items correlate freely, as do the (residual terms of the) intention items. Using this model, we can investigate whether the difference in labeling of the attitude items affects model estimates.

We first verify that the 9 remainder intention measures are invariant across conditions in terms of means, variances and covariances. This seems to be the case as the nested chi square invariance tests are all insignificant: $chi^2(9) = 8.21$, $p=.513$ for the means, $chi^2(9)=13.28$, $p = .150$ for the variances, and $chi^2(36)=34.94$, $p = .519$ for the covariances. Thus, any subsequent

[13] Observed scores consist of one's true score plus random and systematic error variance. One can assume that the respondents' true score is free of any bias and as such comparable across conditions (Baumgartner & Steenkamp, 2006).

violation of cross-group invariance in the model can be attributed to the responses to the attitude questions.

TABLE 8:

MODEL FIT INDICES FOR INVARIANCE TESTS BETWEEN ALL LABELED AND EXTREMES LABELED CONDITIONS (STUDY 2)

	Chi² test			Chi² difference test		
Model	Chi²	DF	p	ΔChi²	ΔDF	Δp
Unconstrained	158.13	144	0.199			
Attitude means	169.01	153	0.178	10.88	9	0.284
Attitude variances	177.92	153	0.082	19.79	9	0.019
Intention intercepts	175.71	153	0.101	17.58	9	0.040
Regression weights	187.95	153	0.029	29.81	9	0.000

Literature (i.e., Cheung & Rensvold, 2000; Little, 2000) suggests to test invariance based upon a series of nested models. The assumption behind this kind of testing is that there is invariance between weights, intercepts and variances across groups. However, in this study we do expect that weights, intercepts and variances differ across groups. So, we start from the assumption that there is non-invariance. As a consequence, our models are not nested, since each aspect, attitude weights, intercepts and variances, needs to be tested independently. The unconstrained model fits the data well (see unconstrained model in Table 8) and we use this unconstrained model as the reference model against which we test invariance restrictions. The invariance restrictions test the hypotheses that parameter estimates are the same in the two conditions (all categories labeled versus extremes labeled). In the first model ('attitude means'), the chi square difference test tests the null hypothesis that the means of the 9 attitude items are equal across the two experimental conditions. This hypothesis is not rejected (p = .284). The subsequent tests (also using the unconstrained model as the reference model) indicate that invariance is rejected for the attitude variances, the intention intercepts and the

regression weights from attitude to intention items (all $p < .05$). The model estimates for the latter parameters (that are not the same across conditions) are shown in Table 9. The data were coded as follows: 'Strongly disagree' = -2; 'Slightly disagree = -1'; 'Neutral' = 0; 'Slightly agree' = 1; 'Strongly agree = 2'. Consequently, the intercept term is the expected intention score corresponding to a neutral attitude. The last four columns of Table 9 contain an index based on the ratio of the estimate in the all categories labeled condition over the estimate in the extremes labeled condition.

With one exception, the regression weights in the extremes condition are greater than the regression weights in the all condition. The R^2 estimates are consistently greater in the extremes condition. The intercepts are greater in the extremes condition for 7 out of 9 behaviors[14]. The variances are greater in the extremes condition for 6 out of 9 behaviors. Overall, these results support the notion that the attitude measures in the all labeled condition show higher NARS and lower ERS.

Importantly, the explained variance, which indicates criterion validity, is consistently and substantially higher in the extremes condition. The model implied regression slopes are shown in Figure 2, illustrating the higher intercept and slope for the Extremes condition. In sum, the results of this follow-up study indicate that the extremes only scale format performs better than the fully labeled scale format in terms of criterion validity, and that NARS due to full labeling is more problematic than ERS due to endpoints only labeling.

[14] We note that the intention intercept test is more sensitive than the attitude means test (as attitude serves as a covariate of the experimental effect for the former).

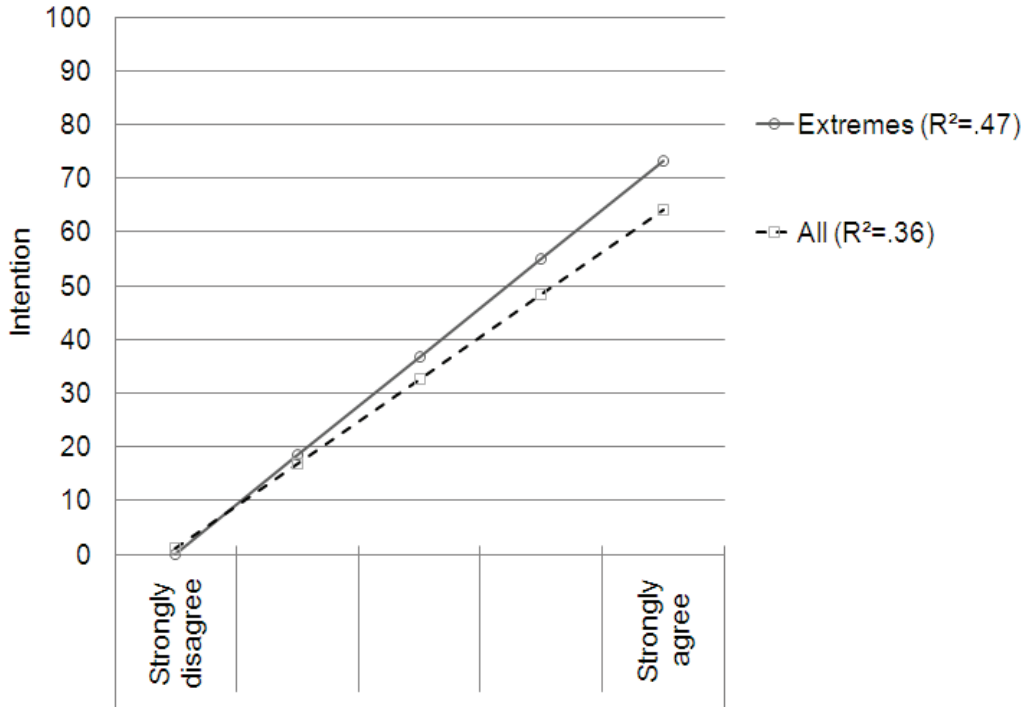
TABLE 9:

REGRESSION MODEL ESTIMATES BY CONDITION (STUDY 2)

	All labeled (G1)				Extremes labeled (G2)				Index(G2/G1)			
	B (s.e.)	R^2	Intercept (s.e.)	Var (s.e.)	B (s.e.)	R^2	Intercept (s.e.)	Var (s.e.)	B	R^2	Intercept	Var
Go shopping	7.6 (2.0)	0.07	69.3 (3.6)	1.2 (0.2)	13.8 (1.8)	0.27	68.1 (3.0)	1.6 (0.2)	182%	417%	98%	127%
Invite friends	18.1 (2.6)	0.23	27.6 (3.9)	0.9 (0.1)	18.0 (1.8)	0.37	30.4 (3.2)	1.7 (0.2)	100%	160%	110%	192%
Attend a concert	6.0 (1.3)	0.16	7.9 (1.8)	2.0 (0.3)	8.1 (1.4)	0.20	10.2 (2.2)	2.0 (0.3)	135%	124%	129%	100%
Go for a walk	24.7 (1.9)	0.44	34.3 (3.6)	1.1 (0.1)	25.7 (1.8)	0.53	38.0 (3.4)	1.4 (0.2)	104%	121%	111%	127%
Go to the gym	12.0 (1.5)	0.34	23.5 (2.6)	1.7 (0.2)	19.1 (1.1)	0.66	33.5 (2.1)	2.2 (0.3)	159%	194%	142%	131%
Play computer game(s)	21.3 (1.1)	0.68	38.1 (2.3)	2.8 (0.4)	22.9 (1.2)	0.69	44.3 (2.3)	2.6 (0.3)	108%	102%	116%	94%
Online with friends	22.3 (1.6)	0.43	45.7 (3.4)	1.3 (0.2)	18.9 (1.5)	0.46	55.4 (3.0)	1.9 (0.2)	85%	107%	121%	146%
Go to the cinema	10.2 (1.3)	0.33	15.3 (2.1)	2.2 (0.3)	15.3 (1.4)	0.45	18.8 (2.4)	2.1 (0.3)	150%	136%	122%	94%
Go to a bar	19.7 (1.5)	0.54	32.5 (2.5)	2.2 (0.3)	22.7 (1.6)	0.57	31.7 (2.8)	2.1 (0.3)	115%	105%	98%	98%
Average	15.8 (1.6)	0.36	32.7 (2.9)	1.7 (0.2)	18.3 (1.5)	0.47	36.7 (2.7)	2.0 (0.3)	126%	163%	116%	123%

FIGURE 2:

LABELING RESPONSE OPTIONS LEADS TO DIFFERENT REGRESSION FUNCTIONS (STUDY 2)



IIa.5 General discussion

In recent years, a growing number of researchers have used questionnaires with Likert-type rating scales in order to understand, explain and predict the behavior of participants. However, researchers often use different rating scale formats with varying numbers of response categories and labels since they have only few guidelines when selecting a specific format. This article examines the effects of these scale format characteristics on the response distributions and the level of MR in order to provide better insight in the optimal scale format choice.

In study 1, we experimentally manipulated the rating scale format of items, varying the number of the response categories from 4 up till 7 and the labels of the response categories (all labeling versus endpoints only). Our results demonstrate significant effects of scale

format characteristics on NARS, ERS and MR, and thereby shed light on the processes that are involved in such effects.

NARS is higher in conditions where all response categories are labeled. We attribute this effect to the clarity of a fully labeled version which enhances the effect of positivity bias (Tourangeau et al., 2000). A fully labeled scale format also leads to lower ERS scores due to the increased salience and attractiveness of the intermediate options. In addition, labeling all response categories leads to less MR. When only the end categories are labeled; respondents have to mentally map the rating scale by assigning meanings to the unlabeled response categories. This leads to ambiguity and a higher cognitive load, both of which may result in higher levels of MR (Krosnick, 1991; Swain et al., 2008).

Including a midpoint led to an increase in NARS due to a disproportional movement of otherwise negative response options to the midpoint, when provided. Ambivalent respondents who are forced to take sides tend to react negatively (Gilljam & Granberg, 1993). This finding is in concordance with the findings of Nowlis et al. (2002) in that the distribution shift is evoked by ambivalent respondents. However, it is not only the focus on the most important attribute that determines the choice of response category; also the task-related negative emotions play an important role.

The inclusion of a midpoint also resulted in lower levels of MR and ERS. The effect of the inclusion of a midpoint on data quality is bigger in fully labeled formats as compared to endpoint labeled formats, in that MR is even lower when an odd scale format is fully labeled. In contrast with our expectations, the inclusion of a midpoint in combination with a fully labeled scale format did not affect the level of NARS or the level of ERS. This may relate to

the perception respondents have of the rating scale format when a midpoint is added. According to Marsh and Parducci (1978), respondents perceive a scale as more equidistant when a midpoint is added irrespective of whether the scale is fully labeled or not. This implies that through this perception of equidistance, respondents have clarity concerning all response options. It also implies that the amount of ambivalent and truly neutral respondents that opt for the midpoint does not depend on the labeling of the rating scale.

Adding gradations of (dis)agreement does not translate into an alteration in the level of NARS and MR as the addition of extra response categories will not change the valence of the respondent's response choice (Marsh & Parducci, 1978). When only the endpoints are labeled, an addition of extra response categories led to higher MR as the valence of the intermediate response categories for this scale format is unclear. Furthermore, MR increases with an increasing number of gradations conditional on the presence of a midpoint. Therefore, the decrease in MR when a midpoint is offered will be lower when there are more gradations of (dis)agreement. In terms of ERS, the presence of extra intermediate response categories and the possibility to better qualify the strength of a response reduces the level of ERS. This effect is strengthened when all response categories are labeled or when a midpoint has been offered.

Study 2 focused on the labeling effect on ERS and NARS. Findings replicate study 1 in that a fully labeled scale format led to higher NARS and lower ERS. More importantly, we find that criterion validity is higher in the extreme labeled condition, meaning that the latter provides better data for estimation of linear models. It should be noted that Study 2 is only a first, preliminary study into the topic of labeling. We discuss some suggestions for further research in the last section of the current paper.

IIa.5.1 Implications

It is clear that the response format characteristics affect the central tendency, spread and internal consistency of self-report data. Consequently, data obtained with different formats are not comparable and interpretations of Likert data are always relative: the probability that respondents agree with an item depends on how such agreement can be expressed. In setting up studies, researchers need to make a well-considered choice for a specific format and they need to explicitly report upon this choice. Meta-analyses will have to take into account response format as a factor influencing estimates.

The practice of reporting survey results by means of percentages of respondents who agree with a statement ('top two boxes' or 'top three boxes') has to be treated with great caution. As shown in Table 6, the percentage of respondents with positive trial and purchase intentions varied widely across formats (from 22.6% through 60.6%). Also for regressions, differences in format lead to differences in model estimates and model fit. As shown in Table 9, formats with endpoint labels only, lead to a stronger linear relation between attitudes and intention compared to fully labeled formats.

The current findings advance our theoretical understanding of NARS, ERS, MR and rating scale formats in several ways. First, our study provides additional insights in the age-old debates of whether to label all options, whether to include a midpoint, and the right amount of response options. Our findings highlight the importance of making the right choices when constructing a survey scale. We posit that the question of whether or not to include a midpoint depends not only on the particular research goals (Nowlis et al., 2002) but also on the risk for MR in the data. The inclusion of a midpoint led to a reduction in MR. A 4 or 6-point scale

format can be used only in cases where respondents have clear-cut answers (so neither ambivalence nor indifference can arise) and where no reversed coded items are present in the scale. Overall, we suggest avoiding scales without a midpoint, unless particular and relevant reasons present themselves.

Our study contributes to the response bias literature by identifying a previously unrecognized antecedent of MR. This relates back to the four cognitive processes respondents perform when answering an item: (1) comprehension, (2) retrieval, (3) judgment, and (4) response (Tourangeau et al., 2000). Previous work has focused on MR due to problems in comprehension (Schmitt & Stults, 1985), retrieval (Weijters et al., 2009) and/or judgment (Swain et al., 2008; Weijters et al., 2009). Our findings demonstrate that MR can also be caused by problems in mapping a judgment onto a specific response category, i.e., difficulties in the response process.

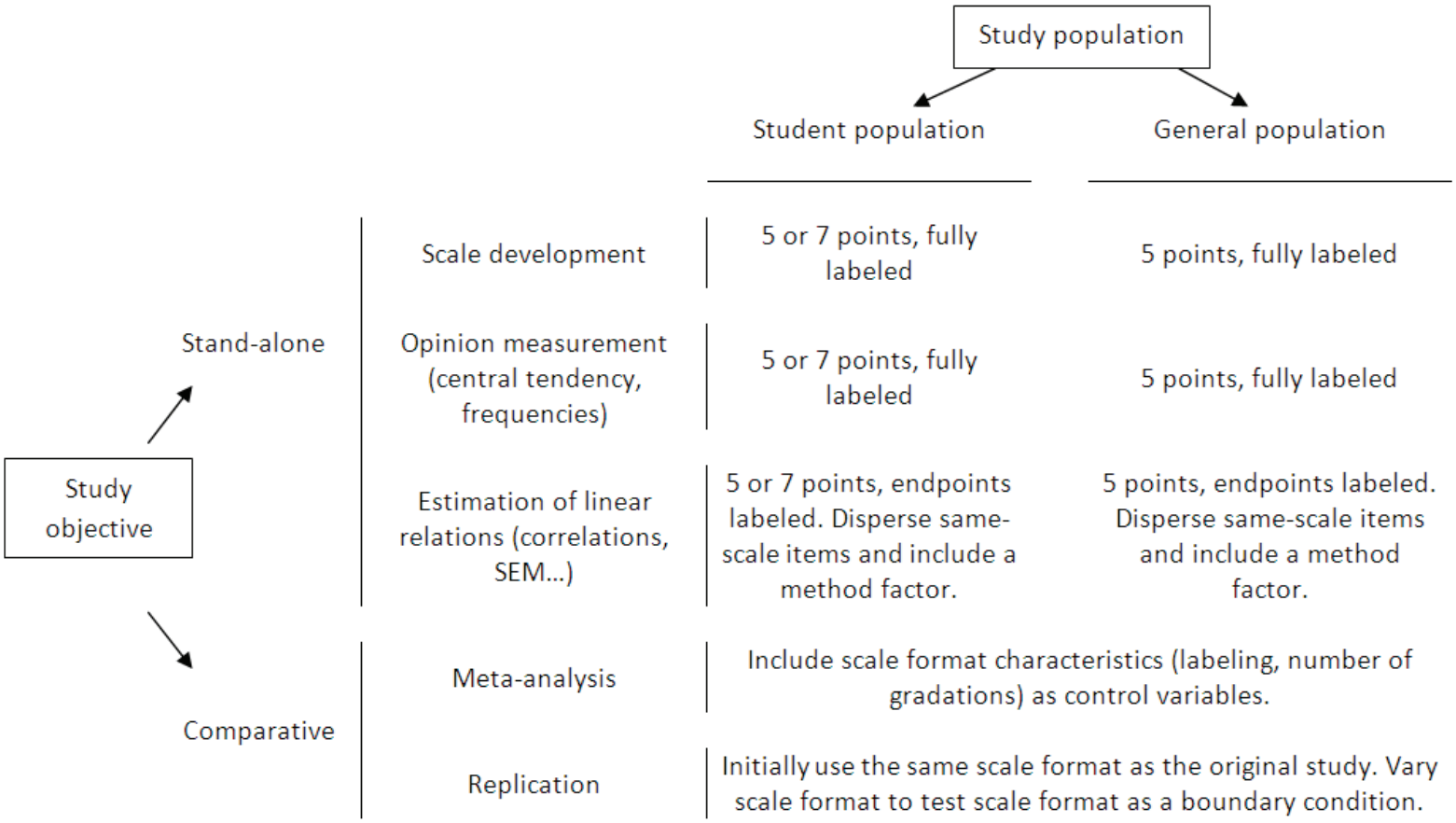
IIa.5.2 Preliminary framework for selecting a response scale format

We propose a preliminary framework for selecting a response scale format. The current results are not conclusive[15], and the framework can serve as a guideline when choosing a scale format until further evidence becomes available. Also, it may provide avenues for further methodological enquiries into scale format choice. We base this framework on the extant literature on the topic, complemented by the two empirical studies we presented in this paper. The framework is shown in Figure 3.

[15] We want to stress the preliminary aspect of Figure 3 since we have not checked whether this framework holds cross-culturally or for different languages. Therefore, the use of this framework may not be taken as granted. For instance, in extraverted countries the level of ERS is higher making the fully labeled scale format more interesting.

FIGURE 3:

PRELIMINARY DECISION FRAMEWORK FOR SELECTING A RESPONSE SCALE FORMAT



As shown in Figure 3, we distinguish studies based on two dimensions: the study population and the study objective. As for the study population, we focus on student populations versus general populations because these cover many instances of marketing research and because students tend to be relatively high in terms of cognitive and verbal ability and in terms of experience with questionnaires. These factors are likely to facilitate processing and make respondents less prone to response biases (Knauper, 1999; Krosnick, 1991; Marsh, 1996).

In selecting the optimal number of gradations, a tradeoff presents itself between maximizing the potential information transmission (Garner, 1960; Green & Rao, 1970) versus minimizing respondent demands (Krosnick, 1991; Weathers et al., 2005). We suggest it may be less problematic to use scales with more response categories (specifically 7 categories) for student populations (and other populations that rate high on cognitive and verbal ability and/or experience with questionnaires). For studies among the general population, it may be safer to stick to 5 point scales. In the current study (general population), 5 point scales led to slightly less MR. We note that for rating scales having at least five response options, linear models seem to be able to approximate the data quite well (Bollen & Barb, 1981; Srinivasan & Basu, 1989; Mullen, 1995; Hu & Bentler, 1999).

The choice for a particular scale format is further modulated by the study objective. When developing a new scale, researchers may want to reduce the risk of MR by fully labeling their scales. Otherwise, results may be biased against the inclusion of reversed items. There are some concerns when using a fully labeled scale though. Researchers need to be aware that agreement levels tend to be higher in this format. We also advise researchers to counter the cross-cultural disadvantage of labeling since it is not easy to find equivalent labels both in meaning and frequency of usage (Weijters, Geuens, & Baumgartner, 2010). If a researcher

wants to report direct summaries of responses (i.e. opinion measurement) by using means or percentages (e.g. top boxes), it may be better to opt for a fully labeled 5 point scale format (or fully labeled 7 point format for students) as labeling makes the scale more direct interpretable (e.g. a "5" means for both the researcher and respondents "strongly agree"). Though respondents tend to be internally consistent in this format, the downside is that they may be positively biased, so estimates should be interpreted as representing an optimistic scenario. We also stress the inherent relativity of scale responses. If a researcher wants to relate variables and estimate linear relations using correlations, regression models, Structural Equation Models (SEM), etc., an endpoint only 5 (or 7) point scale is the best choice since this format is used in a way that better conforms to linear models, thus providing higher criterion validity (cf. Study 2). When estimating linear relations, researchers can include a method factor that captures the response bias. Consequently, data can be purified from response styles by regressing the method factor on the regression of interest. If a researcher decides to include reversals, he/she should place the reversal at some distance from its affirmation to avoid an increase in MR (i.e., disperse same scale items) (Weijters, Geuens, & Schillewaert, 2009).

In a meta-analysis, the analyst can of course not select a scale format, but it is key to take scale format into account even so, in particular by including scale format characteristics as covariates (number of gradations, labeling). In replication studies, it may be safe to initially use the same scale format as the study one is replicating. Afterwards, it may in some instances be interesting to vary scale format as a boundary condition (especially in studies on factor structure).

IIa.5.3 Limitations and future research

To conclude, we note some limitations of our study that offer opportunities for future research. We only studied Likert-type items in this study. Future research might also examine the effects of labeling and the number of response categories in other formats, like semantic differentials.

An important limitation of Study 2 is the use of a self-report measure for assessing criterion validity. One might argue that this leaves open the possibility that 5-point Likert scales with labeled endpoints are more similar to %-scales than are 5-point Likert scales with labels for all response categories. We admit this as a limitation and we are in favor of further research into this topic, possibly using other criterion variables (like third rater reports, for example). However, there are several good reasons to believe that the current empirical context makes the likelihood that the results are due to a confound small. (1) There were filler tasks in between the two measures. This reduces the chance for carryover effects of response styles, as previous research has shown that there is a significant auto-regressive component to response styles, i.e., response styles in adjacent parts of a questionnaire are more similar than in distant parts of a questionnaire (Weijters, Geuens & Schillewaert, 2009). (2) The response formats (5 point Likert scale versus % scale) are very differently experienced by respondents, resulting in different response tactics and response quality (Weathers et al., 2005; Preston & Colman, 2000). In line with this, and referring to the work by Podsakoff et al. (2003) and Lindell & Whitney (2001), Rindfleisch et al. (2008, p. 263) recently recommended the use of different formats to minimize Common Method Variance (CMV): *"[…] surveys that employ a single-scale format (e.g., a seven-point Likert scale) and common-scale anchors (e.g., "strongly disagree" versus "strongly agree") are believed to be especially prone to CMV bias. […], the influence of measurement procedures can be reduced through measurement separation in a*

cross-sectional approach by employing different formats and scales for predictors versus outcomes […]." (3) For the intention question, respondents had to fill out a percentage themselves, rather than having to pick an option from a given set. (4) The difference in R^2 is large and consistent. In sum, we consider the use of a self-report for assessing criterion a limitation rather than a fatal flaw. Nevertheless, Study 2 is a first, preliminary investigation into this topic, as surely, more research is needed before we can draw solid conclusions.

A final intriguing question that remains unanswered is whether scale format interacts with culture in affecting response styles. We conducted Study 1 with Dutch speaking respondents and Study 2 with English speaking respondents. The observation that the findings from Study 1 carried over to the findings from Study 2 provides evidence in support of generalizability of our findings across at least the two languages under study. Further research needs to address generalizability beyond these contexts.

Appendix IIa.1 Items Study 1

Hartelijk dank om deel te nemen aan dit onderzoek. Het invullen van deze vragenlijst invullen vraagt slechts 5 tot 10 minuten van uw tijd.

Vergeet op het einde vooral niet mee dingen voor de digitale camera van Hewlett Packard.

Likert

4-Point All labeled

Volgende uitspraken hebben betrekking op het product, op de productcategorie of op het merk. In welke mate gaat u akkoord met deze stellingen?

	Helemaal niet akkoord	Niet akkoord	Akkoord	Helemaal akkoord
Ik zou dit product willen proberen	❑	❑	❑	❑
De volgende keer dat ik een aankoop doe in deze productcategorie, zal ik het getoonde product overwegen	❑	❑	❑	❑
Ik zal actief op zoek gaan naar dit product om het aan te kopen	❑	❑	❑	❑
Vergeleken met andere producten is dit product belangrijk voor mij	❑	❑	❑	❑
Ik ben niet geïnteresseerd in dit product	❑	❑	❑	❑
Als ik een merk koop uit die productcategorie, kies ik heel zorgvuldig	❑	❑	❑	❑
Ik hecht belang aan dit product	❑	❑	❑	❑
Ik hou van dit merk	❑	❑	❑	❑
Ik vind het een zeer slecht merk	❑	❑	❑	❑
Ik zou het een merk aan anderen aanraden	❑	❑	❑	❑
Ik denk dat het één van de beste merken uit zijn productklasse is	❑	❑	❑	❑
Dit merk is echt iets voor mij	❑	❑	❑	❑
In geen geval zou ik dit merk gebruiken	❑	❑	❑	❑

In welke mate gaat u akkoord met volgende stellingen?

	Helemaal niet akkoord	Niet akkoord	Akkoord	Helemaal akkoord
Ik ben er gerust in dat ik technologie-gerelateerde vaardigheden kan aanleren	❑	❑	❑	❑
Ik winkel omdat dingen kopen me gelukkig maakt	❑	❑	❑	❑
Menselijk contact bij het verlenen van diensten maakt het proces prettig voor de consument	❑	❑	❑	❑
Ik vind het heel belangrijk om het boodschappen doen goed te organiseren	❑	❑	❑	❑
Ik koop geen producten die overdreven verpakt zijn	❑	❑	❑	❑
De zaken die ik bezit zijn niet zo erg belangrijk voor mij	❑	❑	❑	❑
Ik beschouw mezelf als een merkentrouwe consument	❑	❑	❑	❑
Ik kleed me vaak op een manier die tegen de stroom ingaat, zelfs al zijn anderen daardoor verontwaardigd	❑	❑	❑	❑
In een groep mensen ben ik zelden het middelpunt van de belangstelling	❑	❑	❑	❑
Ik vind dat een geordend en regelmatig leven bij mij aard past	❑	❑	❑	❑
Luchtvervuiling is een belangrijk wereldwijd probleem	❑	❑	❑	❑
Een buitenshuis werkende vrouw met jonge kinderen is nog steeds een goed moeder	❑	❑	❑	❑
In het algemeen vind ik dat ik erg gelukkig ben	❑	❑	❑	❑
Financiële zekerheid is erg belangrijk voor me	❑	❑	❑	❑
TV-kijken is mijn belangrijkste vorm van ontspanning	❑	❑	❑	❑
We ervaren een achteruitgang in de levenskwaliteit	❑	❑	❑	❑
Ik voel me vaak misleid door reclame	❑	❑	❑	❑
In de winkel het prijsetiket van een product veranderen, vind ik volstrekt ontoelaatbaar	❑	❑	❑	❑
Ik ben erg met mijn gezondheid begaan	❑	❑	❑	❑
Ik heb het gevoel voortdurend in tijdnood te zijn	❑	❑	❑	❑

5-point All labeled

Volgende uitspraken hebben betrekking op het product, op de productcategorie of op het merk. In welke mate gaat u akkoord met deze stellingen?

	Helemaal niet akkoord	Niet akkoord	Neutraal	Akkoord	Helemaal akkoord
Ik zou dit product willen proberen	❑	❑	❑	❑	❑
De volgende keer dat ik een aankoop doe in deze productcategorie, zal ik het getoonde product overwegen	❑	❑	❑	❑	❑
Ik zal actief op zoek gaan naar dit product om het aan te kopen	❑	❑	❑	❑	❑
Vergeleken met andere producten is dit product belangrijk voor mij	❑	❑	❑	❑	❑
Ik ben niet geïnteresseerd in dit product	❑	❑	❑	❑	❑
Als ik een merk koop uit die productcategorie, kies ik heel zorgvuldig	❑	❑	❑	❑	❑
Ik hecht belang aan dit product	❑	❑	❑	❑	❑
Ik hou van dit merk	❑	❑	❑	❑	❑
Ik vind het een zeer slecht merk	❑	❑	❑	❑	❑
Ik zou het een merk aan anderen aanraden	❑	❑	❑	❑	❑
Ik denk dat het één van de beste merken uit zijn productklasse is	❑	❑	❑	❑	❑
Dit merk is echt iets voor mij	❑	❑	❑	❑	❑
In geen geval zou ik dit merk gebruiken	❑	❑	❑	❑	❑

In welke mate gaat u akkoord met volgende stellingen?

	Helemaal niet akkoord	Niet akkoord	Neutraal	Akkoord	Helemaal akkoord
Ik ben er gerust in dat ik technologie-gerelateerde vaardigheden kan aanleren	❑	❑	❑	❑	❑
Ik winkel omdat dingen kopen me gelukkig maakt	❑	❑	❑	❑	❑
Menselijk contact bij het verlenen van diensten maakt het proces prettig voor de consument	❑	❑	❑	❑	❑
Ik vind het heel belangrijk om het boodschappen doen goed te organiseren	❑	❑	❑	❑	❑
Ik koop geen producten die overdreven verpakt zijn	❑	❑	❑	❑	❑
De zaken die ik bezit zijn niet zo erg belangrijk voor mij	❑	❑	❑	❑	❑
Ik beschouw mezelf als een merkentrouwe consument	❑	❑	❑	❑	❑
Ik kleed me vaak op een manier die tegen de stroom ingaat, zelfs al zijn anderen daardoor verontwaardigt	❑	❑	❑	❑	❑
In een groep mensen ben ik zelden het middelpunt van de belangstelling	❑	❑	❑	❑	❑
Ik vind dat een geordend en regelmatig leven bij mij aard past	❑	❑	❑	❑	❑
Luchtvervuiling is een belangrijk wereldwijd probleem	❑	❑	❑	❑	❑
Een buitenshuis werkende vrouw met jonge kinderen is nog steeds een goed moeder	❑	❑	❑	❑	❑
In het algemeen vind ik dat ik erg gelukkig ben	❑	❑	❑	❑	❑
Financiële zekerheid is erg belangrijk voor me	❑	❑	❑	❑	❑
TV-kijken is mijn belangrijkste vorm van ontspanning	❑	❑	❑	❑	❑
We ervaren een achteruitgang in de levenskwaliteit	❑	❑	❑	❑	❑
Ik voel me vaak misleid door reclame	❑	❑	❑	❑	❑
In de winkel het prijsetiket van een product veranderen, vind ik volstrekt ontoelaatbaar	❑	❑	❑	❑	❑
Ik ben erg met mijn gezondheid begaan	❑	❑	❑	❑	❑
Ik heb het gevoel voortdurend in tijdnood te zijn	❑	❑	❑	❑	❑

6-point All labeled

Volgende uitspraken hebben betrekking op het product, op de productcategorie of op het merk. In welke mate gaat u akkoord met deze stellingen?

	Helemaal niet akkoord	Niet akkoord	Eerder niet akkoord	Eerder akkoord	Akkoord	Helemaal akkoord
Ik zou dit product willen proberen	❑	❑	❑	❑	❑	❑
De volgende keer dat ik een aankoop doe in deze productcategorie, zal ik het getoonde product overwegen	❑	❑	❑	❑	❑	❑
Ik zal actief op zoek gaan naar dit product om het aan te kopen	❑	❑	❑	❑	❑	❑
Vergeleken met andere producten is dit product belangrijk voor mij	❑	❑	❑	❑	❑	❑
Ik ben niet geïnteresseerd in dit product	❑	❑	❑	❑	❑	❑
Als ik een merk koop uit die productcategorie, kies ik heel zorgvuldig	❑	❑	❑	❑	❑	❑
Ik hecht belang aan dit product	❑	❑	❑	❑	❑	❑
Ik hou van dit merk	❑	❑	❑	❑	❑	❑
Ik vind het een zeer slecht merk	❑	❑	❑	❑	❑	❑
Ik zou het een merk aan anderen aanraden	❑	❑	❑	❑	❑	❑
Ik denk dat het één van de beste merken uit zijn productklasse is	❑	❑	❑	❑	❑	❑
Dit merk is echt iets voor mij	❑	❑	❑	❑	❑	❑
In geen geval zou ik dit merk gebruiken	❑	❑	❑	❑	❑	❑

In welke mate gaat u akkoord met volgende stellingen?

	Helemaal niet akkoord	Niet akkoord	Eerder niet akkoord	Eerder akkoord	Akkoord	Helemaal akkoord
Ik ben er gerust in dat ik technologie-gerelateerde vaardigheden kan aanleren	❑	❑	❑	❑	❑	❑
Ik winkel omdat dingen kopen me gelukkig maakt	❑	❑	❑	❑	❑	❑
Menselijk contact bij het verlenen van diensten maakt het proces prettig voor de consument	❑	❑	❑	❑	❑	❑
Ik vind het heel belangrijk om het boodschappen doen goed te organiseren	❑	❑	❑	❑	❑	❑
Ik koop geen producten die overdreven verpakt zijn	❑	❑	❑	❑	❑	❑
De zaken die ik bezit zijn niet zo erg belangrijk voor mij	❑	❑	❑	❑	❑	❑
Ik beschouw mezelf als een merkentrouwe consument	❑	❑	❑	❑	❑	❑
Ik kleed me vaak op een manier die tegen de stroom ingaat, zelfs al zijn anderen daardoor verontwaardigt	❑	❑	❑	❑	❑	❑
In een groep mensen ben ik zelden het middelpunt van de belangstelling	❑	❑	❑	❑	❑	❑
Ik vind dat een geordend en regelmatig leven bij mij aard past	❑	❑	❑	❑	❑	❑
Luchtvervuiling is een belangrijk wereldwijd probleem	❑	❑	❑	❑	❑	❑
Een buitenshuis werkende vrouw met jonge kinderen is nog steeds een goed moeder	❑	❑	❑	❑	❑	❑
In het algemeen vind ik dat ik erg gelukkig ben	❑	❑	❑	❑	❑	❑
Financiële zekerheid is erg belangrijk voor me	❑	❑	❑	❑	❑	❑
TV-kijken is mijn belangrijkste vorm van ontspanning	❑	❑	❑	❑	❑	❑
We ervaren een achteruitgang in de levenskwaliteit	❑	❑	❑	❑	❑	❑
Ik voel me vaak misleid door reclame	❑	❑	❑	❑	❑	❑
In de winkel het prijsetiket van een product veranderen, vind ik volstrekt ontoelaatbaar	❑	❑	❑	❑	❑	❑
Ik ben erg met mijn gezondheid begaan	❑	❑	❑	❑	❑	❑
Ik heb het gevoel voortdurend in tijdnood te zijn	❑	❑	❑	❑	❑	❑

7-point All labeled

Volgende uitspraken hebben betrekking op het product, op de productcategorie of op het merk. In welke mate gaat u akkoord met deze stellingen?

	Helemaal niet akkoord	Niet akkoord	Eerder niet akkoord	Neutraal	Eerder akkoord	Akkoord	Helemaal akkoord
Ik zou dit product willen proberen	❑	❑	❑	❑	❑	❑	❑
De volgende keer dat ik een aankoop doe in deze productcategorie, zal ik het getoonde product overwegen	❑	❑	❑	❑	❑	❑	❑
Ik zal actief op zoek gaan naar dit product om het aan te kopen	❑	❑	❑	❑	❑	❑	❑
Vergeleken met andere producten is dit product belangrijk voor mij	❑	❑	❑	❑	❑	❑	❑
Ik ben niet geïnteresseerd in dit product	❑	❑	❑	❑	❑	❑	❑
Als ik een merk koop uit die productcategorie, kies ik heel zorgvuldig	❑	❑	❑	❑	❑	❑	❑
Ik hecht belang aan dit product	❑	❑	❑	❑	❑	❑	❑
Ik hou van dit merk	❑	❑	❑	❑	❑	❑	❑
Ik vind het een zeer slecht merk	❑	❑	❑	❑	❑	❑	❑
Ik zou het een merk aan anderen aanraden	❑	❑	❑	❑	❑	❑	❑
Ik denk dat het één van de beste merken uit zijn productklasse is	❑	❑	❑	❑	❑	❑	❑
Dit merk is echt iets voor mij	❑	❑	❑	❑	❑	❑	❑
In geen geval zou ik dit merk gebruiken	❑	❑	❑	❑	❑	❑	❑

In welke mate gaat u akkoord met volgende stellingen?

	Helemaal niet akkoord	Niet akkoord	Eerder niet akkoord	Neutraal	Eerder akkoord	Akkoord	Helemaal akkoord
Ik ben er gerust in dat ik technologie-gerelateerde vaardigheden kan aanleren	❑	❑	❑	❑	❑	❑	❑
Ik winkel omdat dingen kopen me gelukkig maakt	❑	❑	❑	❑	❑	❑	❑
Menselijk contact bij het verlenen van diensten maakt het proces prettig voor de consument	❑	❑	❑	❑	❑	❑	❑
Ik vind het heel belangrijk om het boodschappen doen goed te organiseren	❑	❑	❑	❑	❑	❑	❑
Ik koop geen producten die overdreven verpakt zijn	❑	❑	❑	❑	❑	❑	❑
De zaken die ik bezit zijn niet zo erg belangrijk voor mij	❑	❑	❑	❑	❑	❑	❑
Ik beschouw mezelf als een merkentrouwe consument	❑	❑	❑	❑	❑	❑	❑
Ik kleed me vaak op een manier die tegen de stroom ingaat, zelfs al zijn anderen daardoor verontwaardigt	❑	❑	❑	❑	❑	❑	❑
In een groep mensen ben ik zelden het middelpunt van de belangstelling	❑	❑	❑	❑	❑	❑	❑
Ik vind dat een geordend en regelmatig leven bij mij aard past	❑	❑	❑	❑	❑	❑	❑
Luchtvervuiling is een belangrijk wereldwijd probleem	❑	❑	❑	❑	❑	❑	❑
Een buitenshuis werkende vrouw met jonge kinderen is nog steeds een goed moeder	❑	❑	❑	❑	❑	❑	❑
In het algemeen vind ik dat ik erg gelukkig ben	❑	❑	❑	❑	❑	❑	❑
Financiële zekerheid is erg belangrijk voor me	❑	❑	❑	❑	❑	❑	❑
TV-kijken is mijn belangrijkste vorm van ontspanning	❑	❑	❑	❑	❑	❑	❑
We ervaren een achteruitgang in de levenskwaliteit	❑	❑	❑	❑	❑	❑	❑
Ik voel me vaak misleid door reclame	❑	❑	❑	❑	❑	❑	❑
In de winkel het prijsetiket van een product veranderen, vind ik volstrekt ontoelaatbaar	❑	❑	❑	❑	❑	❑	❑
Ik ben erg met mijn gezondheid begaan	❑	❑	❑	❑	❑	❑	❑
Ik heb het gevoel voortdurend in tijdnood te zijn	❑	❑	❑	❑	❑	❑	❑

4-point Extreme labeled

Volgende uitspraken hebben betrekking op het product, op de productcategorie of op het merk. In welke mate gaat u akkoord met deze stellingen?

	Helemaal Niet Akkoord 1	2	3	Helemaal Akkoord 4
Ik zou dit product willen proberen	❑	❑	❑	❑
De volgende keer dat ik een aankoop doe in deze productcategorie, zal ik het getoonde product overwegen	❑	❑	❑	❑
Ik zal actief op zoek gaan naar dit product om het aan te kopen	❑	❑	❑	❑
Vergeleken met andere producten is dit product belangrijk voor mij	❑	❑	❑	❑
Ik ben niet geïnteresseerd in dit product	❑	❑	❑	❑
Als ik een merk koop uit die productcategorie, kies ik heel zorgvuldig	❑	❑	❑	❑
Ik hecht belang aan dit product	❑	❑	❑	❑
Ik hou van dit merk	❑	❑	❑	❑
Ik vind het een zeer slecht merk	❑	❑	❑	❑
Ik zou het een merk aan anderen aanraden	❑	❑	❑	❑
Ik denk dat het één van de beste merken uit zijn productklasse is	❑	❑	❑	❑
Dit merk is echt iets voor mij	❑	❑	❑	❑
In geen geval zou ik dit merk gebruiken	❑	❑	❑	❑

In welke mate gaat u akkoord met volgende stellingen?

	Helemaal Niet Akkoord 1	2	3	Helemaal Akkoord 4
Ik ben er gerust in dat ik technologie-gerelateerde vaardigheden kan aanleren	❑	❑	❑	❑
Ik winkel omdat dingen kopen me gelukkig maakt	❑	❑	❑	❑
Menselijk contact bij het verlenen van diensten maakt het proces prettig voor de consument	❑	❑	❑	❑
Ik vind het heel belangrijk om het boodschappen doen goed te organiseren	❑	❑	❑	❑
Ik koop geen producten die overdreven verpakt zijn	❑	❑	❑	❑
De zaken die ik bezit zijn niet zo erg belangrijk voor mij	❑	❑	❑	❑
Ik beschouw mezelf als een merkentrouwe consument	❑	❑	❑	❑
Ik kleed me vaak op een manier die tegen de stroom ingaat, zelfs al zijn anderen daardoor verontwaardigt	❑	❑	❑	❑
In een groep mensen ben ik zelden het middelpunt van de belangstelling	❑	❑	❑	❑
Ik vind dat een geordend en regelmatig leven bij mij aard past	❑	❑	❑	❑
Luchtvervuiling is een belangrijk wereldwijd probleem	❑	❑	❑	❑
Een buitenshuis werkende vrouw met jonge kinderen is nog steeds een goed moeder	❑	❑	❑	❑
In het algemeen vind ik dat ik erg gelukkig ben	❑	❑	❑	❑
Financiële zekerheid is erg belangrijk voor me	❑	❑	❑	❑
TV-kijken is mijn belangrijkste vorm van ontspanning	❑	❑	❑	❑
We ervaren een achteruitgang in de levenskwaliteit	❑	❑	❑	❑
Ik voel me vaak misleid door reclame	❑	❑	❑	❑
In de winkel het prijsetiket van een product veranderen, vind ik volstrekt ontoelaatbaar	❑	❑	❑	❑
Ik ben erg met mijn gezondheid begaan	❑	❑	❑	❑
Ik heb het gevoel voortdurend in tijdnood te zijn	❑	❑	❑	❑

5-point Extreme labeled

Volgende uitspraken hebben betrekking op het product, op de productcategorie of op het merk. In welke mate gaat u akkoord met deze stellingen?

	Helemaal Niet Akkoord 1	2	3	4	Helemaal Akkoord 5
Ik zou dit product willen proberen	❑	❑	❑	❑	❑
De volgende keer dat ik een aankoop doe in deze productcategorie, zal ik het getoonde product overwegen	❑	❑	❑	❑	❑
Ik zal actief op zoek gaan naar dit product om het aan te kopen	❑	❑	❑	❑	❑
Vergeleken met andere producten is dit product belangrijk voor mij	❑	❑	❑	❑	❑
Ik ben niet geïnteresseerd in dit product	❑	❑	❑	❑	❑
Als ik een merk koop uit die productcategorie, kies ik heel zorgvuldig	❑	❑	❑	❑	❑
Ik hecht belang aan dit product	❑	❑	❑	❑	❑
Ik hou van dit merk	❑	❑	❑	❑	❑
Ik vind het een zeer slecht merk	❑	❑	❑	❑	❑
Ik zou het een merk aan anderen aanraden	❑	❑	❑	❑	❑
Ik denk dat het één van de beste merken uit zijn productklasse is	❑	❑	❑	❑	❑
Dit merk is echt iets voor mij	❑	❑	❑	❑	❑
In geen geval zou ik dit merk gebruiken	❑	❑	❑	❑	❑

In welke mate gaat u akkoord met volgende stellingen?

	Helemaal Niet Akkoord 1	2	3	4	Helemaal Akkoord 5
Ik ben er gerust in dat ik technologie-gerelateerde vaardigheden kan aanleren	❑	❑	❑	❑	❑
Ik winkel omdat dingen kopen me gelukkig maakt	❑	❑	❑	❑	❑
Menselijk contact bij het verlenen van diensten maakt het proces prettig voor de consument	❑	❑	❑	❑	❑
Ik vind het heel belangrijk om het boodschappen doen goed te organiseren	❑	❑	❑	❑	❑
Ik koop geen producten die overdreven verpakt zijn	❑	❑	❑	❑	❑
De zaken die ik bezit zijn niet zo erg belangrijk voor mij	❑	❑	❑	❑	❑
Ik beschouw mezelf als een merkentrouwe consument	❑	❑	❑	❑	❑
Ik kleed me vaak op een manier die tegen de stroom ingaat, zelfs al zijn anderen daardoor verontwaardigt	❑	❑	❑	❑	❑
In een groep mensen ben ik zelden het middelpunt van de belangstelling	❑	❑	❑	❑	❑
Ik vind dat een geordend en regelmatig leven bij mij aard past	❑	❑	❑	❑	❑
Luchtvervuiling is een belangrijk wereldwijd probleem	❑	❑	❑	❑	❑
Een buitenshuis werkende vrouw met jonge kinderen is nog steeds een goed moeder	❑	❑	❑	❑	❑
In het algemeen vind ik dat ik erg gelukkig ben	❑	❑	❑	❑	❑
Financiële zekerheid is erg belangrijk voor me	❑	❑	❑	❑	❑
TV-kijken is mijn belangrijkste vorm van ontspanning	❑	❑	❑	❑	❑
We ervaren een achteruitgang in de levenskwaliteit	❑	❑	❑	❑	❑
Ik voel me vaak misleid door reclame	❑	❑	❑	❑	❑
In de winkel het prijsetiket van een product veranderen, vind ik volstrekt ontoelaatbaar	❑	❑	❑	❑	❑
Ik ben erg met mijn gezondheid begaan	❑	❑	❑	❑	❑
Ik heb het gevoel voortdurend in tijdnood te zijn	❑	❑	❑	❑	❑

6-point Extreme labeled

Volgende uitspraken hebben betrekking op het product, op de productcategorie of op het merk. In welke mate gaat u akkoord met deze stellingen?

	Helemaal Niet Akkoord 1	2	3	4	5	Helemaal Akkoord 6
Ik zou dit product willen proberen	❑	❑	❑	❑	❑	❑
De volgende keer dat ik een aankoop doe in deze productcategorie, zal ik het getoonde product overwegen	❑	❑	❑	❑	❑	❑
Ik zal actief op zoek gaan naar dit product om het aan te kopen	❑	❑	❑	❑	❑	❑
Vergeleken met andere producten is dit product belangrijk voor mij	❑	❑	❑	❑	❑	❑
Ik ben niet geïnteresseerd in dit product	❑	❑	❑	❑	❑	❑
Als ik een merk koop uit die productcategorie, kies ik heel zorgvuldig	❑	❑	❑	❑	❑	❑
Ik hecht belang aan dit product	❑	❑	❑	❑	❑	❑
Ik hou van dit merk	❑	❑	❑	❑	❑	❑
Ik vind het een zeer slecht merk	❑	❑	❑	❑	❑	❑
Ik zou het een merk aan anderen aanraden	❑	❑	❑	❑	❑	❑
Ik denk dat het één van de beste merken uit zijn productklasse is	❑	❑	❑	❑	❑	❑
Dit merk is echt iets voor mij	❑	❑	❑	❑	❑	❑
In geen geval zou ik dit merk gebruiken	❑	❑	❑	❑	❑	❑

In welke mate gaat u akkoord met volgende stellingen?

	Helemaal Niet Akkoord 1	2	3	4	5	Helemaal Akkoord 6
Ik ben er gerust in dat ik technologie-gerelateerde vaardigheden kan aanleren	❑	❑	❑	❑	❑	❑
Ik winkel omdat dingen kopen me gelukkig maakt	❑	❑	❑	❑	❑	❑
Menselijk contact bij het verlenen van diensten maakt het proces prettig voor de consument	❑	❑	❑	❑	❑	❑
Ik vind het heel belangrijk om het boodschappen doen goed te organiseren	❑	❑	❑	❑	❑	❑
Ik koop geen producten die overdreven verpakt zijn	❑	❑	❑	❑	❑	❑
De zaken die ik bezit zijn niet zo erg belangrijk voor mij	❑	❑	❑	❑	❑	❑
Ik beschouw mezelf als een merkentrouwe consument	❑	❑	❑	❑	❑	❑
Ik kleed me vaak op een manier die tegen de stroom ingaat, zelfs al zijn anderen daardoor verontwaardigt	❑	❑	❑	❑	❑	❑
In een groep mensen ben ik zelden het middelpunt van de belangstelling	❑	❑	❑	❑	❑	❑
Ik vind dat een geordend en regelmatig leven bij mij aard past	❑	❑	❑	❑	❑	❑
Luchtvervuiling is een belangrijk wereldwijd probleem	❑	❑	❑	❑	❑	❑
Een buitenshuis werkende vrouw met jonge kinderen is nog steeds een goed moeder	❑	❑	❑	❑	❑	❑
In het algemeen vind ik dat ik erg gelukkig ben	❑	❑	❑	❑	❑	❑
Financiële zekerheid is erg belangrijk voor me	❑	❑	❑	❑	❑	❑
TV-kijken is mijn belangrijkste vorm van ontspanning	❑	❑	❑	❑	❑	❑
We ervaren een achteruitgang in de levenskwaliteit	❑	❑	❑	❑	❑	❑
Ik voel me vaak misleid door reclame	❑	❑	❑	❑	❑	❑
In de winkel het prijsetiket van een product veranderen, vind ik volstrekt ontoelaatbaar	❑	❑	❑	❑	❑	❑
Ik ben erg met mijn gezondheid begaan	❑	❑	❑	❑	❑	❑
Ik heb het gevoel voortdurend in tijdnood te zijn	❑	❑	❑	❑	❑	❑

7-point Extreme labeled

Volgende uitspraken hebben betrekking op het product, op de productcategorie of op het merk. In welke mate gaat u akkoord met deze stellingen?

	Helemaal Niet Akkoord 1	2	3	4	5	6	Helemaal Akkoord 7
Ik zou dit product willen proberen	❑	❑	❑	❑	❑	❑	❑
De volgende keer dat ik een aankoop doe in deze productcategorie, zal ik het getoonde product overwegen	❑	❑	❑	❑	❑	❑	❑
Ik zal actief op zoek gaan naar dit product om het aan te kopen	❑	❑	❑	❑	❑	❑	❑
Vergeleken met andere producten is dit product belangrijk voor mij	❑	❑	❑	❑	❑	❑	❑
Ik ben niet geïnteresseerd in dit product	❑	❑	❑	❑	❑	❑	❑
Als ik een merk koop uit die productcategorie, kies ik heel zorgvuldig	❑	❑	❑	❑	❑	❑	❑
Ik hecht belang aan dit product	❑	❑	❑	❑	❑	❑	❑
Ik hou van dit merk	❑	❑	❑	❑	❑	❑	❑
Ik vind het een zeer slecht merk	❑	❑	❑	❑	❑	❑	❑
Ik zou het een merk aan anderen aanraden	❑	❑	❑	❑	❑	❑	❑
Ik denk dat het één van de beste merken uit zijn productklasse is	❑	❑	❑	❑	❑	❑	❑
Dit merk is echt iets voor mij	❑	❑	❑	❑	❑	❑	❑
In geen geval zou ik dit merk gebruiken	❑	❑	❑	❑	❑	❑	❑

In welke mate gaat u akkoord met volgende stellingen?

	Helemaal Niet Akkoord 1	2	3	4	5	6	Helemaal Akkoord 7
Ik ben er gerust in dat ik technologie-gerelateerde vaardigheden kan aanleren	❑	❑	❑	❑	❑	❑	❑
Ik winkel omdat dingen kopen me gelukkig maakt	❑	❑	❑	❑	❑	❑	❑
Menselijk contact bij het verlenen van diensten maakt het proces prettig voor de consument	❑	❑	❑	❑	❑	❑	❑
Ik vind het heel belangrijk om het boodschappen doen goed te organiseren	❑	❑	❑	❑	❑	❑	❑
Ik koop geen producten die overdreven verpakt zijn	❑	❑	❑	❑	❑	❑	❑
De zaken die ik bezit zijn niet zo erg belangrijk voor mij	❑	❑	❑	❑	❑	❑	❑
Ik beschouw mezelf als een merkentrouwe consument	❑	❑	❑	❑	❑	❑	❑
Ik kleed me vaak op een manier die tegen de stroom ingaat, zelfs al zijn anderen daardoor verontwaardigt	❑	❑	❑	❑	❑	❑	❑
In een groep mensen ben ik zelden het middelpunt van de belangstelling	❑	❑	❑	❑	❑	❑	❑
Ik vind dat een geordend en regelmatig leven bij mij aard past	❑	❑	❑	❑	❑	❑	❑
Luchtvervuiling is een belangrijk wereldwijd probleem	❑	❑	❑	❑	❑	❑	❑
Een buitenshuis werkende vrouw met jonge kinderen is nog steeds een goed moeder	❑	❑	❑	❑	❑	❑	❑
In het algemeen vind ik dat ik erg gelukkig ben	❑	❑	❑	❑	❑	❑	❑
Financiële zekerheid is erg belangrijk voor me	❑	❑	❑	❑	❑	❑	❑
TV-kijken is mijn belangrijkste vorm van ontspanning	❑	❑	❑	❑	❑	❑	❑
We ervaren een achteruitgang in de levenskwaliteit	❑	❑	❑	❑	❑	❑	❑
Ik voel me vaak misleid door reclame	❑	❑	❑	❑	❑	❑	❑
In de winkel het prijsetiket van een product veranderen, vind ik volstrekt ontoelaatbaar	❑	❑	❑	❑	❑	❑	❑
Ik ben erg met mijn gezondheid begaan	❑	❑	❑	❑	❑	❑	❑
Ik heb het gevoel voortdurend in tijdnood te zijn	❑	❑	❑	❑	❑	❑	❑

Uw profiel

Wat is uw geslacht?

- ❑ man
- ❑ vrouw

Tot welke leeftijdsgroep behoort u?

- ❑ minder dan 15 jaar
- ❑ 15 - 17 jaar
- ❑ 18 - 24 jaar
- ❑ 25 - 34 jaar
- ❑ 35 - 44 jaar
- ❑ 45 - 54 jaar
- ❑ 55 - 64 jaar
- ❑ 65 jaar of ouder

Wat is het hoogste diploma dat u behaalde?

- ❑ lager onderwijs
- ❑ lager middelbaar (ASO/VSO)
- ❑ hoger middelbaar (ASO/VSO)
- ❑ lager middelbaar beroeps/technisch
- ❑ hoger middelbaar beroeps/technisch
- ❑ hoger niet-universitair korte type
- ❑ hoger niet-universitair lange type
- ❑ universitair
- ❑ post-universitair

Sinds wanneer bent u actief op het Internet?

- ❑ geen idee
- ❑ voor 1998
- ❑ sinds 1998
- ❑ sinds 1999
- ❑ sinds 2000
- ❑ sinds 2001
- ❑ sinds 2002
- ❑ sinds 2003
- ❑ sinds 2004

Van welk type aansluiting thuis maakt u gebruik bij uw voornaamste provider?

- ❑ ik heb thuis geen internetaansluiting
- ❑ gratis aansluiting via modem (mijn telefoonkosten worden betaald door mijn werkgever)
- ❑ gratis aansluiting via modem (ik betaal uitsluitend telefoonkosten)
- ❑ betalende dial-in (ik betaal een bedrag aan mijn provider voor Internet toegang en extra telefoonkosten aan mijn telefoonoperator)
- ❑ betalende dial-in via een ISDN modem (ik betaal een bedrag aan mijn provider voor Internet toegang en extra telefoonkosten aan mijn telefoonoperator)
- ❑ Collecting abonnement - ik betaal één bedrag voor zowel Internet toegang als telefoonkosten
- ❑ Breedband light (Versatel, Telenet, Scarlet) = Een goedkopere breedband verbinding aan een lagere snelheid
- ❑ ADSL (Belgacom ADSL, Tiscali ADSL, ...)
- ❑ Kabel (Telenet,...)
- ❑ Ander
- ❑ Geen idee

Appendix IIa.2 Questionnaire Study 2

5-point Extreme labeled

How likely is it that you will do the following activities at least once during the next 2 weeks? Please indicate a number from 0% to 100%.

0% means 'definitely not' (i.e. there is no chance I will do this the next two weeks) and 100% means 'definitely will' (i.e. it is certain that I will do this activity in the next two weeks).

Go shopping	…. %
Go to a restaurant	…. %
Invite friends at my place	…. %
Attend a concert	…. %
Go for a walk	…. %
Go to the gym	…. %
Play computer game(s)	…. %
Communicate online with friends (chat, e-mail, facebook,…)	…. %
Go to the cinema	…. %
Go to a bar to have a drink with friends	…. %

Please indicate to what extent you (dis)agree with the following statements.

In general , I like to …	Strongly disagree				Strongly agree
… go shopping	0	0	0	0	0
… go to a restaurant	0	0	0	0	0
… invite friends at my place	0	0	0	0	0
… attend a concert	0	0	0	0	0
… go for a walk	0	0	0	0	0
… go to the gym	0	0	0	0	0
… play computer game(s)	0	0	0	0	0
… communicate online with friends (chat, e-mail, facebook,…)	0	0	0	0	0
… go to the cinema	0	0	0	0	0
… go to a bar to have a drink with friends	0	0	0	0	0

5-point ALL labeled

How likely is it that you will do the following activities at least once during the next 2 weeks? Please indicate a number from 0% to 100%.

0% means ‘definitely not’ (i.e. there is no chance I will do this the next two weeks) and 100% means ‘definitely will’ (i.e. it is certain that I will do this activity in the next two weeks).

Go shopping	…. %
Go to a restaurant	…. %
Invite friends at my place	…. %
Attend a concert	…. %
Go for a walk	…. %
Go to the gym	…. %
Play computer game(s)	…. %
Communicate online with friends (chat, e-mail, facebook,…)	…. %
Go to the cinema	…. %
Go to a bar to have a drink with friends	…. %

Please indicate to what extent you (dis)agree with the following statements.

In general , I like to …	Strongly disagree	Slightly disagree	Neutral	Slightly agree	Strongly agree
… go shopping	0	0	0	0	0
… go to a restaurant	0	0	0	0	0
… invite friends at my place	0	0	0	0	0
… attend a concert	0	0	0	0	0
… go for a walk	0	0	0	0	0
… go to the gym	0	0	0	0	0
… play computer game(s)	0	0	0	0	0
… communicate online with friends (chat, e-mail, facebook,…)	0	0	0	0	0
… go to the cinema	0	0	0	0	0
… go to a bar to have a drink with friends	0	0	0	0	0

Chapter IIb

Who said that looks do not matter?
The effects of scale format on response styles

IIb.1 Chapter outline

Questionnaires are an important source of data in marketing research. Unfortunately, survey data is often confounded by response styles such as acquiescence response style, disacquiescence response style, extreme response style and midpoint response style. Researchers can use different rating scale formats, which basically differ on two major dimensions, namely Polarity (unipolar versus bipolar) and Anchoring (only positive numbers or negative and positive numbers). To investigate which scale format performs best in terms of minimizing different response styles, we set up an experiment in which we manipulate Polarity and Anchoring. An online survey (N=337) shows strong effects of Polarity and Anchoring on response distributions and provides evidence for the superiority of the unipolar scale format with positive anchors.

IIb.2 Introduction

Survey data remain an important source of information in marketing research and has laid the basis for much of what we know about consumers' internal states and traits (like attitudes, intentions, values, etc.). The quality of survey data is therefore a major concern for marketing researchers. A rich stream of research has identified sources of method bias in survey data and has tried to come up with solutions for it (Baumgartner & Steenkamp, 2001; Podsakoff et al., 2003).

In this respect, Rindfleisch et al. (2008) recently pointed out that marketing researchers invest disproportionate efforts in solving method bias post hoc (i.e., after the data have been collected) by using statistical solutions. They call for more care during survey design to minimize method bias, rather than trying to cope with it once it arises. A specific decision that merits more consideration during survey design is the choice for a particular scale format (Podsakoff et al., 2003; Wong, Rindfleisch & Burroughs, 2003; Rindfleisch et al., 2008). Likert type scale formats have the advantage of being easy to construct, as each item consists of a single statement that needs to be rated in terms of agreement (Likert, 1932). Every Likert item taps into one pole of the underlying construct (e.g., "I like this product"), not both poles (I like – dislike the product). Concerning numbers used to anchor response categories, a Likert scale typically uses positive integer numbers only. An alternative measurement approach, gaining importance lately, offer respondents both poles of the construct in each item (Wong et al., 2003). A well-known format that uses this approach is the semantic differential (Osgood, Suci & Tannenbaum, 1957; Menezes & Elbert, 1979). In contrast to Likert scales, semantic differentials usually represent response categories by means of negative and positive numbers, centered around zero.

Despite the recommendation to use different scale formats like Likert scales and semantic differentials - possibly within the same questionnaire (Rindfleisch et al., 2008) - much remains to be learned about how respondents map their judgments onto different types of response scales. It is not clear to what extent one scale format shows a differential use of response categories than another scale format. In the current study, we compare unipolar (Likert) and bipolar (Semantic differential) scale formats in terms of the response frequencies of categories expressing agreement (acquiescence response style or ARS), disagreement (disacquiescence response style or DARS), extremity (extreme response style or ERS) and neutrality (midpoint response style or MRS). Moreover, we disentangle scale format (i.e. Polarity; unipolar vs. bipolar) and the numbers used to anchor response categories (only positive numbers or negative and positive numbers) by also studying the effect of Anchoring. Basically, these two dimensions cover most of the differences between scales (Dawis, 1987; Churchill & Iacobucci, 2005). Although it is common to use only positive numbers (e.g., from "1" to "7" on a 7-point scale) in Likert items and negative and positive numbers (e.g., from "-3" to "+3") in semantic differentials, in the current study we orthogonally manipulate Polarity and Anchoring to better understand their effect on the way respondents map judgments on scales.

We expect that both Polarity (unipolar versus bipolar) and Anchoring (positive numbers only versus positive and negative numbers, further also referred to as mixed options) affect the way respondents interpret response categories and subsequently map their judgments onto the available options (Tourangeau, Rips & Rasinski, 2000). Previous research has only partly addressed this issue and either focused on response styles in one specific format (e.g., Greenleaf, 1992a; Baumgartner & Steenkamp, 2001; Albaum et al., 2007; Weijters, Schillewaert & Geuens, 2008), or on the effect of a particular characteristic of a specific

format (e.g., the number of response categories in Likert items; Preston & Colman, 2000; Weathers, Sharma & Niedrich, 2005; or the inclusion of a midpoint in Likert items; Nowlis, Khan & Dhar, 2002; or both midpoint inclusion and number of response categories (Weijters, Cabooter & Schillewaert, 2010).

In the next section, we formulate hypotheses on the effect of scale format on response styles and set up an experiment in which we manipulate Polarity (unipolar / bipolar) and Anchoring (positive anchors / negative and positive anchors) of scale formats, and measure the resultant levels of ARS, DARS, ERS and MRS. We then present the findings of our experimental study with a total of 337 respondents. The experimental approach will allow us to formulate recommendations on the choice of an optimal scale format, and to draw preliminary conclusions on the comparability of data obtained by means of different scale formats.

IIb.3 Theoretical background and hypotheses development

IIb.3.1 Response styles

Response styles are defined as the tendency to respond in a systematic way to items independent of the content (Cronbach, 1950; Rorer, 1965; Paulhus, 1991). The central tendency of scale format measures is directly influenced by ARS and DARS (Baumgartner & Steenkamp, 2001). ARS concerns the extent to which respondents tend to show agreement, irrespective of content. DARS is the behavioral tendency to disagree with items. The spread of the observed scores is biased by the response styles ERS and MRS. ERS describes a general tendency of respondents to favor answering in extremes on scale formats (Cronbach, 1950; Greenleaf, 1992b). MRS is defined as the tendency to use the middle scale category

regardless of content. Response styles both affect reliability and validity of results (Baumgartner & Steenkamp, 2001; Arce-Ferrer, 2006).

IIb.3.2 Scale format and response styles

According to Tourangeau et al. (2000, 2004, 2007) and Schwarz et al. (1991), people rely on a range of cues (verbal, numerical and/or visual) to make inferences about the clarity and meaning of the response options in rating scales. Subsequently, these inferences affect the selection probability of each response option (Tourangeau et al., 2004). So, item ratings are a function of the interpretation of the rating scale options. For instancethe two contrasting poles of a construct, compared to only one pole, makes the meaning of the intermediate options clear thereby enhancing their chance of being selected (Arce-Ferrer, 2006; Schaeffer & Presser, 2003)

Our hypotheses rely on two main mechanisms through which the characteristics Polarity and Anchoring affect the clarity and meaning of the response options and consequently determine the level of response style bias. First, response rating scales differ in their perception of scale symmetry (Gannon & Ostrom 1996). Consequently, different rating scales may result in very different cognitive meanings and contents being retrieved from memory (Tourangeau & Rasinski, 1988). Second, response rating scales also vary in the extent to which the endpoint response options are interpreted as being conceptually further apart from each other. This effect is called the intensity effect and affects the obtained response distribution as the more intense the scale end points labels, the fewer respondents will choose those options (Judd & Harackiewicz, 1980; Wildt & Mazis, 1978). Both effects imply differences in the perceived meaning and salience of response options, thus changing the probability of them being selected (Arce-Ferrer, 2006; Schaeffer & Presser, 2003)

According to the scale symmetry mechanism, the presence of two contrasting poles of a construct determines whether a scale is perceived as symmetric. A study of Gannon and Ostrom (1996) provided empirical evidence for this process of categorical activation of two contrasting poles which results in symmetric rating scale perception. This means that for bipolar scales, which explicitely provide two poles, respondents activate both categories which each gives meaning for half of the rating scale response options. For instance, in the example of the opposite poles 'good' and 'bad' of the construct behavior, the response options on the left side of the rating scale (1, 2 and 3 on a 7-point scale) are gradations of 'bad', whereas the right hand side response categories (5, 6 and 7 on a 7-point scale) are gradations of 'good'. In addition, respondents also generate examples that best fit the explicit poles (Gannon & Ostrom, 1996). For instance, 'donating money for children in need' is a better fit for 'good' pole of behavior than 'helping an elder person to cross the street'. The best fitting examples, for the 'good' and for the 'bad' pole, represent the rating scale endpoints. The intermediate response options (2,3 for bad and 5,6 for good) are given meaning by determining the degree to which they are related to the best fitting category exemplar. For unipolar scales with positive numbers, two categories are activated as well, but these are not symmetrical. Here, participants use the explicitly activated category to correspond to the majority of the scale intervals; another not explicitly activated category is then used for the lower end (i.e., left side[16]) of the rating scale (Gannon & Ostrom, 1996; Mayo, Schul & Burnstein, 2004; Schwarz, 1999). Applied to our example, most response options in a unipolar scale will be interpreted in terms of their level of 'good' behavior, with the example of 'donating money for children in need' as best fitting exemplar. The interpretation of the implicit category used for the lower end points is left open to the

[16] In some countries (e.g., Germany) and in some studies, the lower end side of the scale is rather the higher end side of the scale (i.e., the right side).

respondent. For instance, the opposite of ‘good’ can be ‘not good’ (which is the absence of good) or ‘bad’ (which refers to failure of good behavior). The dominance of ‘not good’ versus ‘bad’ depends on the existence of a readily available schema for the inverse, e.g. ‘bad’. However, according to Mayo, Schul & Burnstein (2004), in unipolar scales the opposite will most often be interpreted as the absence of the explicit pole, in our example ‘not good’. As a result, compared to bipolar rating scales, a different knowledge structure is activated for the implicit pole in unipolar rating scales with positive anchors, even when the anchor is labeled as ‘strongly disagree’ (Mayo, Schul & Burnstein, 2004) . In contrast to a unipolar scale with positive anchors, a unipolar rating scale with mixed anchors, which provide a continuum from negative to positive response options, conveys a bipolar dimension in respondents’ minds (Gannon & Ostrom, 1996; Schwarz et al., 1991). It are the negative values on the left side of the rating scale that suggest that this side of the scale does not reflect the absence of the explicit pole, but the presence of the opposite of the explicit pole.

The second mechanism, also called the intensity effect, indicates that the poles of the rating scale are perceived more intensively which lead respondents away from the ends of the scale. So, the psychological width of the rating scale is defined by the extremity of the end point labels (Wildt & Mazis, 1978). Therefore, the meaning of the poles and the numeric values accompanying these poles can lead to differences in interpretation of the rating scale (Lam & Stevens, 1994; Schwarz et al., 1991; Wildt & Mazis, 1978). According to Judd & Harackiewicz (1980), the endpoints of a scale that differ on two dimensions, for instance value and sign, will be perceived more distinct from each other compared to a rating scale where the endpoints only differ on one dimension. For example, consider again the example of ‘good’ versus ‘bad’ above. If the endpoints ‘good’ and ‘bad’ differ on two dimensions, such as label and sign (i.e.,‘good +3’, ‘bad -3’), than the same level of good behavior should

lead respondents to select a lower number on a seven-point bipolar scale than in case of ‘good 7’ and ‘bad 1’. The reason for this is that the same experienced good behavior is lower relative to the more extreme anchor ‘good +3’ than to the less extreme anchor ‘good 7’.

As the most fundamental choice relates to the Polarity of the scale format, we first focus on the main effect of Polarity on response styles. Next to this main effect, we also discuss how the choice of anchors may alter the effects of Polarity, i.e. the possible interaction effect between Polarity and Anchoring on response styles.

IIb.3.2.1 Polarity

As stated before, the meaning of the response options of uni- and bipolar scales differs considerably. If a bipolar response format is offered, respondents activate two poles symmetrically and generate examples that best fit those explicit poles (Gannon & Ostrom, 1996). By retrieving examples from memory that define the endpoints, the endpoints of the scale become more salient and accessible, in turn leading to a higher probability of them being selected (Posavac, Sanbonmatsu & Fazio, 1997; Posavac, Herzenstein & Sanbonmatsu, 2003). In addition, as compared to unipolar scales, respondents will be more confident to use the extreme categories in bipolar rating scales since they have more information about the exact range of the rating scale to base their judgment on (Eiser & Osmon, 1978). Hence, we expect that the typical symmetric appearance of a bipolar scale may prompt a greater extreme response bias.

This dual poles activation for bipolar scales not only makes it easier to generate exemplars for the endpoint poles, it also means that the transition from the negative half to the positive half of the scale provides a salient anchor for the middle of the response scale compared to

unipolar scales where for most of the response options a single pole is activated. According to McCroskey et al. (1986), the midpoint in a bipolar scale is the point corresponding with the lowest intensity. Therefore, the midpoint on bipolar scales represents the midpoint (neither left nor right) in attitude. On the other hand, the meaning of the midpoint in a unipolar response format is not clear at all. Respondents can use this midpoint to either indicate moderate agreement, neutrality, ambivalence, or even a missing category (Gagné & Godin, 2000). Due to different meanings that can be attached to the midpoint in unipolar scales, we can expect a higher selection of this response option. Consequently, we hypothesize higher levels of MRS for unipolar scales compared with bipolar scales.

H1: Unipolar scales versus bipolar scales lead to lower levels of ERS.

H2: Unipolar scales versus bipolar scales lead to higher levels of MRS.

Another consequence of the symmetry in bipolar scales is that it makes the valence of the response options clear and salient, i.e. the left of the scale has a negative valence whereas the right of the scale has a positive valence[17]. As respondents have an inherent desirability to show agreeableness (Schuman & Presser, 1981; McClendon, 1991), the clarity and salience of the positive response options in bipolar scales (half of the response options situated on the right of the scale) could strengthen the felt pressure to agree. In case of a unipolar scale, the majority of the response categories represent gradations of the activated pole (the explicit pole), which is positive (Gannon & Ostrom, 1996). So, respondents can use more response options to show their agreeableness; this includes even options on the lower end side of the 7-point scale, namely options 3 and 4, which will lead to a decrease in ARS. On top of the fact that respondents will make more use of the lower end of the scale since some of these

[17] Although in some countries the opposite is true, namely negative valence for the right of the scale and positive valence for the response options on the left of the scale.

response options still have a positive connotation (e.g., option 3 on a 7-point scale), the lowest end point(s) of the unipolar rating scale (1 and possibly 2 on a 7-point scale) represent the contrast of the explicit pole and can have a different meaning for different respondents: 'the absence of the positive pole' (e.g., not good) or 'the opposite of the positive pole' (e.g., bad) (Mayo et al., 2004). The different meanings of the left points of the unipolar scale (e.g., 1, 2 and 3 on a 7-point scale) make the selection of these response categories more likely which increases DARS. Hence, we hypothesize that bipolar scales constitute higher levels of ARS compared to unipolar scales whereas the latter will enhance the level of DARS.

H3: Unipolar scales versus bipolar scales lead to lower levels of ARS.

H4: Unipolar scales versus bipolar scales lead to higher levels of DARS.

IIb.3.2.2 Polarity by Anchoring

The use of mixed anchors usually suggests that the rating scale is symmetric (Marsh & Parducci, 1978; Schwarz et al., 1991). The use of mixed anchors is as such informative (Tourangeau, Couper & Conrad, 2007) and can have an impact on the response distribution. For bipolar scales, the polarity is rather clear, although some respondents may remain uncertain about the intended range of the scale (e.g., how far apart the two ends of the scale are). The addition of positive and negative numbers makes the bipolar scale wider and the extreme response options more intense (Judd & Harackiewicz, 1980). Therefore, we expect a decrease of ERS for bipolar scales due to the influence of the intensity of the positive (+3) and negative anchors (-3). In terms of MRS, for bipolar scales, the presence of the natural midpoint already makes the midpoint clear even when this response option has the numerical value of 4 (on a 7-point scale). Therefore, we do not expect any differences in terms of MRS by changing the numerical anchors from positive to both positive and negative in bipolar scales.

Respondents who are faced with unipolar scales perceive the rating scale as asymmetric. Consequently, respondents will draw on the numeric values to disambiguate the meaning of the verbal endpoints and the other response options (Schwarz & Hipller, 1995; Schwarz et al., 1991) The inclusion of the zero point in the middle of the scale perceptually anchors the scale at three points (i.e., the midpoint as well as both endpoints). As a consequence, the positive and negative sides reflect symmetry around the midpoint (Marsh & Parducci, 1978). This contributes to the clarity of the middle response option, which can be expected to reduce MRS, since the midpoint will now only be chosen when one has an ambivalent or neutral judgment and not as mild agreement. When anchors are mixed, respondents also have more certainty about the exact range of the scale since it is clearer that the meaning of the implicit category signifies the contrast of the explicit category. The endpoints of the unipolar scale become more intense and are less likely to be chosen, according to the intensity effect. Therefore, we expect that the main effect of Polarity on ERS and MRS will be qualified by a significant Polarity by Anchoring effect in the sense that the differences in MRS between bipolar scales and unipolar scales reduce when mixed anchors are used. In terms of ERS we assume that the difference between bipolar scales and unipolar scales remain the same. Hence, we hypothesize:

H5a: Bipolar scales with positive and negative anchors lead to lower levels of ERS than bipolar scales with positive anchors.

H5b: Unipolar scales with positive and negative anchors lead to lower levels of ERS than unipolar scales with positive anchors.

H6: Unipolar scales with positive anchors lead to higher levels of MRS versus unipolar scales with positive and negative anchors, bipolar scales with positive anchors and bipolar scales with positive and negative anchors.

As stated earlier, for both bipolar rating scales, the position of the positive options as well as the negative options are clear because of the perception of symmetry. As a result, we do not expect to find any differences between both rating scales in terms of ARS or DARS.

For a unipolar scale with mixed anchors, respondents will eagerly use these numerical values to disambiguate the valence of each response option (Schwarz & Hippler, 1995). As a result, it becomes clear that options '-3' till '-1' have a negative connotation, the response options '+1' till '+3' a positive connotation and response option '0' signifies neutrality. The presence of the negative numbers will change respondents' judgment towards more positive response options since a downward shift in the meaning of the verbal end anchor results in an upward shift in the frequency ratings of the response options on the right side of the scale (Schaeffer & Presser, 2003). So through the clarity of the positive response options, their probability of being selected enhances. In contrast, respondents who are faced with a unipolar scale with positive anchors will make use of the majority of the response options to express agreement. Therefore, we can expect an increase in ARS compared to the unipolar scale with only positive response options. As a result, we hypothesize that the main effect of Polarity on ARS is qualified by a significant Polarity by Anchoring interaction. More specifically, we expect that the difference in ARS between bipolar and unipolar scales is attenuated when mixed anchors are used. For the lower end of a unipolar scale, the connotation is already negative, but not for response option '3' which has a positive connotation (Gannon & Ostrom, 1996). In contrast, for unipolar scales with mixed anchors, the third response option in row '-1' has a clear negative connotation. In addition, this rating scale is interpreted as more intense which likely leads to a reduction of the number of extreme response options including the negative response option '-3'. Consequently we do expect differences in terms of DARS for unipolar rating scales when the response options are mixed. Hence, we hypothesize:

H7a: Unipolar scales with positive and negative anchors versus unipolar scales with positive anchors will lead to higher levels of ARS.

H7b: Unipolar scales with positive and negative anchors versus unipolar scales with positive anchors will lead to lower levels of DARS

IIb.4 Methodology

IIb.4.1 Design

To test our hypotheses, we conducted an online survey in which we experimentally manipulated the rating scale characteristics Polarity (bipolar scale or unipolar scale) and Anchoring (only positive anchors or positive and negative anchors) (see Figure 1). The respondents were randomly assigned to one of the four conditions which resulted in the following cell counts: Unipolar-Positive anchors (N=78), Unipolar-Mixed anchors (N=75), Bipolar-Positive anchors (N=85), Bipolar-Mixed anchors (N=99). We used a 7-point rating scale with labeled endpoint response categories since this scale seems to score best in terms of reliability and discriminates well between the scale values (Cox, 1980; Krosnick & Fabrigar, 1997). The 7-point rating scale is also the most frequently used rating scale in marketing, based on an analysis of the rating scales used in the marketing scale inventory by Bruner et al. (2001), 55.2%, and research published in the International Journal of Research in Marketing, Journal of Consumer Research and Journal of Marketing Research between 2004 and 2009 (±43.1%).

FIGURE 1:

EXAMPLES OF THE DIFFERENT RESPONSE FORMATS USED

Unipolar - positive anchors	Good	Strongly Disagree 1	2	3	4	5	6	Strongly Agree 7	
Unipolar - positive and negative anchors	Good	Strongly Disagree -3	-2	-1	0	1	2	Strongly Agree 3	
Bipolar - positive anchors	Bad	1	2	3	4	5	6	7	Good
Bipolar - positive and negative anchors	Bad	-3	-2	-1	0	1	2	3	Good

IIb.4.2 Sample

The sample was randomly drawn from a panel which is representative for local Internet users. The sample comprised 337 respondents. Age ranged from 18 to 72 years with a median of 30. 59.9 % of respondents did not have any formal education after secondary school. 73% of the respondents were female.

IIb.4.3 Instrument

The questionnaire consisted of 52 heterogeneous items, designed to measure response styles, as well as an attitude and intention measure to be used for illustrative purposes. We used heterogeneous items to avoid confounding between content and style (Greenleaf, 1992b; Baumgartner & Steenkamp, 2001; Weijters et al., 2008). In particular, we randomly sampled the items by selecting only one item per unrelated marketing scale in Bearden and Netemeyer (1999) and Bruner, James and Hensel (2001). It is reasonable to assume that the scales from which the items are drawn have acceptable levels of discriminant validity, as all scales have been subjected to a thorough validation process (Bruner et al., 2001). We made sure that the contents of these items had no substantial true correlations. This is confirmed by the low inter-item correlations, ranging from .05 to .08 (average inter-item r = .06).

Deese's method (1965) was used to determine the antonyms for the statements. This method consists of two phases. During the first phase, 99 respondents were each shown 15 of 60 statements and were asked to formulate the statement's linguistic contrast. In the second phase, we asked a different sample of 87 respondents to formulate the linguistic contrast to the antonym given most frequently by the first group[18]. We dropped statements which did not

[18] Most frequently given antonym is the antonym that more than 80% of the respondents had given in the first group.

yield consistent and identifiable contrasts from the study. In total, we withdrew eight of the original 60 statements. Table 1 shows examples of the items used in the study.

TABLE 1:

EXAMPLES OF HETEROGENEOUS ITEMS AND THEIR SEMANTIC OPPOSITE

The work I do is valuable	The work I do is useless
A woman working out of home with children is a good mother	A woman working out of home with children is a bad mother
In general, strangers are reliable	In general, strangers are unreliable

To create operational measures of the response styles ARS, DARS, MRS and ERS, we used the formulas of Baumgartner and Steenkamp (2001) and applied them to the 52 heterogeneous items:

ARS= (#option 5*1 + #option 6*2 + #option 7*3))/ #statements,[19] (1)
DARS = (# option 1*3 + #option 2*2 + #option 3*1))/ #statements, (2)
ERS = (#option 1 + #option 7)/ #statements, (3)
MRS = (#option 4)/ #statements, (4)
# = represents the frequency of the response option

The second part of the questionnaire consisted of three attitude items on a 7-point scale and an intention question scaled on 100 points. Attitude items were measured on the same format as the 52 heterogeneous items. The (bipolar) attitude items included to illustrate the impact of scale format and response bias were "I find this product a good (bad) product (S1)", "I am convinced that this product is valuable (worthless) (S2)", "This product is appealing (not appealing) to me (S3)" (Bruner et al., 2001) (α = .80). The stimulus was a fictitious deodorant. The intention item was "How likely is it that you will buy this product in the

[19] Same results were obtained when we used non-weighted formulas of ARS and DARS.

future? Please indicate a number from 0% to 100%" and was kept constant across conditions. This question uses a specific format that has an objective meaning. Therefore, we assume that it does not share substantial method bias with the attitudinal measures which varies in scale formats (Greenleaf, 1992a). We designed this part of the questionnaire to show the effect of scale formats on response styles in a practical application.

IIb. 5 Results

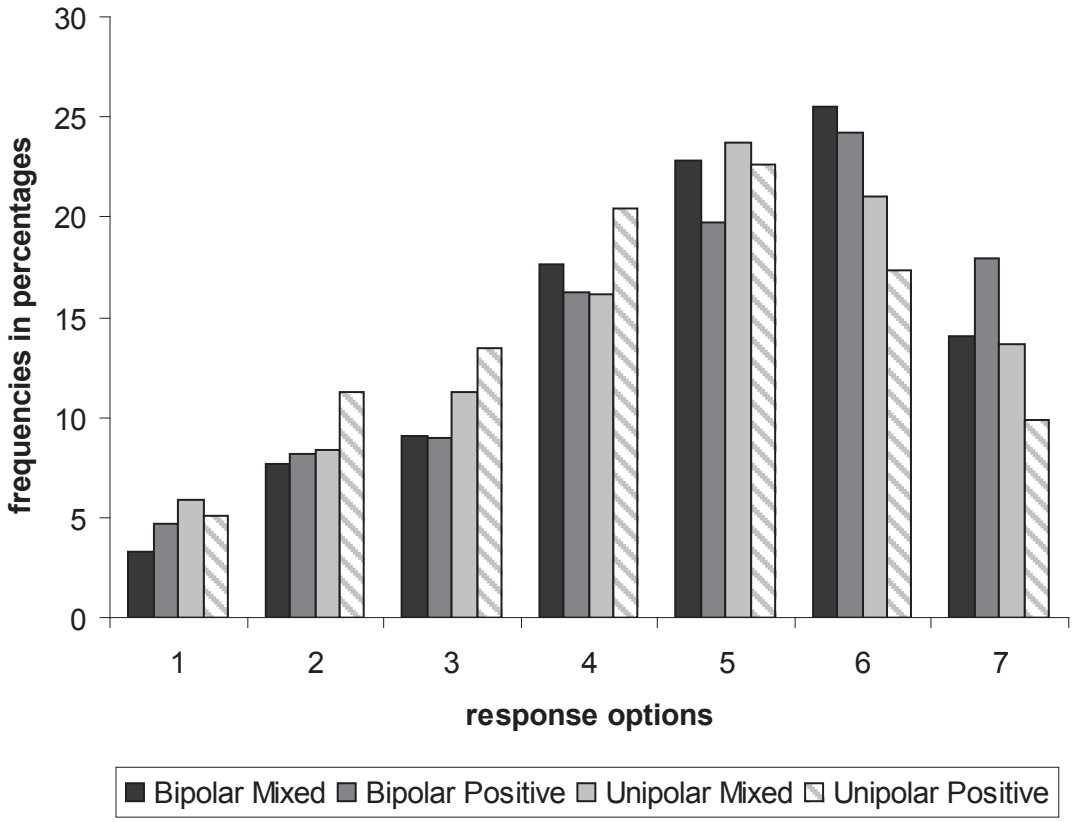
As can be seen in Figure 2, differences in rating rating scale do have an effect on the response distributions. In order to find out where the differences are situated, we test the hypotheses based upon a 2 x 2 Analysis of covariance (ANCOVA)[20][21], in which the two between-subjects variables were Polarity (bipolar vs. unipolar) and Anchoring (positive anchor vs. positive and negative anchors). The covariates in our model were age, gender and education. Each covariate has its own effect on response styles (Bachman & O'Malley, 1984; Hamilton, 1968; Greenleaf, 1992b; Krosnick & Fabrigar, 2003; Baumgartner & Steenkamp, 2001) and therefore we controlled for it (see additional analysis for further information on the covariates).

[20] The same results can be obtained by means of a Mancova

[21] The assumptions, for performing Ancova's have been tested. None of the assumptions were violated, that is the presence of homogeneous variances and normal distributions of the error term. The p-values for the Levene's tests were p=0.229 for ARS; p=0.139 for DARS; p=0.090 for ERS and p=0.906 for MRS.

FIGURE 2:

FREQUENCIES BY RESPONSE FORMAT CONDITION



The main effect of Polarity is significant for ERS ($F(1,330)= 3.970$, $p=0.047$), MRS ($F(1,330)= 9.246$, $p=0.002$), ARS ($F(1,330)= 48.984$, $p<0.001$) and DARS ($F(1,330)= 27.248$, $p<0.001$). The Polarity by Anchoring interaction is significant for ERS ($F(1,330)= 12.602$, $p<0.001$), MRS ($F(1,330)= 8.357$, $p=0.004$), ARS ($F(1,330)= 17.268$, $p<0.001$), but not for DARS ($F(1,330)= 0.009$, $p=0.924$). An overview of the different contrast effects on the interaction between Polarity and Anchoring can be found in Table 2.

TABLE 2:

OVERVIEW OF THE CONTRAST EFFECTS OF THE INTERACTION BETWEEN POLARITY AND ANCHORING

	ARS		ERS		MRS	
	F	p-value	F	p-value	F	p-value
Bipolar mixed - Bipolar positive	2.073	.151	7.758	.006	1.097	.296
Bipolar mixed - Unipolar mixed	4.201	.041	1.264	.265	1.168	.281
Bipolar mixed - Unipolar positive	44.142	.000	1.605	.206	4.234	.040
Bipolar positive - Unipolar mixed	10.951	.000	2.364	.125	0.003	.956
Bipolar positive - Unipolar positive	60.320	.000	14.879	.000	8.849	.003
Unipolar mixed - Unipolar positive	18.533	.000	5.032	.026	8.718	.003

IIb.5.1 Polarity

Polarity significantly influences ERS and MRS. Bipolar scales lead to more use of ERS ($M_{bipolar}$=0.200 vs. $M_{unipolar}$=0.172), whereas unipolar scales show higher levels of MRS ($M_{bipolar}$=0.164 vs. $M_{unipolar}$=0.187), supporting H1 & H2. Also for ARS and DARS, there is a significant main effect of Polarity. Whereas bipolar scales (M=1.189) lead to higher levels of ARS than unipolar scales (M=0.968), the opposite is true for DARS ($M_{bipolar}$=0.370 vs. $M_{unipolar}$=0.483), providing evidence in support of H3 and H4.

IIb.5.2 Polarity By Anchoring

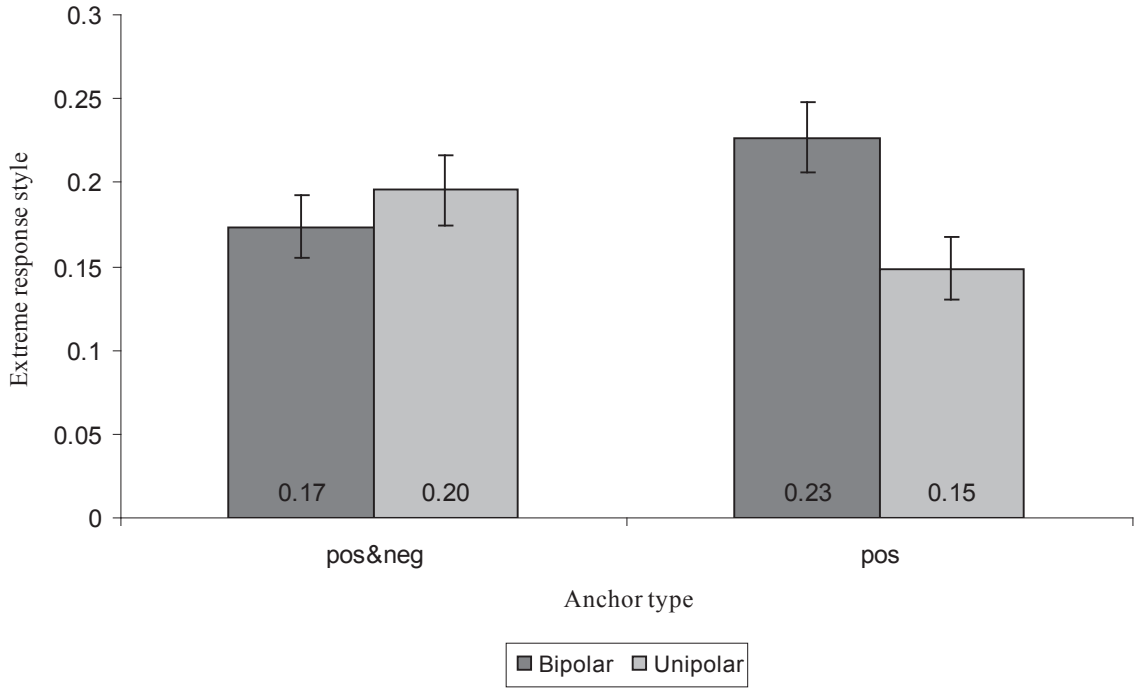
The effect of polarity is moderated by the anchors for both ERS and MRS. Bipolar scales only lead to more ERS than unipolar scales when positive anchors are used, but this difference disappears in case of mixed anchors. Apparently, the intensity effect of positive and negative numbers in symmetric scales like bipolar scales lowers the level of ERS. However, this intensity effect is not present in unipolar scales. There, the presence of mixed anchors leads to an increase in ERS. So, although the range of the scale becomes more extreme, the mixed anchors provide respondents information about the exact range of the rating scale. As a result, respondents will generate examples for both endpoints and the endpoints will become more salient and accessible, which results in a higher selection. In addition, respondents become more confident in selecting the extreme response options when the range of the scale is clear (Eiser & Osmon, 1978).

More specifically, unipolar scales with positive anchors (M=0.149) show lower levels of ERS compared to unipolar scales with mixed anchors (M=0.195), not supporting H5b. Bipolar scales with positive anchors (M=0.227) lead to higher levels of ERS compared to bipolar scales with mixed anchors (M=0.173), supporting H5a (see Figure 3). On the other hand, the unipolar scale with positive anchors (M=0.202) enhances the level of MRS compared to the

three other scale formats ($M_{bipolarpos}$=0.169 vs. $M_{bipolarpos\&neg}$=0.158 vs. $M_{unipolarpos\&neg}$=0.173), providing evidence for H6 (see Figure 4). So, bipolar scales lead to less MRS than unipolar scales, but again only when positive anchors are used and not when mixed anchors are used. For an overview of the hypotheses see Table 3.

FIGURE 3:

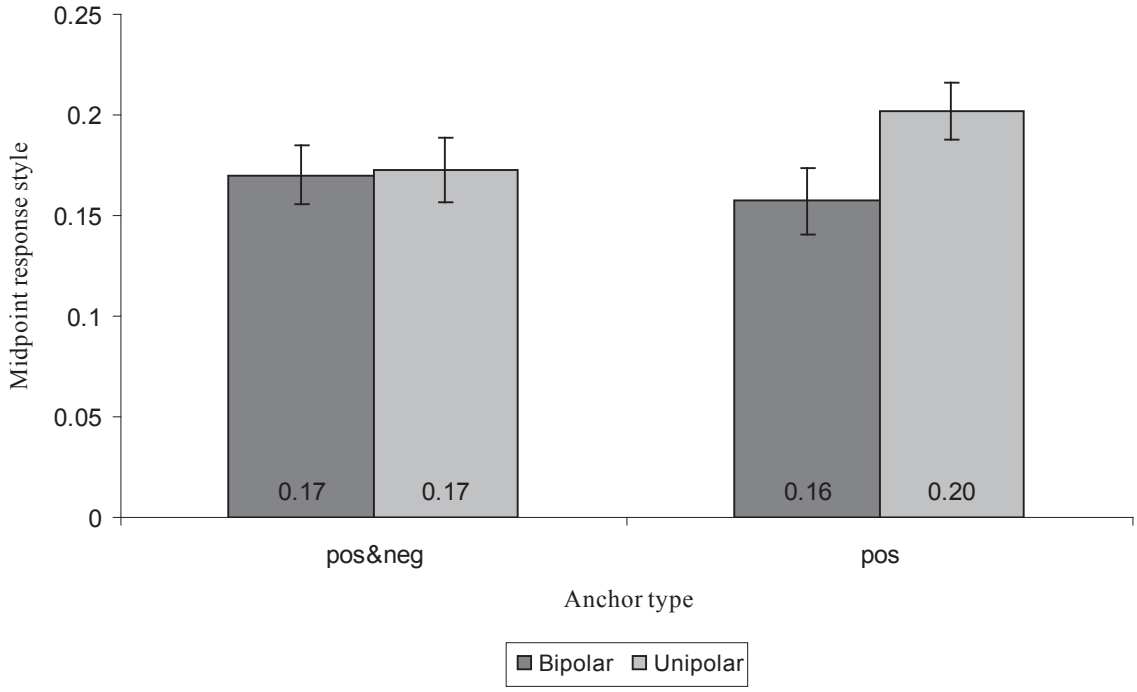
EXTREME RESPONSE STYLE BY RESPONSE FORMAT CONDITION



The bars represent the standard error (±SE) of ERS for each rating scale.

FIGURE 4:

MIDPOINT RESPONSE STYLE BY RESPONSE FORMAT CONDITION



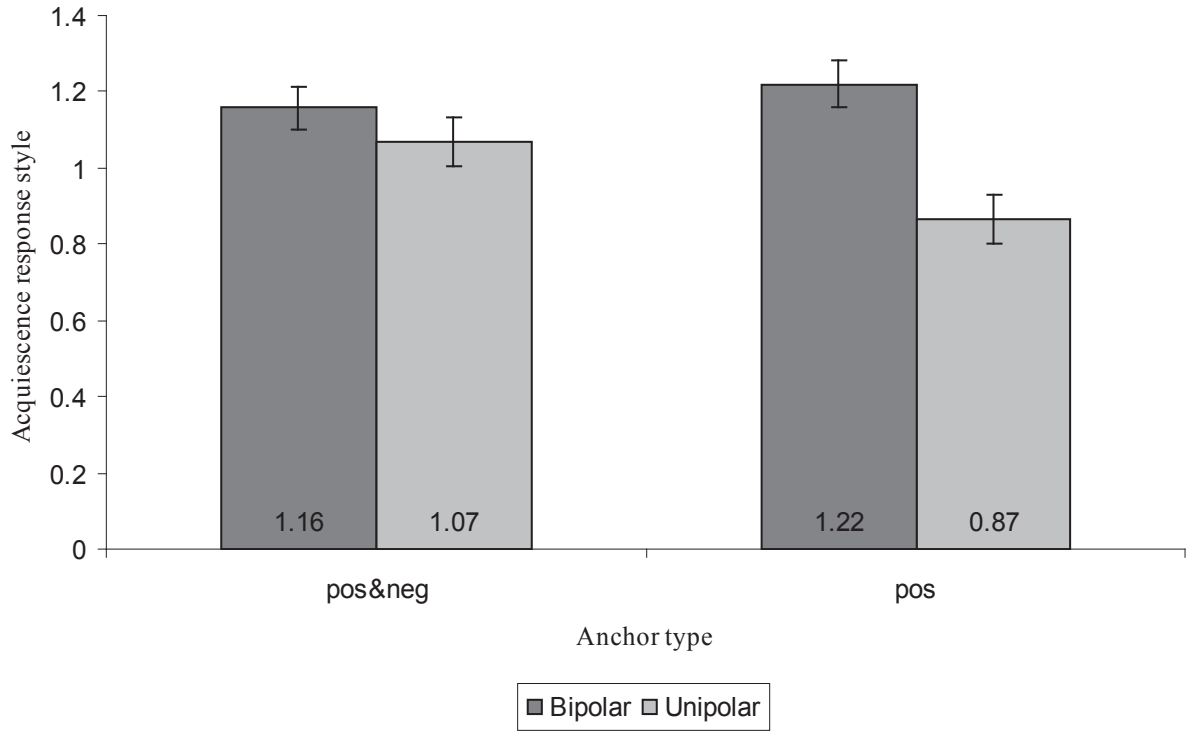
The bars represent the standard error (±SE) of MRS for each rating scale.

The main effect of Polarity on ARS is also qualified by a significant Polarity x Anchoring interaction effect in the sense that the differences between bipolar and unipolar scales are much more pronounced in case of positive anchors as compared to mixed anchors. More precisely, bipolar scales with mixed anchors (M=1.158) and bipolar scales with positive anchors (M=1.220) lead to higher ARS compared to unipolar scales with mixed anchors (M=1.068) and unipolar scales with positive anchors (M=0.876). The bipolar rating scales do not differ in terms of ARS nor in terms of DARS. The unipolar scale with positive anchors also lead to less use of ARS compared with unipolar scales with mixed anchors, lending support for H7a (see Figure 5). The interaction between Polarity and Anchoring was not

significant for DARS, which is not in line with hypothesis H7b. We expected a reduction in DARS due to the intensity effect and due to a reduction for the response option '-1', which has a negative connotation, in unipolar scales with mixed anchors compared to the response option '3', which still has a positive connotation, for unipolar scales with positive anchors. Whereas the latter did appear (reduction in response option 3) (see Figure 2), the intensity effect did not. Consequently, there is no significant difference in terms of DARS between unipolar rating scales.

FIGURE 5:

ACQUIESCENCE RESPONSE STYLE BY RESPONSE FORMAT CONDITION



The bars represent the standard error (±SE) of ARS for each rating scale.

TABLE 3:

OVERVIEW OF THE HYPOTHESES

Hypothesis	Test	Decision
H1: Unipolar scales versus Bipolar scales lead to lower levels of ERS	F=3.970, p=0.047	Support
H2: Unipolar scales versus Bipolar scales lead to higher levels of MRS	F=9.246, p=0.002	Support
H3: Unipolar scales versus Bipolar scales lead to lower levels of ARS	F=48.984, p<0.001	Support
H4: Unipolar scales versus Bipolar scales lead to higher levels of DARS	F=27.248, p<0.001	Support
H5a: Bipolar scales with positive and negative anchors lead to lower levels of ERS than Bipolar scales with positive anchors	F=7.758, p=0.006	Support
H5b: Unipolar scales with positive and negative anchors lead to lower levels of ERS than Unipolar scales with positive anchors	F=5.032, p=0.026	No Support
H6: Unipolar scales with positive anchors lead to higher levels of MRS versus		
Unipolar scales with positive and negative anchors	F=8.718, p=0.003	Support
Bipolar scales with positive anchors	F=8.849, p=0.003	Support
Bipolar scales with positive and negative anchors	F=4.234, p=0.040	Support
H7a: Unipolar scales with positive and negative anchors versus Unipolar scales with positive anchors will lead to lower levels of ARS	F=18.533, p<0.001	Support
H7b: Unipolar scales with positive and negative anchors versus Unipolar scales with positive anchors will lead to lower levels of DARS	F=3.054, p=0.081	No support

IIb.5.3 Additional analysis: impact of socio-demographics

Since gender, age and education have an impact on response styles (Greenleaf, 1992b; Krosnick & Fabrigar, 2003), we need to further investigate whether the significant effects of scale formats on response styles depend on socio-demographics; that is, do scale formats induce a response bias for specific demographic segments.

Earlier findings on demographics and response styles are not always consistent (see demographics in Chapter I).

We specified a path-model in Amos 17.0 in which ARS, DARS, ERS and MRS are the dependent variables and gender, age and education the independent variables. Scale format is used as the grouping variable (4 groups). We defined education as a variable that takes on the value of 0 for lower education and a value of 1 for higher education. Gender is coded as binary with the value of 0 for male and the value of 1 for female, age is defined as a continuous variable. We specified the response styles ARS, DARS, MRS and ERS as observed variables. We tested this model in AMOS 17.0 using multiple group path-analyses. For the hypothesis tests, we report p-values with alpha set to 0.05 as the threshold for significance.

We test invariance restrictions against the unconstrained model. The invariance restrictions test the hypotheses that parameter estimates are the same in the four conditions (Polarity x Anchoring). Since the reference model is saturated, we use the structural weights model to test the null hypothesis that the effects of the socio-demographics on response styles are equal across the four experimental conditions, this hypothesis is accepted and the fit of the model is good ($chi^2(36)=40.900$, $p = 0.264$; CFI= 0.992; TLI= 0.980; RMSEA= 0.027).

Overall, this means that the differences found between scale formats are not related with specific demographic segments[22]. Looking at the weights (see Table 4), we find a consistent significant effect between education and ERS/ARS, i.e. the higher educated respondent answers less in terms of ERS and ARS. So, the effect of education on response styles can not be attributed to differences in scale format. In addition, the effect of education on response styles is quite robust. Other significant differences found in demographics are that female versus male respondents answer more in terms of ARS and less in terms of DARS. No significant differences in gender on ERS or MRS are found. Some of our findings conflicts with earlier findings. This can perhaps be related to suboptimal measures but may also be explained by differences in the operationalization employed compared to our study such as the use of online questionnaires, even-point scales, fully labeled scales, content-driven items and student samples. Our findings contribute to the literature on socio-demographics by providing evidence of socio-demographic effects on response styles that are stable and robust across different rating scales whereas earlier research could not be generalized.

[22] Using the structural weights model as reference model, we then test whether the response styles have equal intercepts across the conditions. The structural intercept model is non-invariant ($chi^2(48)=97.564$, $p < 0.001$; CFI= 0.914; TLI= 0.849; RMSEA= 0.075), indicating that there is a robust main effect of scale format on response styles which is another confirmation for our findings based on Ancova's.

Table 4:

BETA COEFFICIENTS OF EDUCATION, AGE AND GENDER ON RESPONSE STYLES

		B	S.E.	P-value
Education	ARS	-0.17	0.054	0.002
	DARS	0.06	0.034	0.094
	ERS	-0.07	0.024	0.003
	MRS	-0.01	0.016	0.627
Age	ARS	0.00	0.002	0.411
	DARS	0.00	0.001	0.610
	ERS	0.00	0.001	0.231
	MRS	0.00	0.001	0.429
Gender	ARS	0.13	0.053	0.013
	DARS	-0.06	0.033	0.065
	ERS	0.04	0.023	0.085
	MRS	-0.01	0.015	0.431

IIb.5.4 Additional analysis 2: Impact of format on attitude and intention measures

The Ancova's demonstrate that the scale format components Polarity and Anchoring affect ARS, DARS, ERS and MRS. In particular, this analysis revealed that both bipolar scales with mixed anchors or positive anchors and the unipolar scale with mixed anchors show more ERS than the unipolar scale with positive anchors. The latter shows more MRS compared to the other scale formats. Bipolar scales versus unipolar scales also show more ARS, whereas unipolar scales versus bipolar scales show more DARS. Based on these results it is not clear which scale format is most preferable. Therefore, we extend our analysis by relating an intention measure on a %-scale to attitudes measured on bipolar or unipolar scales that either have positive or mixed anchors. This will allow us to study how Polarity and Anchoring affect estimates in simple regression models of a type that is quite common in marketing research.

To test this, we ran a SEM in AMOS 17.0. We first verify that the intention measure is invariant across conditions in terms of intercepts and variances. This seems to be the case as the nested chi square invariance tests are all insignificant: $chi^2(3) = 1.235$, $p=0.745$ for the

intercepts, and chi²(3)= 0.363, p = 0.948 for the variances. Thus, any subsequent violation of cross-group invariance in the model can be attributed to the responses to the attitude measures. We expect that attitude measures will differ across conditions through the appearance of ARS/DARS and/or ERS/MRS in the model. The presence of ARS/DARS could translate in a shift in central tendency, or in other words, in different intercept terms (Cheung & Rensvold, 2000), which would suggest scalar non-invariance. Since the groups are randomly assigned, we do not expect differences in the true attitude but in the observed attitude. Therefore, in the models, we fix the variance of the attitude factor to 1 and the mean of the attitude factor to zero (McArdle & McDonald, 1984). This method allows the latent means and variances of the corresponding constructs to be freely estimated relative to the fixed mean and variance in the first group. It has the advantage above the classic marker method (where one of the indicators is fixed to be zero and the loading of the indicator is fixed to one), that the results are not dependent on the indicator that has been chosen as marker variable (Little, Slegers & Card, 2006). Metric non-invariance, i.e. differences in the measurement weights, suggests the existence of between-group ERS differences (Cheung & Rensvold, 2000; Little, 2000).

The model with fixed attitude factor variance and mean fits the data rather well (see Model with fixed Attitude factor mean and variance in Table 5) and we use this model as the reference model against which we test model invariance. In the attitude item intercepts model, the chi square difference test tests the null hypothesis that the 3 attitude items are equal across the four scale format conditions. This hypothesis is rejected (p=0.018). A subsequent test indicates that invariance is borderline accepted for the attitude item loadings (p=0.07) (Hu & Bentler, 1999) (also see Table 5). The model estimates for the parameters are shown in Table 6 and group differences in Table 7.

TABLE 5:

MODEL FIT INDICES FOR INVARIANCE TESTS BETWEEN THE FOUR CONDITIONS

		Chi² test			Chi² difference test		
	Model	Chi²	DF	p	△Chi²	△DF	△P
A	Model with fixed Attitude factor mean and variance	20.18	8	0.010			
B	Attitude item loadings	64.41	17	0.005	32.63	9	0.070
C	Attitude item intercepts	80.59	17	0.001	48.81	9	0.018
D	Regression model	163.05	35	<0.001	131.27	23	<0.001

TABLE 6:

MODEL ESTIMATES BY CONDITION

Variable x Parameter	Bipolar Mixed	Bipolar Positive	Unipolar Mixed	Unipolar Positive
Intercepts (s.e.)	4.74 (0.1)	4.76 (0.1)	4.40 (0.2)	3.95 (0.2)
Standardized loadings (s.e.)	0.67 (0.1)	0.71 (0.1)	0.70 (0.2)	0.84 (0.2)
Average variance extracted	0.46	0.50	0.50	0.71
Composite reliability	0.71	0.75	0.74	0.88
R-square of intention	0.60	0.51	0.57	0.70

TABLE 7:

INTERGROUP DIFFERENCES

	Model	
	Attitude item intercepts	Attitude item loads
Bipolar pos - Bipolar mixed	$chi^2(3) = 1.43$; p=0.697	$chi^2(3) = 1.24$; p=0.744
Bipolar pos - Unipolar pos	$chi^2(3) = 13.2$; p=0.004	$chi^2(3) = 6.61$; p=0.09
Bipolar pos - Unipolar mixed	$chi^2(3) = 4.14$; p=0.247	$chi^2(3) =3.05$; p=0.384
Bipolar mixed - Unipolar pos	$chi^2(3) = 14.8$; p=0.002	$chi^2(3) = 9.73$; p=0.02
Bipolar mixed - Unipolar mixed	$chi^2(3) = 4.17$; p=0.244	$chi^2(3) =5.51$; p=0.138
Unipolar mixed -Unipolar pos	$chi^2(3) = 5.57$; p=0.135	$chi^2(3) = 3.07$; p=0.380

The measurement intercept model shows that higher intercepts (average intercepts of S1, S2 and S3) are obtained with bipolar scales compared to unipolar scale with positive anchors (see Table 7 and 8). The average standardized loadings are larger for the unipolar scale with positive anchors. This means that unipolar scales with positive anchors show lower ERS and ARS. Importantly, the explained variance, which indicates criterion validity, and the composite reliability score are higher for the unipolar scale with positive anchors. This indicates that the latter performs better than the other scale formats. In order to find out whether the problems in the measurement model spill-over in the regression model, we need to fix all measurement parameters (attitude means, attitude variances, attitude intercepts, factor means, and factor variances). The regression model is non-invariant ($p<0.05$) (also see regression model in Table 5). This means that differences in the measurement model translate into differences in the relationship between attitude and intention. To conclude, unipolar scale with positive anchors outperforms the other scale formats: the lower level of ARS and the higher reliability results in higher explained variance.

TABLE 8:

INTERCEPT DIFFERENCES BETWEEN THE CONDITIONS WITH REFERENCE GROUP: UNIPOLAR POSITIVE EXTREME

	Bipolar Mixed	Bipolar Positive	Unipolar Mixed	Unipolar Positive
S1	125%	121%	115%	100%
S2	115%	118%	104%	100%
S3	120%	122%	116%	100%
Average	120%	120%	111%	100%

IIb.6 Discussion

In this study, we experimentally manipulated the scale format of items, varying the scale format on two components: Polarity (unipolar versus bipolar) and Anchoring (positive valued anchors versus positive and negative valued anchors) (Schwarz et al., 1991; O'Muircheartaigh, Gaskell & Wright, 1995). We studied the effect of these manipulations on acquiescence response style (ARS), disacquiescence response style (DARS), extreme response style (ERS) and midpoint response style (MRS).

Our findings indicate that the scale format components Polarity and Anchoring have an important impact on response bias. In particular, the results of our study show that bipolar scales with positive anchors and the unipolar scale with mixed anchors show more ERS than the unipolar scale with positive anchors. The unipolar scale with positive anchors, in turn, shows more MRS compared to all the other scale formats. Bipolar scales versus unipolar scales, especially unipolar scales with positive anchors, also show more ARS, whereas unipolar scales versus bipolar scales show more DARS.

The mechanism behind these findings could be attributed to the interpretation respondents give to these different scale formats: for bipolar scales, two categories are symmetrically activated making the signal (positive, negative, extreme or neutral) of the different response categories more clear. In contrast, unipolar scales activate one pole for most of the response options, resulting in positive signals for the majority of the response categories. Our results also show that the intensity effect only occurs with bipolar rating scales not when the rating scale is unipolar. Apparently, for unipolar rating scales, the presence of mixed anchors provides extra information both about the range of the scale and the meaning of the implicit

pole. Consequently, the extreme response options become more salient and respondents will feel more confident in chosing those options.

Each scale produces different levels of measurement bias and has as such certain disadvantages and advantages compared to the other scales. A first additional analysis shows that the results described are robust and have no link with specific demographic segments. A second additional analysis shows higher criterion validity for the unipolar scale with positive anchors, meaning that this scale format provides better data for estimation of linear models.

The notion that response styles can be heightened through scale formats is an important message for the marketing community. Marketing scholars mostly have focused on eliminating the response style bias in data. The elimination techniques have proven to be very useful. However, revealing the relation between scale formats and response styles, contributes not only to the understanding of response styles, but also allows a better prediction of the extent to which response styles can affect data or can help researchers to avoid or minimize response styles upfront.

IIb.6.1 Implications

Researchers often forget to take into account that different formats can lead to different answers. However, our data clearly demonstrate that the relationship between variables, such as attitudes and intentions, can be significantly altered by response bias caused by the type of scale format. In our studies, we clearly indicate that the design of the format determines the pattern of response tendencies, potentially resulting in misleading conclusions.

An informal consensus seems to exist that a Likert format, or unipolar scale with positive anchors, is the most appropriate means of assessment. This scale is also one that is most frequently used. However, researchers have been using this scale format in the absence of evidence on its validity. Our study provides evidence for the superiority of the unipolar scale format with positive anchors (Likert scale) compared to the bipolar scales and the unipolar scale with mixed anchors. Therefore, our study is the first to provide empirical validation of the unipolar scale with positive anchors.

Nevertheless, the choice of a scale format should be based on the researchers' knowledge about the relevant response distribution of the sample. Suppose that the sample mainly consists of higher educated respondents, whom show less ARS and ERS, then a bipolar scale format would provide more meaningful response alternatives than a unipolar scale format. In contrast, if people are more likely to respond in terms of ARS than DARS, then a unipolar scale format with positive anchors will provide more meaningful response alternatives. A bipolar scale format can also be more appropriate (for reasons beyond measurement bias) when a researcher wants to be certain that the meaning conveyed by the scale format is clear and the same for both the respondent and the researcher. So, the choice for a scale format should be based upon the match between question interpretation and researcher interest. If one is interested in whether a respondent has experienced both failure and success, then two unipolar scale formats are more appropriate. On the other hand, if the researcher is interested in whether a respondent has more success than failure in life (or the other way round) than a bipolar scale format is more appropriate. Therefore, researchers need to consider which poles they intend to activate when selecting the scale format. However, when using scale formats, one need to bear in mind that response styles bias data. Therefore, one needs to correct afterwards for the presence of measurement bias.

Our study further contributes to the literature by fine-tuning the often suggested fit between polarity and anchors that a bipolar scale, for example, matches best with mixed anchors and a unipolar scale matches best with only positive anchors (Schwarz et al., 1991). For bipolar scales, this rule is not completely in line with our results. Both scale formats have identical effects on measurement bias except for ERS. This suggests that for bipolar scales words and numbers are less checked for consistency (O'Muircheartaigh et al., 1995). The rule is more applicable for unipolar scales than for bipolar scales. Unipolar scales with only positive anchors perform better in terms of measurement bias than unipolar scales with mixed anchors.

Our results also have important implications for questionnaire design. Some researchers advise, for example, to use heterogeneous scale formats to disrupt consistency biases and to increase validity (Rindfleisch et al., 2008). For example, they advise bipolar scales with mixed anchors for the independent variable and unipolar scales with positive anchors for the dependent variable. However, in this case the relation between the independent and dependent variable could be underestimated. On the other hand, when both variables are measured via the same scale format the relation between the variables could be overestimated. At first the option with different formats seems more logic. However, our study suggests that the second option, the same scale format, could be the better choice, if response styles are controlled for. In addition, our study indicates, through an invariant measurement model, that the results obtained from different scale formats (i.e., in meta-analyses) via between or within study designs cannot be compared due to the presence of response styles. When conducting, for instance, meta-analyses, it is important that one takes scale formats into account as a factor influencing estimates.

IIb.6.2 Limitations & future research

To conclude, we note some limitations of our study that offer opportunities for future research. An important limitation of the second additional analysis is the use of a self-report measure for assessing criterion validity. One might argue that there is a possibility that a 7-point unipolar scale with positive anchors is more similar to %-scales than bipolar scales. We admit this as a limitation and we recommend further research on this topic, possibly using other criterion variables (like third rater reports, for example). However, we have reasons to believe that the current empirical context makes the likelihood that the results are confounded rather small. The response formats (e.g., a 7 point unipolar scale with positive anchors versus % scale) are experienced differently by respondents, resulting in different response behavior and response quality. (Preston & Colman, 2000; Weathers et al., 2005) For the intention question, respondents had to fill out a percentage rather than choosing a response option from a given set. Thirdly, the observed difference in R^2 is large. In sum, we consider the use of a self-report for assessing criterion a limitation rather than a fatal flaw. Nevertheless, this additional analysis is a first, preliminary investigation into this topic, as surely, more research is needed before we can draw solid conclusions.

Second, in our study we only used a 7-point scale format. Future research might also examine the effects of Polarity and Anchoring in other formats, like the 5-point scale format. Although, we believe that the differences in ARS, DARS, MRS between a 7-point scale format and a 5-point scale format will be rather small since both scale formats lead to equal levels of observed response styles. Only the level of ERS is expected to be somewhat higher in a 5-point scale format (Weijters, Cabooter & Schillewaert, 2010).[23]

[23] This is under the condition that only the endpoints of the scale format are labeled and that there are no reversals among the items.

Third, concerning the effect of socio-demographics, we found some differences compared to earlier studies. It would therefore be interesting to perform a meta-analysis on the effect of socio-demographics on response styles while taking into account the scale format.

Fourth, since respondents also rely on visual cues, further research could also take these cues into account such as the orientation of the scale (i.e, vertical versus horizontal) (Mazaheri & Theuns, 2009) or the form of the scale (e.g., pyramid) (Tourangeau et al., 2007). According to Mazaheri and Theuns (2008), the orientation of the scale also determines the response distribution. However, they could only find effects when respondents rated dissatisfaction with life, not when the explicit pole was affirmative.

Chapter III

The effect of cognitive load on yeah-saying and nay-saying

III.1 Chapter outline

Respondents often fill out questionnaires under cognitive load (e.g., under time-pressure or while listening to music or watching TV). This paper investigates the impact of cognitive load on Net Acquiesence Response Style (NARS). In two studies we show that cognitive load increases the level of the Net Acquiesence Response Style when respondents are under high load but not when respondents perceive the cognitive load as moderate or low. To conclude, this paper shows the pervasive impact cognitive load can have on the validity of questionnaire responses.

III.2 Introduction

Respondents fill out questionnaires under different circumstances. Whereas researchers often assume that respondents, who fill out a questionnaire at home, do this undisturbed, in a quiet room without any distractions, this situation may well be the exception rather than the rule. Respondents may fill out a survey under time pressure, when doing multiple tasks at the same time (e.g., listening to music or talking to someone while filling out a questionnaire), while frequently being interrupted by e-mail notification, etcetera.

The latter situations all impose cognitive load on respondents. Cognitive load can be defined as a multidimensional construct representing the load that performing a particular task imposes on the respondent's cognitive system (Paas & van Merriënboer, 1994a). Researchers have long been intrigued by the effects of cognitive load on a host of consumer phenomena such as persuasion and decision making (e.g., Suri & Monroe, 2003; Svenson, Edland & Slovic, 1990; Zakay, 1990). However, a question that largely remains unanswered is whether cognitive load influences consumer responses to surveys? In other words, are response distributions sensitive to these situational conditions?

Baumgartner and Steenkamp (2001), Cronbach (1950) and Tversky & Kahneman (1974) suggest that the tendency to engage in stylistic responding can indeed be influenced by situational determinants such as cognitive load and called for further research on the matter. However, concrete evidence of response styles due to cognitive load remains scarce in the marketing literature. An important reason for this gap is that most research on response styles has focused on post-hoc elimination techniques (Baumgartner & Steenkamp, 2001; De Jong et al., 2008; Greenleaf, 1992b; Podsakoff et al., 2003; Wong, Rindfleisch & Burroughs, 2003)

and not on the cause of response styles. A notable exception is the study of Knowles and Condon (1999). Knowles and Condon (1999) found that under cognitive load, the level of the acquiescence response style increased. However, their findings are tentative due to some characteristics of the studies they report: (1) the focus is on binary scales, which are uncommon in marketing research, (2) they study a very specific measure of the acquiescence response style: their conceptualization of acquiescence is one of acquiescence-based misresponse (MR) (Swain et al., 2008) in which a respondent answers both positively on an affirmation and its negation. However, a growing body of evidence indicates that MR (acquiescence-based MR and/or disacquiescence-based MR) cannot be equated with NARS (Wong et al., 2003; Swain et al., 2008; Weijters et al., 2009). Finally, Knowles and Condon's methodology did not allow to disentangle the effects of content and style, which is a necessary prerequisite to draw valid conclusions on response styles (e.g., Andrews, 1984; Alwin & Krosnick, 1991; De Beuckelaer et al., 2009).

Given the fact that (1) preliminary results indicate that cognitive load could have a pervasive effect on the quality of survey data, that (2) for many academics and probably the majority of business surveys a high percentage of the sample may fill out the survey under cognitive load, that (3) cognitive load is a variable often manipulated in marketing experiments, and that (4) hardly any systematic research has been carried out on this issue, a rigorous investigation of the effects of cognitive load on response styles is called for. If cognitive load indeed induces response styles in typical marketing research questionnaires, then erroneous conclusions may have been drawn from a multitude of academic and business studies.

In this paper, we present two studies that provide evidence that a moderate level of cognitive load does not pose problems, but that high cognitive load significantly enhances

acquiescence. As a consequence, the results of several previous studies should be handled with care. Also, our findings can form an alternative explanation for cognitive load effects found in earlier studies.

III.3 Conceptual background

III.3.1 Response styles

Response styles are defined as tendencies to respond in a systematic way to items independent of the content (Paulhus, 1991). They are viewed as contaminating variables having undesirable effects on the reliability and validity of tests, which should therefore be controlled or eliminated from the test variance (Cronbach, 1946, 1950; Baumgartner & Steenkamp, 2001).

The most well known and widely studied response styles are the acquiescence response style and disacquiescence response style. The difference between these response styles is called Net acquiescence response styles (NARS) (e.g., Greenleaf, 1992a; Baumgartner & Steenkamp, 2001). This response style concerns the extent to which respondents systematically tend to show greater acquiescence (tendency to agree) rather than disacquiescence (tendency to disagree) with items, irrespective of content (Greenleaf, 1992a).

Researchers have agreed on the detrimental effects of NARS on data analysis and interpretation of results. NARS affects (1) the central tendency of rating scale measures (through biasing the intercept) (Greenleaf, 1992a; Rossi, Gilula & Allenby, 2001; Podsakoff et al., 2003) (2) scale reliability and validity (Cronbach, 1946), (3) the correlations between scales: the latter effect may result in biased estimates, factor analysis (i.e., it can lead to

factors composed exclusively of negatively worded items) and regression analysis (e.g., Chun et al., 1974; Lorr & Wunderlich, 1980; Heide & Grønhaug, 1992; Baumgartner & Steenkamp, 2001).

In summary, the above implications highlight the practical importance of studying NARS in marketing research. Because, it has been suggested that NARS is a cognitive style rather than a motivational one (McGee, 1967; Knowles & Nathan, 1997), cognitive load is a likely antecedent of NARS.

III.3.2 Cognitive load

Cognitive load is a cognitive variable that affects working memory capacity (Gilbert & Osbourne, 1989; Paulhus et al., 1989). Working memory can be broken down into three constructs: the central executive (i.e., the attentional capacity controller) and two subsystems: the visuo-spatial sketchpad (retention of color, shapes, dynamic information…) and the phonological loop (retention of speech-based information) (Baddeley, 2002). Each of these subsystems is competing for the attentional resources available to the individual. Therefore, placing additional load on one or both of these subsystems will result in impaired working memory and will consequently reduce the respondents' ability to perform the tasks.

Previous studies have induced cognitive load by placing respondents under time pressure or giving respondents a demanding secondary task that requires them to divide their attention (e.g., Svenson, Edland & Slovic, 1990; Dhar & Nowlis, 1999; Knowles & Condon, 1999; Paas et al., 2003; Lalwani, 2009). Time pressure, an often used cognitive load manipulation, creates a cognitive discrepancy between the time available and the time required to perform a task, this discrepancy is also called subjective time pressure. Time pressure in decision

making is mostly created by allowing individuals insufficient time to deliberate (Maule & Hockey, 1993; Suri & Monroe, 2003). Such instructions suggest that attention is divided between the passage of time and the decision process (Zakay, 1990).

Another manipulation of cognitive load is dual tasking, where respondents are requested to perform simultaneously two tasks (e.g., Festinger & Maccoby, 1964; Knowles & Condon, 1999; Lalwani, 2009). Here, the addition of a second task places the respondent under cognitive load since the second task will reduce the working memory capacity to attend to the primary task. The addition of for instance a computer game which burdens the visuo-sketchpad can reduce the cognitive capacity to respond to auditory questions which put a load on the phonological loop. Another example of dual tasking is requesting participants to indicate which song is played by a piano while they have to answer questions.

A number of studies (e.g., Simon, 1981; Kruglanski & Freund, 1983; Sanbonmatsu & Fazio, 1990; Dhar & Nowlis, 1999; Dhar, Nowlis & Sherman, 2000) on cognitive load find that under load persuasive messages become less effective, performance decreases, unique features become more important, respondents are less data driven, focus more on attitude-based decision strategies and are less likely to select the compromise option.

Possible explanations that account for these effects include the possibility (1) that under cognitive load less counterarguments are formed, (2) that cognitive load interferes with message comprehension, (3) that cognitive load limits the accessibility of information, (4) that respondents accelerate their pace and (5) that respondents use more simplifying heuristics such as non-compensatory rules, the brand-name heuristic or the price/quality heuristic (e.g.,

Festinger & Maccoby, 1964; McGuire, 1969; Wright, 1975; Payne, Bettman & Johnson, 1988; Edland & Svenson, 1993; Suri & Monroe, 2003).

The construct of cognitive load reflects both task characteristics and respondent characteristics (Paas & van Merriënboer, 1994a). Consequently, the presence of cognitive load per se does not imply that a consumer's ability to process information is limited (Lalwani, 2009; Suri & Monroe, 2003; Maule & Hockey, 1993). For instance, for some students studying and listening to music goes easily together, whereas for others the sound of a radio while studying can impose cognitive load. So, simply providing all respondents with the same level of cognitive load does not mean that they all perceive this load to a similar extent (Eysenck, 1983; Maule & Hockey, 1993). Cognitive load is internally, subjectively perceived and is not primarily based on the level of externally imposed load (Maule & Hackey, 1993). If a respondent considers the load of a task to be excessive he/she may behave as though he/she is overloaded, even though the task demands were objectively low. As a result, one should take into account the level of subjective load and not only the level of load that has been set. This has not been done in previous research.

III.3.3 Impact of Cognitive load on response styles

Answering a question consists of four stages: (1) understanding the question, (2) bringing to mind relevant information, (3) integrating this information into a judgment, and (4) selecting and reporting an answer (Tourangeau, 1984). Normally, for many attitude questions, respondents methodically employ these four steps (Eagly & Chaiken, 1993). However, attention is necessary for completing these mental processes while ignoring or filtering out other aspects of the environment/situation (e.g., Hunt & Ellis, 1999). When load is induced, respondents will use these situational cues to formulate their answer (Lalwani, 2009; Svenson & Edland, 1987; Trott & Jackson, 1967; Bettman & Johnson, 1988). Consequently, these

additional demands on attention will detract respondents from their initial task (Gilbert & Osborne, 1989; Paulhus et al., 1989). So, the exact set of response processes a respondent will carry out partly depends on the level of task load placed upon the respondent (Krosnick, 1991; Tourangeau et al., 2000).

Most cognitive load studies (e.g. Kruglanski & Freund, 1983; Sanbonmatsu & Fazio, 1990; Knowles & Condon, 1999) focused on two conditions of cognitive load: either no load or low load versus high load. However, based on the respondent's ability to cope with the level of task load, we can expect more than two levels of perceived cognitive load. One can expect that although respondents have been put under low, moderate or high load (externally imposed load), the perceived level of cognitive load for some is different (internally imposed load) than the load that was posed. There are also reasons to assume that respondents under moderate load will react differently compared to respondents under high perceived load (e.g. Tversky & Kahneman, 1974; Pham, 1996). When a moderate level of cognitive load is induced and perceived, respondents will probably react by accelerating their pace of responding (Edland & Svenson, 1993). Although these respondents will divide their attention across two tasks, the activation of alertness will stimulate them to process the questions systematically (Eysenck & Calvo, 1992). Therefore, we believe that they will use the same response process as respondents in the low load condition but at a faster rate.

However, when the constraint is more severe, such as under high load, and also perceived as more severe by the respondent, then acceleration will not be sufficient (Bettman, Johnson & Payne, 1990; Krosnick, 1991). These noticeable differences between moderate and high cognitive load will likely incite respondents to use heuristics to simplify the task (Dhar, Nowlis & Sherman, 2000; Kaplan, Whansula & Zanna, 1993; Tversky & Kahneman, 1974;

Wright, 1975). Consequently, respondents will eliminate parts of the four staged response process and are likely to automatically accept the statement (Gilbert, 1991), such as base their response on positive evaluations (Edland, 1993; Dhar, Nowlis & Sherman, 2000) since this heuristic requires few cognitive resources. Therefore, people are expected to answer more positively than negatively, or in other words to show higher levels of NARS, under perceptions of high cognitive load. Hence, we hypothesize:

> *H1: NARS increases under conditions of high cognitive load, but not under conditions of moderate cognitive load.*

We test our hypothesis in two studies. Study 1 focuses on cognitive load manipulated by means of time pressure; whereas Study 2 uses a dual task to replicate the findings of the first study.

All of our studies test the hypothesis that an externally imposed load leads to increases in NARS, but only if the perceived cognitive load is high (and not low to moderate) (H1). Different from former studies is that we determine the level of cognitive load based upon both the level of load externally placed upon the respondent and the level of load perceived by the respondent. We hereby refer to the subjective nature of cognitive load (Paas & van Merriënboer, 1994a; Maule & Hackey, 1993).

III.4 Study 1

III.4.1 Method

III.4.1.1 Participants and Design

One-hundred and fifteen undergraduate students (54.8% female) filled out the questionnaire to fulfill part of a research requirement. All participants were tested individually in a research room and the influence of other distracters was controlled for. While responding to the items of the questionnaire, some participants were placed under time pressure[24] (high cognitive load condition), whereas others were not (low cognitive load condition)[25].

III.4.1.2 Externally imposed cognitive load

At the start of the experiment, participants in the load condition were told that the experiment dealt with how people react under divided attention. We used time pressure as manipulation since it has been identified as an exogenous variable capable of influencing consumer behavior (e.g., Wright, 1975; Howerd & Seth, 1969). Based on pre-tests, time pressure levels were set at 4 minutes. Respondents received instructions that the time was shorter than usual (i.e., a pretest indicated that respondents needed on average 8 minutes to fill out the survey) but that it was still sufficient for completing the survey. This instruction automatically induces time pressure (cfr. Svenson & Benson, 1993). Further, each 30-second increment respondents were visually given feedback on the elapsed time until no time was left (Svenson & Benson, 1993). In the low load condition, respondents completed the questionnaire at own pace. Consequently, two load levels were externally imposed. However, in the analysis we take into

[24] Attentional load was used to manipulate cognitive load, as in numerous previous studies (e.g., Knowles & Condon, 1999; Suri & Monroe, 2003).

[25] Low time pressure is used instead of no time pressure since there is always an internal time clock that is in operation (MacGregor, 1993).

account the respondents' subjective load instead of the level of load that was posed on the respondent.

III.4.1.3 Perceived cognitive load

At the end of the experiment, respondents were asked to fill in the Task-Loading index (TLX) to assess perceived cognitive load (Hart & Staveland, 1988). It has been demonstrated that people are quite capable of giving a numerical indication of their perceived burden (Gopher & Braune, 1984) and subjective measures of task difficulty highly correlate with other load measures such as physiological techniques and performance based techniques (O'Donnell & Eggemeier, 1986)[26]. In addition, subjective scores may come closest to tapping the essence of perceived cognitive load and provide the most valid and practical indicator (Hart & Staveland, 1988). The TLX is a scale that provides a summary of the level of cognitive load perceived by the respondents. Respondents were asked to give a number from 0-100 on (1) how much mental activity the task required, (2) how much physical activity (e.g., pushing, pulling, …) the task required, (3) how much time pressure they felt due to the rate at which the task had to be fulfilled, (4) how successful they were in accomplishing the task, (5) how hard they had to work to accomplish the task and (6) how insecure, discouraged and stressed they felt during the task. The level of physical load was excluded for the calculation of TLX (α =.69) since this subscale was not relevant for this setting. Respondents reported that more cognitive effort was required to fulfill the questionnaire under external cognitive load than under low load (M_{low} = 30.59 vs M_{load}= 61.20; t(38.97) =-9.155, p<0.001). So, although it seems that the manipulation was successful, and that the level of cognitive load posed on the respondent was

[26] There are several reasons to believe that the TLX measure is response style free. First, respondents have to fill out a percentage rather than to pick an option from a given set. Second, the TLX correlates highly with physiological techniques. Since the latter are free of response styles, we expect that the former to also be free of response styles. However, we admit that this is a limitation and call for more research on this topic.

sufficient, we still expect different reactions based on the level of respondents' subjective load.

III.4.1.4 Dependent measure

The questionnaire consisted of items that were included with the specific aim of measuring NARS. In particular, we used sixty-nine heterogeneous items on a 7-point likert scale format, randomly selected from as many unrelated marketing scales in Bearden and Netemeyer (1999) and Bruner et al. (2001). To illustrate, the item sample included items like "I sometimes have the feeling that people use me", "Television is my primary form of entertainment" and "I am good at sports". Thus, we made sure that the contents of these items had no substantial true correlations. This was confirmed by the low inter-item correlations ($r_{average}$ = .05). This procedure guarantees that content and response styles are not confounded. To create a measure of NARS we used the formula from Greenleaf (1992a).

NARS = (# agreements - # disagreements), (1)
#agreements=(option 5*1, option 6*2, option 7*3)
disagreements=(option 1*3 , option 2*2, option 3*1)

Based on the respondent's subjective load (Zakay, 1990), we expect to find different results. To test this, we specify a regression model that explains the level of NARS as a function of subjective load.

$$NARS_j = \beta_0 + \beta_1 \times TLX_j + \beta_2 \times Z_j + \varepsilon_j,^{27} \quad (2)$$

Where $NARS_j$ is the level of NARS for a subject *j*.

[27] In a second analysis, we controlled for the manipulation of objective load. Objectiveloadj is a dummy variable with 0 referring to the low load condition and 1 referring to the load condition. Objective load did not have a significant effect on the level of NARS.

For the independent variables, we created a subjective load variable based on the TLX index of the respondent. To differentiate the effect of low and moderate load from the effect of high load, we used a spline regression specification with one breakpoint (Marsh & Cormier, 2002; Weijters et al., 2009)[28,29]. The breakpoint will be determined based on the data, as we discuss in the results section.

First, the initial TLX measure is included in the regression, TLX_j. Second, we create a dummy variable D_j based on the value of TLX. For D_j, values equal to or below the breakpoint are set to zero; values above the breakpoint are set to one. Using this dummy variable D_j, we create the corresponding spline adjustment variable Z_j as $Z_j = D_j*(TLX_j - \text{breakpoint})$. Whenever TLX_j is below the breakpoint, $D_j = 0$, so Z_j can never be negative. Thus, the intercept β_0 corresponds to the expected level of NARS at a subjective load equal to zero. Finally, the error term (ε_j) captures the variance in NARS that is not explained by the preceding variables.

[28] When a nonlinear relationship is expected between a dependent variable, NARS, and a independent variable, the respondent's subjective load, spline regression models are recommended (Marsh & Cormier, 2002; Rindskopf, 2003). More specifically, this kind of regression allows for changes in direction at special spline knots.

[29] The two breakpoint solution performed worse compared to the one-breakpoint analyisis. For the model with one and two breakpoints, the fit index BIC is respectively 741.54 versus 747.45.

III.4.2 Results

We ran a regression model using different break points for the spline regression. We varied the breakpoint between 30 and 70 (a lower level of TLX=30 would indicate low load, where we are interested in differences between moderate and high load). We compared the resulting R^2 values to determine the optimal point where the model captured the maximum amount of variance in the dependent variable. R^2 reached its optimal point when the breakpoint was at 55.

Table 1 lists the results of the regression analysis. The key assumptions of the multiple linear regression model were met: (1) all condition indices were below 10.6 and the tolerance values are higher than 0.10, and (2) the standardized residuals showed approximately normal distributions as confirmed by the nonparametric Kolmogorov-Smirnov test ($p>0.716$). As hypothesized, there were different effects of the independent variable on the level of NARS.

Results showed that perceived cognitive load induced by time pressure led to a status quo in the level of NARS across the moderate cognitive load group and the low cognitive load group. High perceived cognitive load significantly increased the level of NARS compared to the respondents who perceive the load low or moderately. For instance, the regression estimates indicate that on average, respondents with a TLX score of 75 have a level of NARS of 34, whereas respondents with a TLX score of 25 have a level of NARS of 8.

Together, these findings provide evidence for our hypothesis (H1). So, this confirms our expectation that cognitive load does not have a linear effect on NARS. To illustrate the effect, Figure 1 displays the different response distributions for the different conditions on one of the items. The ratings tend to fall on the negative side of the scale when the perceived load is low

or moderate but not when perceived load is high. So, differences in the level of cognitive load can affect response distributions and consequently bias results.

TABLE 1:

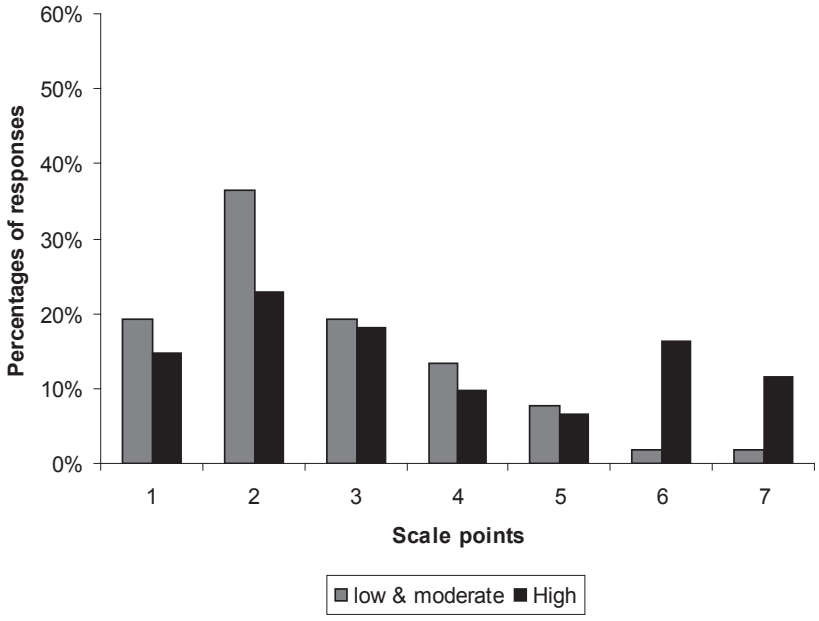
RESULTS OF REGRESSION ANALYSIS STUDY 1

R^2=.16	B	SE	*t*	p
Intercept	4.761	8.778	0.542	0.589
TLX_j	0.142	0.199	0.712	0.478
Z_j	0.931	0.429	2.166	0.032

FIGURE 1:

DISTRIBUTION OF RESPONSES ACROSS THE DIFFERENT CONDITIONS:

"I SOMETIMES HAVE THE FEELING THAT PEOPLE USE ME"



III.5 STUDY 2

In Study 2, we developed another manipulation of cognitive load to show that our findings are robust across different load manipulations. We developed a more subtle manipulation of a dual task where working memory is impaired by placing load on both subsystems: the visuo-spatial sketchpad (i.e., a ping-pong game) and the phonological loop (i.e., answering auditory questions) (Baddeley, 2002). By combining auditory questions with a visual second task, respondents must devote a smaller portion of their total capacity on each of the tasks. Consequently, we can assume that the respondents are under cognitive load.

III.5.1 Method

III.5.1.1 Participants and Design

Seventy-five undergraduate students (58.7% female) of a large European university participated in exchange for class credits. Respondents were invited to a research room in groups of 6. Each participant was assigned his or her own cubicle. At the end of the questionnaire, respondents were asked to complete the Task-Loading index (Hart & Staveland, 1988) (α =.66). Respondents reported that more cognitive effort was required to fulfill the questionnaire under cognitive load than under low load (M_{low} = 28.63 vs M_{load}= 55.13; $t(73) = -5.987$, $p<0.001$).

III.5.1.2 Cognitive load

Participants in the load condition were informed that their task consisted of answering auditory questions they received through headphones while playing a ping-pong game. So, the combination of these auditory questions and visual game lead to a conflict in the respondents' minds, i.e. both slave-systems are competing for attention. Respondents were instructed to list

their responses manually on a 7-point rating scale. The first 5 questions served as an exercise to make sure that every participant understood the instructions. Next, respondents answered thirty questions to measure NARS. The number of each question was preprinted and the questions were spread across three pages. As the items were randomly sampled from existing marketing scales, they were highly heterogeneous, and 30 items could be reasonably assumed to be sufficient to validly measure NARS (Greenleaf, 1992b; Weijters et al., 2008). The ping-pong game and the questionnaire with response buttons for the auditory questions were simultaneously presented to the respondents by using a split screen. The location of the game and questionnaire on the computer screen shifted from the right side to the left after every 10 questions.

In the low load condition, participants received auditory questions only. The auditory questions had a sequence of 5 seconds, which is sufficient since respondents typically take 5 seconds to answer a question (Basilli & Fletcher, 1991; Tourangeau, Rasinski & D'Andrade, 1991). We again tested the spline regression specification in formula (2)[30]. The specifications of the regression model can be found in the method section of study 1.

III.5.2 Results

We ran the regression model using breakpoints varying between 30 and 70. As in Study 1, R^2 reached its maximum when the breakpoint equaled 55. The key assumptions of the multiple linear regression model were met: (1) all condition indices were below 8.4 and the tolerance values are higher than 0.10, and (2) the standardized residuals showed approximately normal distributions as confirmed by the nonparametric Kolmogorov-Smirnov test ($p>0.949$).

[30] Also in this study, a two breakpoint solution performed worse compared to the one-breakpoint analyisis. The fit index BIC was lower for the one-breakpoint solution (357.5) versus the two-breakpoint solution (359.2).

Results showed that the group that perceived the cognitive load moderately did not differ in terms of NARS from the group that perceived low cognitive load and that the high perceived cognitive load group significantly increased the level of NARS compared to the low and moderate load, providing evidence for H1 (see Table 2). To illustrate the effect, respondents with a TLX score of 75 have a level of NARS around 10, whereas respondents with a TLX score of 25 show a level of NARS of 7. Figure 2 displays the effect for the different conditions on one of the items. Ratings tend to fall on the negative side of the scale when the perceived load is low or moderate but not when perceived load is high. Thus, Study 2 replicates the effects of cognitive load on the level of NARS that were identified in Study 1.

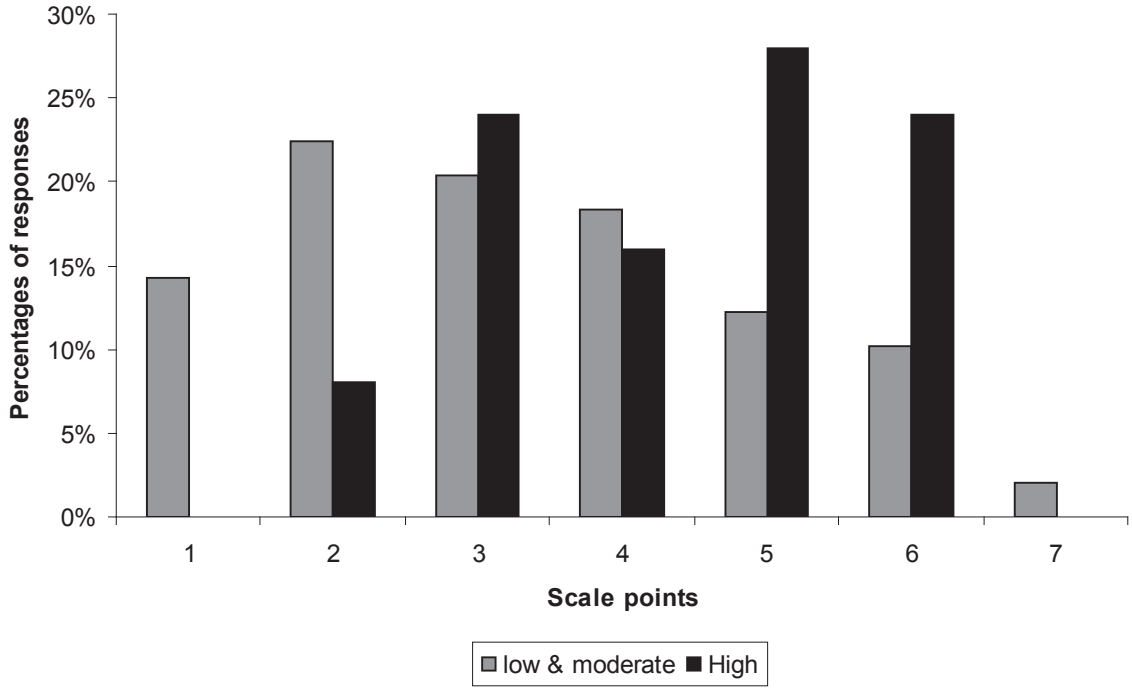
TABLE 2:
RESULTS OF REGRESSION ANALYSIS STUDY 2

R^2=.07	B	SE	*t*	p
Intercept	12.37	3.605	3.433	0.001
TLX_j	-0.172	0.093	-1.863	0.070
Z_j	0.513	0.215	2.381	0.019

FIGURE 2:

DISTRIBUTION OF RESPONSES ACROSS THE DIFFERENT CONDITIONS:

"ANYONE CAN RAISE HIS STANDARD OF LIVING IF ONE TRIES"



III.6 General Discussion

The goal of this article was to improve our understanding of the potential pervasive effects that cognitive load can have on data. Collectively, two studies point out that cognitive load does not have an effect on NARS. In contrast, perceived cognitive load does exert a significant influence on NARS but only if a critical level of load is exceeded. So, a respondent's capacity of dealing with the level of cognitive load (i.e., respondents that deal well with a high level of externally posed load, will perceive the load as moderate or low) determines the level of NARS in data. Compared to earlier studies, our levels of cognitive load are subjectively determined.

III.6.1 Theoretical and Managerial Implications

These findings advance our theoretical understanding of cognitive load and NARS in a number of ways. First, our findings shed light on when cognitive load impacts response bias in terms of NARS. Earlier, researchers claimed that cognitive load does not always pose problems (Dhar, Nowlis & Sherman, 2000; Macrae, Milne & Bodenhausen, 1994). In this regard, the present studies add to this literature stream by highlighting that indeed problems do not arise when respondents do not perceive cognitive load as high. However, NARS and consequently the bias of the survey data do increase under high perceived cognitive load.

Second, the mechanism behind these findings could be attributed to the level of load perceived by the respondent. Under moderate load, respondents react by accelerating their pace of answering without truncating a response phase. So, within the limits of their cognitive capacity, respondents can compensate for an increase in the level of cognitive load thereby maintaining systematic processing. In contrast, respondents under high load cannot keep up with the high pace and will simplify by omitting a great part of the retrieval and judgment phase which results in the use of NARS.

There are many field settings in which our results are particularly relevant. For example, when respondents are interviewed at the entrance/exit of a supermarket or in shopping streets, they are often under time pressure and/or preoccupied by a stream of thoughts such as their concern for frozen products, for being in time at work, for picking up their children in time, or for what they will prepare for dinner that night. As a result, these respondents will boost self-reported measures which affect research conclusions. Consequently, NARS can be an alternative explanation for the obtained results. Also in-home interviews or online research do not occur in a vacuum where for instance interviewer and respondent have complete privacy and can as such be influenced by the presence of distracters like TV/radio and time

constraints. These distractors in combination with an online/offline questionnaire increases the level of cognitive load and can enhance respondents' tendency to respond in terms of NARS. Also in mediation analyses, measurement error on the mediator incited by high cognitive load can lead to an overestimation of the dependent variable. Consequently, successful mediators can be overlooked (Baron & Kenny, 1986).

A possible solution for reducing this bias in data would be to remove all emotional and cognitive distractions and, for example, invite respondents to participate in lab research. However, this can sometimes be too time consuming or too costly. Other techniques can also contribute to better data quality such as informing people correctly of the survey duration so that respondents can judge whether they have the time available to participate, providing a place nearby where respondents can take a seat and take their time, adding a reasoning task to get an indication of cognitive difficulties, etcetera. Strategies as motivating respondents into systematic processing will possibly not have the intended effect because in case of severe cognitive load even highly motivated respondents lack processing (Suri & Monroe, 2003). Accordingly, the presence of motivated respondents does not mean that data is free of bias since cognitive processes mainly determine the level of response bias, i.e. NARS. So, this confirms earlier suggestions that NARS is a cognitive style rather than a motivational one. Finally, next to preventing an increase of NARS, researchers can add a battery of heterogeneous items to the questionnaire to measure the level of NARS and to correct for it a posteriori (Baumgartner & Steenkamp, 2001).

III.6.2 Limitations and suggestions for further research

We would like to point out some of the limitations of our study. First, the cognitive load manipulations used in our studies provide each a continuous cognitive load by focusing on

attentional processes rather than memory processes. One can question whether the results would still hold when a memory based load manipulation (i.e., remember a 10 digit number while answering the questions) is used. In this regard, we believe that memory based loads will probably not replicate our findings since these manipulations put no load on the process of answering questions (i.e., one can easily answer a question and afterwards repeat the number before going to the second question).

Second, more research is needed to further explore the underlying mechanism. In particular, cognitive interviews (DeMaio & Rothgeb, 1996; Jobe & Mingay, 1989) can reveal how and in which processing phase(s) respondents simplify. Another area for future research concerns possible moderators such as individual differences in cognitive sophistication. Cognitive load is a part of many people's daily life, yet persons differ in how well they cope with this daily cognitive burden. An individual's cognitive sophistication is determined both by innate factors, such as stress immunity and by experience. Therefore, respondents who are, for instance, immune to stress, or work well under load, can be expected to be less affected by cognitive load. Also, people who are experienced in coping with cognitive load such as higher educated people holding positions with major responsibilities can be expected to be less affected by a cognitive load task when filling out a questionnaire. On the other hand, children and elderly persons may have more difficulties to cope with dual tasks or time pressure (Borgers et al., 2004).

Third, the TLX index is normally measured by combining the different weights given by the respondent to each subcategory of workload (e.g., mental demand, temporal demand, performance, effort and frustration level) and the respondent's response to each of them (i.e., a number on 100). In our studies, TLX was measured without the use of weights for each

subcategory. Further research could take into account the weighting variable to find out whether our findings remain. However, we do not expect significant differences from our results since the variations in weights are mainly determined by the source of cognitive load posed on the respondent and are less a reflection of individual differences in the subjective importance of different subscales (Hart & Staveland, 1988). As we used the same type of cognitive load tasks (in all our studies) and we used a homogeneous group of respondents, we do not expect any differences.

Based on our results, we expect that earlier findings concerning the effect of cognitive load on consumer phenomena can be partly explained by an enhanced level of NARS. However, in our studies we only focused on changes in the level of NARS under different levels of cognitive load. Further research should point out the detrimental effects that our results can have on, for example, the existing stream of literature on cognitive load.

Chapter IV

How Self-Regulatory Focus Shapes Item Responses Regardless of Content

IV.1 Chapter outline

Response styles, such as the extreme response style and midpoint response style, contaminate questionnaires. The individual antecedents of response styles have proven to be elusive. The authors propose an effect of the important personality trait self-regulatory focus on response styles. Findings point out that self-regulatory focus, measured by means of a uniquely developed combined test, shapes responses to items regardless of their content. Promotion focused people show higher levels of the extreme response style than prevention focused people whereas the reverse is true for the midpoint response style. This article provides evidence for a central link between personality and response styles.

IV.2 Introduction

It is well known that responses to questionnaires are often influenced by content-irrelevant factors named response styles (Baumgartner & Steenkamp, 2001, 2006). However, these systematic errors are still often neglected. Response styles are defined as tendencies to respond systematically to questionnaire items on some basis other than what the items were specifically designed to measure (Paulhus, 1991). As response styles are not content-specific, they can occur in any measurement scale (Bentler, Jackson & Messick, 1971).

In this article, we focus on the Extreme response style (ERS) and the Midpoint response style (MRS). In the extensive literature of response styles, both ERS and MRS have received little attention in journals (cf., Baumgartner & Steenkamp, 2001; De Jong, Steenkamp, Fox & Baumgartner, 2008). The latter is quite astonishing as both response styles have biasing effects that cannot be corrected for in advance (i.e., during scale construction). Therefore, more attention should be placed on ERS and MRS.

Despite their biasing effects, both response styles are not commonly corrected for. This is probably due to the uncertainty of both how to deal with these types of stylistic responding and the mechanisms that underlie them. What drives people to respond in these specific stylistic ways? There is quite some early literature suggesting that the tendency to endorse the extremes or the midpoint is a manifestation of certain personality attributes (e.g., Couch & Keniston, 1960). However, most of these studies did not methodically disentangle content and style or have been subject to much methodological criticism (Hamilton, 1968). First, the studies confounded content and style (e.g., Borgatta & Glass, 1961), which makes it impossible to univocally contribute the observed effects to response style bias. Unlike the

former studies, we will measure response styles in a way that optimally controls for content. Second, many of the studies were largely explorative in nature (e.g., Bachman & O'Malley, 1984). Another contribution of the current study, therefore, is that it proposes and empirically validates hypotheses on the relation between a key personality trait, namely the self-regulatory focus and the response styles, ERS and MRS.

By addressing the two shortcomings listed above, we join a new research method (Naemi, Beal & Payne, 2009) that has only recently emerged. Specifically, in a recent paper, Naemi, et al. (2009) linked personality factors decisiveness, intolerance of ambiguity and simplistic thinking to the use of ERS. However, they used scales to measure the personality traits of which the answers themselves could be contaminated by the use of response styles (Bentler, Jackson & Messick, 1971). Being fully aware of this potential bias, the authors decided to present the personality scales to the peers of the respondents. This technique reduces the likelihood of having shared contamination by response styles of the independent and dependent variables, but there are still some shortcomings. First of all, focal respondents might have been selective in recruiting peers (as there were no specific eligibility criteria). Secondly, as less than half of the respondents were preserved, there could be a problem of selection bias from the part of the peers as well. The combined selection bias can be related to response styles, which can limit validity. Therefore, in the current study, we choose another approach in which we avoid the use of direct measurement scales completely and focus instead on actual behavioral patterns of focal respondents measured in a standardized setting (i.e., using validated tests that do not involve self-reports). In addition, we study a different and increasingly important personality trait, self-regulatory focus, and link it to the use of ERS and MRS as both response styles have major persistent influence on survey data. We believe our approach complements the one proposed by Naemi et al. (2009) and that future

research can benefit from integrating the two methods, namely ours and the one of Naemi et al. (2009), for cross-validating the relation between personality and response styles.

The personality trait, self-regulatory focus, has proven its importance in recent research. This personality trait activates higher level goals and needs that are to be fulfilled and determines the typical strategies that are used to fulfill them (Higgins, 1997). In addition, it is a motivational construct that provides us with important insights about the processes underlying decision-making (Pham & Higgins, 2005). Given its pervasive impact on human decision making, self-regulatory focus can be expected to influence the processes that lead a respondent to select a specific response category in a questionnaire. We subsequently discuss response styles and self-regulatory focus, and explain why ERS and MRS can be linked with self-regulatory focus.

IV.3 Response styles

Response styles are defined as the tendency to respond in a systematic way to items independent of the content (Cronbach, 1950; Rorer, 1965; Paulhus, 1991). They inflate reliability (Greenleaf, 1992a) and threaten the validity of empirical findings by contaminating respondents' answers to substantive questions (Cronbach, 1942). In addition, they influence the correlations between scales (Baumgartner & Steenkamp, 2001). The biased estimates of these correlations may, in turn, bias results from a variety of methods, including regression analysis, factor analysis, and cluster analysis.

The ERS is the tendency to use the extreme options on a rating scale. Researchers have agreed on the detrimental effects of ERS. This response style not only affects the spread of

observed scores (Cheung & Rensvold, 2000), it also induces spurious correlations among otherwise unrelated constructs (Baumgartner & Steenkamp, 2001; Chun, Campbell & Yoo, 1974; Hui & Triandis, 1985).

Whereas some people overuse the extreme response categories, others seem to prefer the moderate response categories or those categories near the middle of the scale. The MRS is the tendency to disproportionally use the midpoint of the response scale. This style also has a biasing effect on scale scores, depending on the deviation of the mean of the scale from the midpoint (Baumgartner & Steenkamp, 2001).

Several authors have argued that response styles are stable individual characteristics (e.g., Billiet & Davidov, 2008; Hamilton, 1968; Messick, 1968). However, most of the variance of the response styles remained unexplained, and it is still unclear how response styles exactly relate to central personality traits. We propose the self-regulatory focus to be a plausible candidate for explaining individual differences in ERS and MRS.

IV.4 Self-regulatory focus

Although there is general agreement that consumers' goals and motivations play a fundamental guiding role in information processing and consumer behavior (Pham & Higgins, 2005), there is no universal way of classifying these goals. However, a classification gaining in importance is the distinction between *ideals* (i.e., aspirations and hopes) and *oughts* (i.e., duties and obligations) which relates to the motivational construct of self-regulatory focus (Higgins, 1997; Pham & Avnet, 2004). Self-regulatory focus theory builds on the hedonic principle of pleasure and pain by distinguishing between two different kinds of self-regulation

in relation to desired or undesired end states (Higgins, 1997). These two self-regulation or motivational systems are labeled promotion and prevention focus. The foci involve different types of goals and strategic orientations. Whereas individuals in a promotion focus are eager to approach ideals, hopes and wishes, individuals in a prevention focus are more vigilant to assure safety and to avoid danger/losses by doing what ought to be done (Crowe & Higgins, 1997; Higgins, 1996, 1997).

IV.5 Self-regulatory focus, MRS and ERS

Self-regulatory focus can be an important antecedent of response styles as it distinguishes in the way people process information based on their differences in strategic orientation. Especially, these specific differences make self-regulatory focus a plausible candidate for explaining individual differences in MRS and ERS.

Prevention focused people concentrate on avoiding failures and are driven to protect themselves against potential threats. They are more vigilant and less willing to accept risks (Crowe & Higgins, 1997). As a result, they exhibit a conservative bias. Accordingly, these individuals will avoid the extreme option as this increases the risk of making a poor choice (Mourali, Böckenholt & Laroche, 2007), and they will rather choose the safer "neutral" option, i.e., an option that avoids failure (Crowe & Higgins, 1997). Choosing such an option will have the consequence that prevention focused individuals are more likely to defer choice and elect the no choice option (Pham & Higgins, 2005).

In contrast with prevention focused people, promotion focused people are more open to risks and are more likely to take action (commission), i.e. to actually make a choice other than the no-choice or the neutral (Crowe & Higgins, 1997). In addition, Chernev (2004) found

evidence for the fact that promotion focused people are less likely to display extremeness aversion because promotion people exhibit a passion for maximizing positive outcomes. Maximizing outcomes is then related to a higher attraction of the extreme values and less of the intermediate options. A promotion focused person has a strategic preference for speed (Förster, Higgins & Bianco, 2003). The latter is found to be related to extreme responding (Paulhus & Lim, 1994). Finally, promotion focused persons have a more individualistic self-view (Aaker & Lee, 2001). It is known that individualists like to distinguish themselves from others by showing conviction in their choice (Cheung & Rensvold, 2000). Demonstrating conviction and being decisive is closely linked with the use of extreme options (Cheung & Rensvold, 2000; Naemi, Beal & Payne, 2009). For all the above reasons, it seems more likely that MRS will be more prevalent among prevention focused people whereas promotion focused people will show higher levels of ERS.

H1: Prevention focused individuals will show higher levels of MRS than promotion focused individuals.

H2: Promotion focused individuals will have higher levels of ERS than prevention focused people.

IV.6 Method

The objective of this study was to examine the link between self-regulatory focus and the response styles MRS and ERS. Instead of using the traditional measures of the self-regulatory focus, we developed another method to measure self-regulatory focus in order to test our hypotheses properly.

IV.6.1 Materials

Self-regulatory focus usually is measured by means of scales, such as the RFQ scale (Higgins et al. 2001), the lockwood scale (Lockwood, Jordan & Kunda, 2002), the BIS/BAS scale (Carver & White, 1994), or the composite scale of chronic SRF (Haws, Dholakia & Bearden, 2010). Because the answers to these scales themselves can be contaminated by response styles (Craig & Douglas, 2000), we chose to follow an alternative procedure to assess self-regulatory focus.

We scrutinized the existing literature for tests that could reveal a person's self-regulatory focus and selected three different tests that in previous studies have proven to discriminate among a promotion and prevention focus. This method is comparable with the one used in Pham and Avnet (2004), where the authors combined several tests as a manipulation check for priming self-regulatory focus. Here, we also use a combination of tests to encompass one's reflection of a self-regulatory tendency. Importantly, the three tests cannot be influenced by response styles as they do not make use of a multi-item measurement scale. The specific criteria used to select the tests were the following. First, each test had the advantage that respondents were unaware of their real purpose namely, giving an indication of the personal dominant self-regulatory focus. Second, all the tests had a close link with self-regulatory focus theory, were widely referred to and were firmly validated.

The first test was the speed/accuracy test developed by Förster, Higgins and Bianco (2003). According to these authors, promotion focused individuals stress speed over accuracy, whereas prevention focused individuals stress accuracy over speed. In the test, dots of a drawing had to be connected. A respondent who is fast will have more dots connected (more speed), but will also have made more mistakes (less accuracy). Therefore, as found in their study, a prevention focused individual who is more conservative and sensitive for mistakes, had fewer dots connected than a promotion focused individual, but did the exercise with less

mistakes. Mistakes were defined as missing a dot when linking it to an adjacent dot. Based on this exercise, we could categorize the respondents, according to their mean speed and accuracy levels, into two groups: prevention or promotion focused. More specifically, we measured average speed and accuracy. If a respondent was faster/slower than the overall speed mean and had a lower/higher amount of accuracy compared with the overall accuracy mean, than he/she was respectively labeled as promotion/prevention focused. Respondents, who were both fast and accurate or slow and inaccurate, were coded as missing.

The second test was a determination task (e.g., Liberman, Idson, Camacho & Higgins, 1999). This test was based on task substitution. Participants had to describe three abstract figures in such a way that another person would be able to recognize each of them in a series of 10 based on their descriptions. Respondents made the descriptions at own pace. All of the respondents were briefly interrupted, by the computer, while describing the third figure. Respondents were informed that the interruption was of short duration. The interruption took place at a pre-tested time when people had invested a lot of time in the exercise, but had not finished. Just after the interruption, respondents were asked to write down the percentage of the figure description they had completed before the interruption. On average, they indicated to have completed around 65% of the task. After engaging briefly in an unrelated task, people were asked to indicate whether they would like to continue with the old figure from the point where they were interrupted or whether they wanted to describe a new figure. A second question asked whether they would continue with the old figure or whether they wanted to describe a new figure, in case unfinished information had been lost due to computer fall-out. Choosing the new figure, for both questions, was indicative of a promotion focus because they favor a risky strategy which implies more openness to change. On the other hand, sticking to the old figure was indicative of a prevention focus, as these individuals are

concerned with safety and consequently favor stability. Respondents were only classified when they showed consistency across both questions.

The final test was a gift task, also from Liberman et al. (1999). The difference between this task and the previous determination task is that the gift task is not based on task substitution, but on object substitution. Two popular gifts were selected from eight objects of the same value by means of own pre-tests. Respondents received one gift (the gift was either the most liked or the second most liked). After receiving the gift, they needed to fulfill a letter task in order to keep the gift. Next, all respondents received negative feedback about their performance on the task. As a consequence, they lost their gift. Respondents could regain their old gift or another gift during a new letter task. They had to indicate for which gift they wanted to fulfill the new letter task. As prevention focused individuals are rather conservative, they are assumed not to swap gifts. Promotion focused people, on the other hand, are expected to switch. As a result, persons who indicated that they wanted the initial 'old' gift were categorized as having a prevention focus. Those who preferred the alternative 'new' gift were categorized as having a promotion focus.

Both regulatory foci are assumed to coexist in a given individual. However, for a given individual, one focus usually is more dominant (Pham & Avnet, 2004). In line with this, we created a measure of dominant self-regulatory focus. The advantages of this dominant focus as a latent construct using the selected tests as indicators are twofold. First, this approach is conceptually and operationally most appropriate as only construct-related variance (common across the tests) is of interest as an antecedent of response styles, whereas the unique variance of each of the tests is not. Secondly, through this approach, we can verify the convergent validity of the tests.

IV.6.2 Participants

Two hundred and twenty-eight undergraduate students (67 males, 161 females) participated in this study. None of the students were excluded. The incentive was a movie ticket (€6) for each participant. Subjects were run in groups of five to eight persons. Each of the respondents was individually assigned to a computer. Respondents were not informed about the subject of the research.

IV.6.3 Dependent variables

In previous research, many authors used existing personality scales in order to capture the link between personality and response styles (e.g., Hamilton, 1968). However, it is almost impossible to determine to what extent extreme responses and midpoint responses represent stylistic tendencies of the individual or a subject's meaningful response when the items are homogeneous. Therefore, it is required to use heterogeneous items to avoid confounding between content and style. If all the items represent different constructs that are unrelated, it can be expected that relations or similarities in individual's responses are mainly due to pure behavioral tendencies unrelated to content, namely response styles. Such approach is advocated by Greenleaf (1992b) and by Baumgartner and Steenkamp (2001).

To maximize heterogeneity, we took a random sample of unrelated items from the scales compilation by Bruner, James and Hensel (2001). It is reasonable to assume that the scales from which the items were drawn have acceptable levels of discriminant validity, as all scales have been subjected to a thorough validation process (Bruner, James & Hensel, 2001). Accordingly, in this study, 70 heterogeneous items were used to capture ERS and MRS (see Table 1). We also verified that the items were very heterogeneous in content (average correlation $r_{avg} = .087$).

TABLE 1:

EXAMPLES OF HETEROGENEOUS ITEMS

Heterogeneous items
My family is very social
There is little I can do to change many of the important things in my life
I am good at sports
A woman working out of home with children is still a good mother
I'm confident that I can learn technology-related skills

The formulas for the response style measures are based on Bachman and O'Malley (1984) and on Baumgartner and Steenkamp (2001) (see Table 2). Specifically, we computed the proportion of midpoint responses (i.e., four on a 7-point scale) across the 70 heterogeneous items as a measure of MRS (M=.148; SD=.081). Similarly, we computed the proportion of extreme responses (i.e., one or seven on a 7-point scale) across the 70 heterogeneous items as a measure of ERS (M=.198; SD=.138).

TABLE 2:

FORMULAS OF MRS AND ERS

Formulas
MRS = [usage frequency of option 4]/total number of items
ERS = [(usage frequency of option 1) + (usage frequency of option 7)]/total number of items

IV.6.4 Procedure

All the respondents were told that the tasks were self-paced, apart from the first speed/accuracy task which had a time limit. A session consisted of the three tests to identify the self-regulatory focus and a questionnaire which contained a wide variety of items that were unrelated to one another in terms of content. As already indicated, the items were randomly chosen from Bruner, James and Hensel (2001), resulting in a highly heterogeneous set. All items were adapted to a 7-point Likert scale anchored by the labels "totally disagree/totally agree" at the endpoints and "neutral" at the midpoint. Participants then responded to socio-demographic questions and submitted their responses by clicking on a submit button. A thank you screen appeared.

IV.7 Results

To test our hypotheses, we specified a model in which ERS and MRS are the dependent variables and Dominant Focus (DF) is the independent variable. As for the latter, we measured DF by means of the three tests that yield binary scores, as described in the material section. The proportions of respondents scoring positively on the three respective tests were .526, .461 and .583 (positive scores indicate prevention focus). We defined the DF construct as a latent factor with three binary indicators. To account for random error, we specified ERS and MRS as latent factors with a single indicator by fixing the residual variance of the ERS and the MRS indicators to $(1-\alpha_{ERS})*s^2_{ERS}$ and $(1-\alpha_{MRS})*s^2_{MRS}$ and by fixing the factor loadings of ERS and MRS to $(\alpha_{ERS})^{½}$ and $(\alpha_{MRS})^{½}$ respectively. We used the split-half reliability of the ERS and MRS measures as a proxy for alpha (cf. Greenleaf, 1992b). The consistency coefficients were .73 for MRS and .74 for ERS, which are both satisfactory.

We tested the model in Mplus 5.1 using the weighted least squares estimator (WLSMV) (Muthén & Muthén, 2007). With this robust WLSMV, probit regressions are estimated for the categorical factor indicators, and linear regressions are estimated for the interval scaled outcomes.

The model fitted the data rather well. The chi square test indicated non-significant misfit ($chi^2(4)=2.771$, p=.597) and the alternative fit indices indicated correspondingly good fit (CFI = 1.000; TLI = 1.088; RMSEA = 0.000). The three DF indicators all loaded positively and significantly on the DF factor, with probit loadings of .34 (p = .046), .54 (p = .014) and .35 (p = .044) respectively.

The results supported the hypotheses. In particular, in line with Hypothesis 1, the regression weight of MRS on DF was significantly positive (Unstandardized regression weight = .018; Standardized regression weight = .224; one-sided t = 1.684, p <.05; R^2 = .05). In other words, for an increase in DF by one standard deviation, respondents tend to use the midpoint for an additional 1.8% of the items (cf. the unstandardized regression weight). Given the MRS intercept of .148, this means that respondents in the 95th DF percentile (i.e., highly dominant prevention focus) answer approximately 18.3% of items with a midpoint response, whereas respondents in the 5th DF percentile answer approximately 11.3% of items with a midpoint response.

Furthermore, Hypothesis 2 is also supported as the regression weight of ERS on DF is significantly negative (Unstandardized regression weight = -.042; Standardized regression weight = -.307; one-sided t = -2.194, p<.01; R^2 = .09). In other words, for a decrease in DF by one standard deviation, respondents tend to use the extreme options for an additional 4.2% of

items. Given the ERS intercept of .198, this means that respondents in the 95th DF percentile answer approximately 11.6% of items with an extreme response, whereas respondents in the 5th DF (i.e. highly dominant promotion focus) percentile answer approximately 28.0% of items with an extreme response.

IV.8 Discussion

What drives people to respond in specific stylistic ways? So far, the often suggested link with personality traits has not been investigated without the use of scales. Self-regulatory focus has already proven to be very influential in how people process and decide on different matters. Consequently, the goal of the current paper was to investigate whether this personality trait has a significant impact on response styles as well. The conclusion of this empirical study is that self-regulatory focus does shape responses to items, regardless of item content.

In particular, our results revealed a significantly higher usage frequency of the midpoint among prevention focused individuals than among promotion focused individuals whereas promotion focused people had a higher tendency of answering in extremes compared to the prevention focused people. These results seem to be driven by a fundamental difference between the two foci in strategic orientations. A prevention focused person's tendency for the neutral is a way of not undertaking the action or not making a choice (Spranca, Minsk & Baron, 1991). The reason behind is that the fear of making a mistake leads to a strategy of avoidance among prevention focused individuals (Dhar, 1997; Higgins, 1997). On the other hand, promotion focused people do want to make a choice and want to show conviction in what they choose (Cheung & Rensvold, 2000). Consequently, they have a higher tendency of answering in extremes compared to the prevention focused people.

These findings support the proposition that the type of response style employed in a certain situation greatly depends on the particular individual (Shulman, 1973). The identification of this antecedent of the response styles ERS and MRS has important implications for research in general and for response style research more specifically. Self-regulatory focus, which is a personality trait, can help explain the stable character of response styles. As all the items are heterogeneous in content, the individual differences found are due to pure behavioral tendencies.

IV.8.1 Implications

Investigating the effect of goal orientation on individuals' tendency to answer in extremes or the midpoint has the potential to expand the understanding of the role of self-regulatory factors in response styles. Several researchers have suggested a link between personal characteristics and response styles (e.g. Baumgartner & Steenkamp, 2001; Podsakoff et al., 2003). In this article we provided empirical prove of a relation between an important personality trait and response styles. This new knowledge deepens our understanding of the phenomenon of response styles. Although the effect sizes are moderate, the levels of explained variance match those of other antecedents like demographics (Greenleaf, 1992b). Our findings have important implications for every researcher who makes use of multi-item scales, no matter what the subject of the research is.

Researchers are particularly concerned with the possibility that one or more of the groups being studied is especially prone to a certain response bias (Stening & Everett, 1984). More specifically, this concern is not only focused on the impact of the response style per se, but even more importantly, on the impact of differences in response styles on data equivalence. The personality of the respondent will also determine the type and amount of method bias

affecting the scores. As a result, intergroup differences in mean scores will then partly be a reflection of stylistic responding (Chun, Campbell & Yoo, 1974). For instance, when intergroup means are rather small or large while controlling for demographics and situational influences, it would be interesting to check for ERS and MRS as a reflection of the group's or individual's self-regulatory focus (Heide & Gronhaug, 1992).

Other implications for researchers can be found in the clustering of data (Cronbach, 1942). When a clustering is performed on collected data, there could be a problem with the quality of the results. If a group contains mainly promotion focused people, they will be classified as extreme partly because of their high usage levels of the extreme response style. They can actually belong in more moderate segments. In contrast, when a cluster group mainly contains prevention focused people, respondents could be misclassified due to their midpoint answering tendency. As a consequence, these individuals who are classified in moderate segments could actually belong to extreme segments. Data will as such be misinterpreted and misclassifications will be made (Greenleaf, 1992a). Furthermore, in studies where self-regulatory focus is correlated with constructs that are measured by means of self-report scales, response styles may provide an alternative explanation for the relations that are observed. So, the effects found in this study can be valid for a variety of settings such as product evaluation, brand evaluation and sensory research (i.e., smell and taste tests). Often different brands are judged on their level of likeability, however we expect that respondents' ratings will be different based on whether the respondent has a promotion versus prevention focus. Promotion oriented respondents will exaggerate in their ratings whereas prevention focused respondents will not. Nevertheless, the differences found for these brand evaluations can mainly be attributed to the presence of response styles instead of real regulatory focus differences.

Both response styles ERS and MRS are not only an important threat to the validity of domestic survey based research but also for cross-cultural research (De Jong, Steenkamp, Fox & Baumgartner, 2008), as researchers have found that self-regulatory focus also varies across countries and cultures (Lee, Aaker & Gardner, 2000). Specifically, collectivistic countries tend to be more interdependent (Hamid, Lai & Cheng, 2001). This self-view encourages the values of fulfilling obligations and responsibilities. Hence, interdependent cultures are closely related with prevention focus (Lee, Aaker & Gardner, 2000). On the contrary, independent cultures go together with promotion focused view. They are more individualistic and focused on uniqueness and determination (Oyserman, Coon & Kemmelmeier, 2002). Since, culture can influence one's personality during development (McCrae & Costa, 1996; Church, 2001), it can be expected that in individualistic (interdependent) countries, more respondents will be promotion (prevention) focused. Consequently, comparability between two or more countries differing on this dimension cannot be guaranteed as the corresponding measurement parameters may not be equivalent, but biased by the different response styles ERS and MRS. However, self-regulatory focus through interdependency/independency only covers one dimension of all dimensions underlying culture (Hofstede, 1980). In addition, an individual's personality is not fully dependent of culture. Therefore, it would be interesting to find out to what extent the variable self-regulatory focus mediates the relationship between culture and response styles.

As a result of all this, it is of major concern that data and scales are purified of response styles[31]. In recent years, important research has been done to provide measures of response styles and solutions to correct for them (Baumgartner & Steenkamp, 2001). Several authors

[31] In some cross-cultural research, researchers already correct for response styles (e.g., van Herk et al., 2004)

have tried to reduce the problem by matching samples on variables related to response styles for example by removing emotional and cognitive distraction or by giving a certain amount of time to respond (e.g., Chen, Lee & Stevenson, 1995; Knowles & Condon, 1999). Unfortunately, response styles still remained. These authors were not aware of personality characteristics such as the self-regulatory focus influencing ERS and MRS. As it is difficult to control for all these variables in advance, it is still advisable to correct for ERS and MRS after the data collection stage, especially when there is reason enough to believe that a certain focus was dominant in the sample or in a certain group. The latter can be done by means of the three-step procedure developed by Baumgartner and Steenkamp (2001). The first step consists of constructing an estimate of a person's ERS and MRS score based on a set of heterogeneous items. Second, after identifying the response styles, researchers need to consider adjustments. We recommend implementing the regression technique. This means that researchers can regress the scores on the contaminating response styles such that the residuals represent purified scores. Alternatively, the response styles measures can also serve as covariates in analyses. Another possible solution is to employ structural equation modeling with the response styles as covariates (Podsakoff et al., 2003). Finally, after obtaining purified scores, further analysis can be done. If resources allow for it, an alternative method can be used. The self-regulatory focus could be used as a proxy for ERS and MRS (such as demographics) and taken into account when analyzing (i.e., controlling for it).

Methodological challenges are the major obstacle in the search for individual difference variables that explain response styles. Scales have to be avoided and tests need to be developed in order to measure the antecedents of response styles. Measuring, for instance, self-regulatory focus, by means of the combined tests, will make it easier in the future to

predict when certain response styles will be most likely to emerge in the data and to control for it.

IV.8.2 Further research and limitations

In this article, we demonstrated that promotion focused people use extremes more often (ERS) whereas prevention focused people tend to more frequently use the midpoint option (MRS). Mechanisms underlying these effects were suggested. More specifically, we assume that the risky behavior of promotion focused individuals versus the conservative behavior of prevention focused individuals provide a plausible explanation. Responding by means of the middle position is an effective strategy when a perceived risk of being wrong is apparent. Nevertheless, more research is needed to validate whether risk behavior and/or independency are moderators, and to which extent, of this relation. In terms of generalizability it would be interesting to find out whether the significant effects of self-regulatory focus hold in a heterogeneous setting when controlling for age and education, which are factors that influence response styles (Greenleaf, 1992b; Krosnick & Fabrigar, 2003). Then we will also know whether the significant effects of self-regulatory focus on response styles depend on socio-demographics; that is, whether self-regulatory focus lead to bias for specific demographic segments.

Finally, the present article is based on 7-point Likert scales (with labels for the midpoint and the extreme categories). Further research could examine the effect of self-regulatory focus on response styles using other scale formats (e.g., semantic differentials). It would be interesting to know whether there are formats available where the influence of such a psychological variable is less apparent.

Chapter V

General Discussion

"Do we know all we need to know about response biases?"

(Watkins & Cheung, 1995)

V.1 Chapter outline

In this concluding chapter, the previous chapters are recapitulated. Based on this overview, the theoretical and practical implications are discussed.

V.2 Recapitulation

Questionnaires with closed-ended questions are often used in consumer research. Researchers expect that the answers on those closed-ended questions are a reflection of a respondent's true answer. However, it has been known for a long time that people's responses are influenced by content-irrelevant factors, such as response styles (Cronbach, 1946). Although the problem of response styles is well known and several techniques have been suggested to correct for these biases (e.g., Baumgartner & Steenkamp, 2001; Van Rosmalen, van Herk & Groenen, 2010) researchers still do not know when and for which type of respondents response styles are most problematic. Consequently, researchers have not given much attention to these biasing effects. Therefore, we have focused in this dissertation on when and for which type of respondents response styles are most likely to occur and which research methods and research settings help to limit the presence of response styles.

In this dissertation, response styles were conceptualized as respondents' stylistic tendencies to respond to questions (Paulhus, 1991). Individuals may exhibit stylistic tendencies to agree (Acquiescence response style or ARS), to disagree (Disacquiescence response style or DARS), to agree over disagree (Net Acquiescence response style or NARS), to answer in extremes (Extreme response style or ERS) or to select the midpoint (Midpoint response style or MRS). The response styles in this dissertation were measured by specific style indicators and their influence of content was corrected for by randomizing content over items.

To answer the question of when response styles are most likely to occur and for which type of respondents, we investigated task characteristics such as the scale format used in surveys (Chapters IIa and IIb) and the level of cognitive load imposed upon respondents (Chapter III).

As for dispositional variables (i.e., type of respondents) we looked at the effect of a respondents' self regulatory focus on response styles (Chapter IV).

Empirical **Study 1** (Chapter IIa) uses different Likert rating scale formats which differ in the number of response categories (from 4 till 7) and the labeling of response categories (labeling all response categories versus labeling the endpoints only) and investigates their effect on response biases: NARS, ERS and misresponse to reversed items (MR).

It was found that the inclusion of a neutral point led to an increase in NARS due to a disproportional movement of otherwise negative response options to the midpoint. Ambivalent respondents who are forced to take sides tend to react negatively. The inclusion of a midpoint also resulted in lower levels of MR and ERS. Consequently, formats with an even number of categories bias data more than formats with an odd number of categories.

The inclusion of a midpoint in combination with a fully labeled scale format did not affect the level of NARS or the level of ERS because respondents perceive a scale as more equidistant when a midpoint is added. In other words, a midpoint attracts an equal amount of (ambivalent or neutral) respondents irrespective of whether the scale is fully labeled or not.

Adding gradations of (dis)agreement did not translate into an alteration in the level of NARS and MR. However, when only the extremes were labeled, adding gradations led to an increase of MR. In terms of ERS, the presence of extra intermediate response categories reduced the level of ERS. This effect was strengthened when a midpoint had been offered or when all response categories were labeled. As a result, a 5-point scale is more preferred than a 7-point scale when only the extremes are labeled.

We also found that NARS is higher and ERS is lower in conditions where all response categories are labeled. In addition, labeling all response categories lead to less MR. Considering an odd numbered scale format without reversals, one can still question whether NARS or ERS have the most detrimental effects on data. Therefore, we studied this labeling effect on ERS and NARS in a second study. The second study used 5-point likert scale formats which varied in labeling. The findings of Study 1 were replicated in that full labeling led to an increase of NARS and a decrease of ERS, but more importantly we found that criterion validity was higher in the extreme labeled condition.

Our findings could be attributed to two main mechanisms through which formats affect response styles. First, different response scale formats imply differences in the perceived meaning and salience of response categories, thus changing the chance of them being selected (Arce-Ferrer, 2006; Schaeffer & Presser, 2003). Second, response scale formats vary in the extent to which they force ambivalent and indifferent or truly neutral respondents to choose sides when responding; this has an effect on response distributions (Nowlis, Khan & Dhar, 2002).

An important implication of the findings in this article is that response style bias in scale formats depends on the number of categories and the effect of labeling. From this study, it is recommended to use a 5-point extreme labeled format since this format performs best in minimizing response styles in linear relations.

Whereas Study 1 (Chapter IIa) focused on the dimensions of response categories, **Study 2** (Chapter IIb) examined the format of the scale, which basically differs on two major dimensions, namely Polarity (unipolar versus bipolar) and Anchoring (only positive numbers

or negative and positive numbers). We studied the effect of these manipulations on ARS, DARS, ERS and MRS. The results indicated that unipolar scales with positive anchors showed less ERS and more MRS compared to bipolar scales with positive anchors and unipolar scale with negative and positive anchors. Bipolar scales versus unipolar scales also showed more ARS, whereas unipolar scales versus bipolar scales showed more DARS. These findings are found to be robust and had no link with specific demographic segments. In addition, higher criterion validity for the unipolar scale with positive anchors in linear relations has been found.

The mechanism behind these findings could be attributed to two mechanisms, the symmetry effect and the intensity effect. Both effects explain how respondents interpret different scale formats. For bipolar scales, the presence of two poles leads to the activation of these poles and makes the scale symmetrical. Consequently, the signals (positive, negative, extreme or neutral) of the different response categories become clearer. This is in contrast with unipolar scales, where only one pole is activated for most of the response options, resulting in positive signals for the majority of the response categories. The intensity effect makes the poles of a bipolar rating scale more intense, which lead respondents away from the extreme response options. For unipolar rating scales, the presence of mixed anchors provides extra information both about the range of the scale and the meaning of the implicit pole. Consequently, the extreme response options become more salient and respondents will feel more confident in chosing those options.

This study implies that, next to the format of the response categories, also the format of the scale is an important determinant in the quest for minimizing response style bias. In this regard, the unipolar scale with positive anchors, also known as the Likert format, performs

best. However, the choice of a scale format should be based on the researcher's knowledge about the relevant distribution of the sample and upon the match between question interpretation and researcher interest.

Study 3 (Chapter III), extended the effect of situational variables on response styles from task characteristics to environmental characteristics, such as the level of cognitive load posed on a respondent. In two studies, we investigated the effect of cognitive load on NARS and showed an increase in the level of NARS when respondents are under high perceived load but not when respondents perceive the cognitive load as moderate or low. In this article, we focused on the subjective load of the respondent which is a better predictor of the findings than the level of objective load posed on the respondent.

Apparently, when respondents perceive the level of cognitive load as moderate, they react by accelerating their pace of answering. This means that respondents are capable of systematic processing under moderate perceived load. In contrast, respondents under high perceived load cannot keep up with the high pace and simplify by omitting a great parts of the response phase which results in a higher level of NARS.

Whereas the former empirical studies focused on situational variables, **Study 4** (Chapter IV) investigated the effect of a dispositional variable, such as the Self-Regulatory focus, on ERS and MRS. To link this personality trait to these response styles, we needed to follow an alternative procedure to assess a respondents' self-regulatory focus since personality scales are contaminated. Consequently, we created the Dominant Focus (DF), measured by means of three tests; the speed/accuracy test, the determination task and the gift task, that each yielding

binary scores. In essence, we found that promotion focused respondents showed a higher level of ERS whereas prevention focused respondents showed a higher level of MRS.

These results seem to be driven by a fundamental difference between the two foci in strategic orientations. A prevention focused person is afraid of making mistakes which leads to a strategy of avoidance. Accordingly, these respondents will rather choose the safer "neutral" option to avoid failure. On the other hand, promotion focused people do want to make a choice and want to show conviction in doing so, therefore they are more likely to choose the extreme options.

V.3 Theoretical Implications

The empirical studies in this dissertation advance our theoretical understanding of NARS, ARS, DARS, ERS and MRS in several ways.

First, the majority of the response styles studies focused on the generality and stability of response styles or on post hoc correction techniques. As a consequence of these publications, interest in the impact of situational and dispositional variables rose. Antecedents as demographics and culture only explain a minor portion of the total variance of response styles, so much variance remained unexplained. In this respect, this dissertation contributes to the response bias literature by identifying four new sources of response styles, that each explains a part of the total variance of response styles. Study 1, 2 and 3 showed substantial differences in response styles between different scale formats and under conditions of high perceived cognitive load versus low or moderate load. Study 4 focuses more on the stable character of response styles by providing evidence for a link between the personality trait self-regulatory focus and the response styles ERS and MRS.

Second, the four sources can be related to one of the four cognitive processes that respondents perform when answering a question item, (1) comprehension (they attend to the question and interpret it), (2) retrieval (they generate a retrieval strategy and then retrieve relevant beliefs from memory), (3) judgment (they integrate the beliefs into a conclusive judgment), and (4) response phase (they map the judgment onto the available response categories and answer the question) (Tourangeau, Rips & Rasinski, 2000). The first two empirical studies, concerning scale format, show that response styles can be caused by a problem in the mapping of a judgment onto a specific response category. Study 3 points out that response styles can already occur earlier in the response process. More specifically, Study 3 showed that the presence of a distracter can lead to the truncation of the retrieval and judgment phase, which results in higher use of response styles.

V.4 Practical implications

From study 1 and 2, it is clear that differences in response format characteristics lead to different answers. These studies demonstrate that the relationship between variables, such as attitude and intention, can be significantly altered by response bias caused by the type of scale format (e.g., the number of items, the labeling of response categories, polarity and anchoring of the scale). Consequently, data obtained with different formats are not comparable and the pattern of response distributions depends on the scale format.

Empirical Study 3 implicates that the presence of distracters while filling out a questionnaire increases the level of perceived cognitive load and can enhance respondents' tendency to respond in terms of NARS. Consequently, in situations where perceived cognitive load is high data will be biased and wrong conclusions will be made. However, future research should

point out the degree to which NARS biases, for instance, persuasion or decision making data (such study is planned).

Next to situational influences, the personality of the respondent can also influence the amount of response style bias affecting scores. Furthermore, in studies where self-regulatory focus is correlated with constructs that are measured by means of self-report scales, response styles may provide an alternative explanation for the relations that are observed. However, since the effects of self-regulatory focus on response styles are rather moderate, future research should find out the degree to which the presence of response styles through self-regulatory focus are deleterious (such study is planned).

In the search for individual difference variables that explain response styles, methodological challenges are the major obstacle. Scales have to be avoided and tests need to be developed in order to measure dispositional antecedents of response styles. Measuring, for instance, self-regulatory focus, by means of combined tests, will make it easier in the future to predict when certain response styles will be most likely to emerge in the data and to control for it.

As a result of all this, it is of major concern that data and scales are purified of response styles.

V.5 Ex ante solutions

Researchers can correct for response styles afterwards by using post-hoc solution techniques or can prevent data from response style bias in advance (ex ante) such as during survey design. This dissertation focuses on the ex ante solutions.

Certain steps can be taken during the design of surveys to minimize the likely impact of response styles. In this regard, our first two empirical studies focused on this matter. Study 1 recommends the use of a 5-point endpoint labeled scale and sheds light on the problematic nature of a 7-point endpoint labeled scale in linear regressions. The problem associated with the 7-point scale format is the higher level of misresponse compared to the 5-point scale with labels at the extremes. Misresponse occurs when respondents answer either positively or negatively on both reversed and non-reversed items. So, misresponse only appears when reversals are present in the questionnaire and increases when these reversals are situated nearby their nonreversed equivalent (Hui & Triandis, 1985; Weijters, Geuens & Schillewaert, 2009). Consequently, when no reversals are present, a researcher can use a 7-point endpoint labeled scale format for linear regressions. When developing a new scale or when respondents want to report direct summaries it may better to choose a fully labeled 5-point scale. Even though this scale format leads to more optimistic scores compared to scale formats with endpoints labeled, the meaning of each response option is equal for both the respondent and the researcher. Study 2 recommends the use of Likert scale formats above the semantic differential.

Both studies advise the use of the same scale format throughout a questionnaire or across questionnaires (e.g., for repeated measures). However, this issue has not been conclusively settled. It has been indicated that when both a dependent and independent variable are measured via the same scale format the relation between variables could be overestimated due to the presence of common method bias. On the other hand, when heterogeneous scale formats are used, the relation between the independent and dependent variable can be underestimated (Rindfleisch et al., 2008). This issue deserves some further attention and future research should find out which of the two settings minimizes data bias.

Next to survey design, Study 3 highlights the importance of increased biasing effects through distractors and recommends the removal of all emotional and cognitive distractions. This conclusion favors the use of laboratory experiments, where a researcher can fully control different influences. Internal validity of such lab experiments is quite high and essential, but external validity could be at stake. Survey research executed in a lab is less susceptible to validity problems compared to manipulations, lab experiments, however, can lead to different results but only under certain conditions. Consequently, the assumption that results of lab experiments are not generalizable to the "real world" environment is not necessarily true (Sawyer et al., 1979). Boundary effects indicate when and under which situations results are not generalizable. For instance, when a field experiment focuses on a different population group than the experiment in the lab, results between the two settings can differ. The challenge here lies in designing lab experiments that can reflect real world situations.

Study 4 recommends random assignment of subjects to experimental conditions since personality affect response styles. Although, it is hard to take into account respondents' personality completely, at least one should randomize promotion and prevention respondents across the different experimental conditions.

However, as response styles are also related to factors outside the immediate control of the researcher, it is not possible to completely eliminate response styles through research design. In addition, establishing that the data is comparable can only happen after the data collection stage. So, a post hoc approach is called for. Such approach consists of two steps. First, researchers need to measure response styles based on a set of heterogeneous items. Including an amount of 20 items could be sufficient to validly measure response styles (Weijters et al., 2008). Having calculated response styles, a second step includes taking these response styles

measures into account in the analysis of the data and treat them as covariates by using analysis of covariance (ANCOVA) instead of analysis of variance (ANOVA) or Structural equation modeling (Greenleaf, 1992a; Baumgartner & Steenkamp, 2001; Podsakoff et al., 2003; Diamantopoulos et al., 2006). Alternatively, a regression technique can be used where researchers can regress the scores on the contaminating response styles such that the residuals represent purified scores or more advanced modeling techniques can be used to purify data from response styles (De Jong et al., 2008; Van Rosmalen, van Herk & Groenen, 2010).

V.6 Limitations and future research

To conclude, we note some limitations of our studies that offer opportunities for future research.

V.6.1 Variance explained

Do we already know everything we need to know about response styles? Unfortunately, the answer to this question is 'no'. The variables who were subject of our studies each explained only a part of the variance of the response styles. The scale format characteristics number of response options and labeling accounted for 7% of the explained variance in NARS and 12% in ERS; the characteristics polarity and anchoring accounted for 20% of explained variance in ARS, 9% in DARS, 7% in ERS and 3% in MRS. Cognitive load and self-regulatory focus explained around 10% of the variance of NARS, ERS and MRS. Combining our studies, we succeeded to solve a part of the variance of response styles, however another part still remains unexplained. Consequently, other situational and/or dispositional variables can explain the remaining part of the total variance of response styles. Our findings demonstrate that response styles can be caused by problems in mapping a judgment onto a specific response category

and by truncating the retrieval and judgment phase. One process that has not been covered yet is the comprehension phase. In this respect, item wording has been found to be somewhat related with response styles (Peabody, 1966; Messick, 1968). However, more research is still called for in this area. For instance, one can find out whether different forms of item wording such as true affirmation ("the work I do is useful"), false affirmation ("the work I do is useless"), true negations ("the work I do isn't useless") and false negations ("the work I do isn't useful) which impact item difficulty (Swain et al., 2008) affect ARS, DARS, MRS and ERS.

Another potential source on the stimulus-side is the difference between visual and verbal scales. Visual analogue scales (VAS) have recently gained popularity and differ from Likert scales on the visual/verbal dimension. In addition, since the VAS scale is rather new in marketing research, we can find out whether familiarity in scale format moderates the effects of scale format on response styles. Consequently, research should find out whether the VAS scale versus the Likert scale performs better or worse in terms of response styles (such study is planned). Also other situational subcategories can be linked to response styles such as the former state of a person (e.g., mood and fatigue), the task definition or social surroundings (i.e., interview effects on response styles). Respondent fatigue can lead to higher levels of response styles as respondents will probably lower their cognitive efforts the more fatigue they are. Another potential contextual influence on response styles involves the influence of mood (such studies are planned). Next to these future situational avenues, we could also have a look at other dispositional variables explaining response styles, for instance the Big Five (i.e., extraversion/surgency, agreeableness, conscientiousness, emotional stability/neurotiscism, openess to experience/intellect). Earlier research on the Big Five did not take into account that the Big Five scale itself is also prone to response styles.

V.6.2 Cross-cultural research

All studies have run in Belgium, except for the second study of Chapter IIa for which we collected data from British respondents. Culture has been indicated as an antecedent of response styles in many studies (see Chapter I). So, it would be very interesting to find out whether our studies extend to a cross-cultural context. In addition, it would also be relevant to find out to what extent scale-format and self regulatory focus mediates the relationship between culture and response styles. Both, cognitive interviews and interviewing bicultural respondents can help in finding the mechanism behind the culture – response style link. Instead of measuring respondents' chronic self-regulatory focus, a certain focus can be temporarily activated by means of priming bicultural respondents. We expect that the bicultural respondents will switch their response style behavior from MRS (ERS) to ERS (MRS). If this occurs then we have even more robust evidence for the impact of self-regulatory focus and its link with culture will be proven.

References

Aaker, J.L., & Lee, A.Y. (2001). 'I' Seek Pleasures and 'We' Avoid Pains: The Role of Self-Regulatory Goals in information Processing and Persuasion. *Journal of Consumer Research, 28*, 33-49.

Albaum, G., Roster, C., Yu, J.H., & Rogers, R.D. (2007). Simple Rating Scale Format: Exploring Extreme Response. *International Journal of Marketing Research*, *49*, 634-650.

Albaum, G., & Murphy, B.D. (1988). Extreme response on a Likert scale. *Psychological Reports, 63,* 501-502.

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

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

Arce-Ferrer, A.J. (2006). An Investigation Into the Factors Influencing Extreme-Response Style. *Educational and Psychological Measurement, 66(3)*, 374-392.

Augustin, C., & Singh, J. (2005). Curvilinear Effects of Consumer Loyalty Determinants in Relational Exchanges. *Journal of Marketing Research, 17,* 96-108.

Austin, E.J., Deary, I.J., & Egan,V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual differences, 40*, 1235-1245.

Bachman, J.G., & O'Malley, P.M. (1984). Yea-saying, Nay-saying, Going to Extremes: Black-White Differences in Response Styles. *Public Opinion Quarterly, 48*, 491-509.

Baddeley, A. (2002). Is Working Memory still Working? *European psychologist, 7(2),* 85-97.

Baron R.M., & Kenny, D.A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

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

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

Baumgartner, H., & Steenkamp, J.B. (2006). An Extended Paradigm for Measurement Analysis of Marketing Constructs Applicable to Panel Data. *Journal of Marketing Research*, *43*, 431-442.

Bearden, W.O., & Netemeyer, R.G. (1999). Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research. CA: Sage.

Belk, R.W. (1975). Situational Variables and consumer behavior. *Journal of Consumer Research, 2(3),* 157-164.

Bentler, P.M., Jackson, D.N., & Messick, S. (1971). Identification of Content and Style: A Two-Dimensional Interpretation of Acquiescence. *Psychological Bulletin, 76*, 186-204.

Bettman, J.R., & Johnson, E.J. (1988). Adaptive Strategy Selection in Decision Making. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14(3),* 534-552.

Bettman, J.R., Johnson, E.J., & Payne, J.W. (1990). A componential analysis of cognitive effort in choice. *Organizational behavior and human decision processes, 45,* 111-139.

Billiet, J.B., & Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive scales in a panel design. *Sociological Methods and Research, 36*, 542-562.

Bland, J.M., & Altman, D.G. (2000). Statistics Notes: The odds ratio. *British Medical Journal*, *320(7247),* 320-1468.

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

Borgers, N., How, J., & Sikkel, D. (2004). Response Effects in Surveys on Children and Adolescents: The Effect of Number of Response Options, Negative Wording and Neutral Mid-Point. *Quality & Quantity, 38*, 17-33.

Borgotta, E.F., & Glass, D.C. (1961). Personality Concomitants of Extreme Response set (ERS*). Journal of Social Psychology*, *55*, 213-221.

Bruner, G.C., James, K.E., & Hensel, P.J. (2001). *Marketing Scales Handbook, A compilation of Multi Item Measures Volume III*. Chicago: American Marketing Association.

Carver, C.S., & White, T.L. (1994). Behavioural Inhibition, Behavioural Activation and Affective Responses to Impending Reward and Punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology, 67,* 319-333.

Chang, L. (1994). A Psychometric evaluation of four-point and six-point Likert-type scale in relation to reliability and validity. *Applied Psychological Measurement, 18*, 205-215.

Chen, C., Lee, S.Y., & Stevenson, H.W. (1995). Response style and cross-cultural comparisons of rating scales among east Asian and north American students. *Psychological Science, 6*, 170-175.

Chernev, A. (2004). Extremeness Aversion and Attribute-Balance Effects in Choice. *Journal of Consumer Research*, *31*, 249-263.

Cheung, G.W., & Rensvold, R.B. (2000). Assessing Extreme and Acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of cross-cultural psychology, 31*, 187-212.

Chun, K.T., Campbell, J.B., & Yoo, J.H. (1974). Extreme response style in cross-cultural research: A reminder. *Journal of Cross-Cultural Psychology*, *5*, 465-480.

Church, T.A. (2001). Personality Measurement in Cross-Cultural Perspective. *Journal of Personality, 69(6),* 979-1006.

Churchill, G.A. (1979). A Paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16(Feb),* 65-73.

Churchill, G.A., & Iacobucci, D. (2005). *Marketing Research: Methodological Foundations*, 9th edn., Mason, OH: Thomson South-Western.

Clarke III, I. (2000). Extreme response style in cross-cultural research: an empirical investigation. *Journal of Social Behavioral Personality, 15(1),* 137-152.

Converse, J.M., & Presser, S. (1986). Survey Questions: Handcrafting the standardized questionnaire. CA: Sage.

Couch, A., & Keniston, K. (1960). Yea-sayers and nay-sayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology*, *60,* 151-172.

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

Craig, C.S., & Douglas, S.P. (2000). *International Marketing Research*. NY: John Wiley & Sons.

Cronbach, L.J. (1942). Studies of Acquiescence as a factor in the True-False test. *The Journal of Educational Psychology*, *33*, 401-415.

Cronbach, L.J. (1946). Response set and test validity. *Educational and Psychological Measurement*, *6(Winter),* 475-494.

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

Crowe, E., & Higgins, T.E. (1997). Regulatory Focus and Strategic Inclinations: Promotion and Prevention in Decision Making. *Organizational Behavior and Human Decision Processes, 69*, 117-132.

Cunningham, W.H., Cunningham, I.C.M., & Green, R.T. (1977). The Ipsative Process to Reduce Response Set Bias. *The Public Opinion Quarterly, 41(3),* 379-384.

Dawis, R.V. (1987). Scale Construction. *Journal of Counseling Psychology*, *34*, 481-489.

De Beuckelaer, A., Weijters, B., & Rutten, A. (2009). Using ad hoc measures for response styles: A cautionary note. *Quality and Quantity.*

Deese, J. (1965). The Structure of Associations in Language and Thought. Baltimore: John Hopkins.

De Jong, M.G., Steenkamp, J.B., Fox, J.P., & Baumgartner, H. (2008). Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation. *Journal of Marketing Research, 45*, 104-115.

De Leeuw, E.D. (1992). Data Quality in Mail, Telephone and Face to Face surveys. Amsterdam: T.T. Publikaties.

DeMaio, T.J., & Rothgeb, J.M. (1996). Cognitive interviewing techniques: In the lab and in the field. In: Schwarz, N., & Sudman S. (Eds.). *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 177-195). San Fransisco: Jossey-Bass.

De Pelsmacker, P., & Van Kenhove, P. (2006). *Marktonderzoek: Methoden en Toepassingen* (pp. 59). Amsterdam: Pearson Education Benelux.

Dhar, R. (1997). Consumer Preference for a No-Choice option. *Journal of Consumer Research, 24(2)*, 215-231.

Dhar, R., & Nowlis, S.M. (1999). The Effect of Time Pressure on Consumer Choice Deferral. *Journal of Consumer Research, 25*, 369-384.

Dhar, R., Nowlis, S.M., & Sherman, S.J. (2000). Trying Hardly or Hardly Trying: An Analysis of Context Effects in Choice, *Journal of Consumer Research, 9,* 189-200.

Diamantopoulos, A., Reynolds, N.L., & Simintiras, A.C. (2006). The impact of response styles on the stability of cross-national comparisons. *Journal of Business Research, 59,* 925-935.

Dillman, D.A., & Christian, L.M. (2005). Survey Mode as a source of Instability in Responses across Surveys. *Field Methods, 17,* 30-51.

Eagly, A.H., & Chaiken, S. (1993). *The Psychology of Attitudes.* Orlando FL: Harcourt Brace Jovanovich College.

Edland, A. (1993). The effects of time pressure on choices and judgments of candidates to a university program. In: Svenson, O., & Maule, A.J. (Eds.). *Time pressure and stress in human judgment and decision making* (pp. 145-156). New York: Plenum Press.

Edland, A., & Svenson, O. (1993). Judgment and decision making under time pressure. In: Svenson, O., & Maule, A.J. (Eds.). *Time pressure and stress in human judgment and decision making* (pp. 27-40). New York: Plenum Press.

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment, 16,* 20-30.

Eiser, J.R., & Osmon, B.E. (1978). Judgmental Perspective and Value Connotation of Response Scale Labels. *Journal of Personality and Social Psychology*, *36*, 491-497.

Eysenck, M.W. (1983). Anxiety and individual differences. In: Hockey, G.R.J. (Eds.). *Stress and Fatigue in human performance* (pp. 273-298). Chinchester: John Wiley & Sons.

Eysenck, M.W., & Calvo, M.G. (1992). Anxiety and Performance: The Processing Efficiency Theory. *Cognition and Emotion, 6(November),* 409-434.

Festinger, L., & Maccoby, N. (1964). On resistance to persuasive communications. *Journal of Abnormal and Social Psychology, 68(4),* 359-366.

Friedman, H.H., Herskovitz, P.J., & Pollack, S. (1994). Biasing Effects of Scale Checking Style in Response to a Likert scale. *Proceedings of the American Statistical Assicuation Annual Conference: Survey Research Methods,* 792-795.

Fornell, C., & Larcker, D.F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research 18(February)*, 39-50.

Förster, J., Higgins, T.E., & Bianco, A.T. (2003). Speed/accuracy decisions in task performance: Built-in trade-off or separate strategic concerns? *Organizational Behaviour and Human Decision Processes*, *90*, 148-164.

Gagné, C., & Godin, G. (2000). The Theory of Planned Behavior: Some Measurement Issues Concerning Belief-Based Variables. *Journal of Applied Social Psychology*, *30*, 2173-2193.

Gannon, K.M., & Ostrom, T.M. (1996). How Meaning is given to Rating Scales: The Effects of Response Language on Category Activation. *Journal of Experimental Social Psychology*, *32*, 337-360.

Garland, R. (1991). The Mid-Point on a Rating Scale: Is It Desirable? *Marketing Bulletin, 2*, 66-70.

Garner, W.R. (1960). Rating scales, discriminability and information transmission. *Psychological Review, 67(6)*, 343-352.

Ghiselli, E.E. (1939). All or None Versus Graded Response Questionnaires. *Journal of Applied Psychology, 23(June),* 405-15.

Gilbert, D.T., & Osborne, R.E. (1989). Thinking Backward: Some Curable and Incurable Consequences of Cognitive Busyness. *Journal of Personality and Social Psychology, 57(6),* 940-950.

Gilbert, D.T. (1991). How mental systems believe. *American Psychologist, 46,* 107-119.

Gilljam, M., & Granberg, D. (1993). Should we take Don't know for an answer? *Public Opinion Quarterly, 57(3),* 348-357.

Gopher, D., & Braune, R. (1984). On the psychophysics of workload: Why bother with Subjective measures? *Human Factors, 26(5),* 519-532.

Green, P.E., & Rao, V.R. (1970). Rating scales and information recovery: how many scales and response categories to use? *Journal of Marketing, 34(July),* 33-39.

Greenleaf, E.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, E.A. (1992b). Measuring extreme response style. *Public Opinion Quarterly, 56(3),* 328-350.

Grimm, S.D., & Church, T.A. (1999). A Cross-Cultural Study of Response Biases in Personality Measures. *Journal of Research in Personality, 33,* 415-441.

Hamid, N.P., Lai, J.C.L., & Cheng, S.T. (2001). Response bias and Public and Private Self-Consciousness in Chinese. *Social Behavior and Personality, 29*, 733-42.

Hamilton, D.L. (1968). Personality attributes associated with Extreme Response Style. *Psychological Bulletin, 69*, 192-203.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Hancock, P.A., & Meshkati, N. (Eds.). *Human Mental Workload* (pp. 239-250). Amsterdam: North Holland Press.

Harzing, A.W. (2006). Response styles in Cross-national Survey Research: A 26-country study. *International Journal of Cross Cultural Management, 6(2),* 243-266.

Haws, K.L., Dholakia, U.M., & Bearden, W.O. (2010). An assessment of Chronic regulatory focus measures. *Journal of Marketing Research, in press.*

Heide, M., & Gronhaug, K. (1992). The impact of response styles in surveys: A simulation Study. *Journal of the Market Research Society*, *34(3)*, 215-230.

Higgins, T.E. (1996). The 'Self Digest': Self-Knowledge Serving Self-Regulatory Functions. *Journal of Personality and Social Psychology*, *71*, 1062-1083.

Higgins, T.E. (1997). Beyond Pleasure and Pain, *American Psychologist, 52*, 1280-1300.

Hippler, H.J., & Schwarz, N. (1987). Response Effects in Surveys. In: Hippler, H.J., Schwarz, N., & Sudman, S. (Eds.). *Social Information Processing and Survey Methodology*. New York: Springer-Verlag.

Hofstede, G.H. (1980). Culture's consequences: International differences in work-related Values. Beverly Hills: CA: Sage.

House, R.J., Hanes, P.J., Javidan, M., Dorfman, P.W., & Gupta, V. (2004). *Culture, Leadership, and Organizations. The Globe Study of 62 Societies,* Thousand Oakds, CA: Sage Publications.

Howerd, J., & Seth, J. (1969). *The theory of buyer behavior,* New York: Wiley.

Hu, L., & Bentler, M. (1999). Cutoff Criteria for Fit indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling*, *6*, 1-55.

Hui, H.C., & Triandis, H.C. (1985). The Instability of Response Sets. *Public Opinion Quarterly, 49(2).* 253-260.

Hui, H.C., & Triandis, H.C. (1989). Effects of Culture and Response Format on Extreme Response Style. *Journal of Cross-Cultural Pshychology, 20(3),* 296-309.

Hunt, R.R., & Ellis, H.C. (1999). *Fundamentals of Cognitive Psychology (6th edition).* Boston: McGraw-Hill College.

Jackson, D.N., & Messick, S. (1962). Response styles on the MMPI: Comparison of clinical and normal samples. *Journal of Abnormal and Social Psychology, 65,* 285-299.

Jobe, J.B., & Mingay, D.J. (1989). Cognitive research improves questionnaires. *American Journal of Public Health, 79,* 1053-1055.

Johnson, T., Kulesa, P., Cho, I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-cultural Psychology*, *36*, 264-277.

Judd, C.M., & Harackiewicz, J.M. (1980). Contrast Effects in attitude judgment: An examination of the accentuation hypotheses. *Journal of Personality and Social Psychology, 38,* 569-578.

Kaplan, M.F., Whansula, T.L., & Zanna, M.P. (1993). Time Pressure and Information Integration in Social Judgment: the effect of need for structure. In: Svenson, O., & Maule, A.J. (Eds.). H*uman judgment and decision making* (pp. 255-267). New York: Plenum Press.

Klopfer, F.J., & Madden, T.M. (1980). The Middlemost Choice on Attitude Items: Ambivalence, Neutrality or Uncertainty. *Personality and Social Psychology Bulletin, 6,* 97-101.

Knauper, B. (1999). The Impact of Age and Education on Response Order Effects in Attitude Measurement. *Public Opinion Quarterly, 63(Fall)*, 347-370.

Knowles, E.S., & Nathan, K.T. (1997). Acquiescent responding in self-reports: Cognitive style or social concern? *Journal of Research in Personality, 31,* 293-301.

Knowles, E.S., & Condon, C. (1999). Why People Say Yes: A Dual-Process Theory Of Acquiescence. *Journal of Personality and Social Psychology*, *77*, 379-386.

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

Krosnick, J.A., & Fabrigar, L.R. (1997). Designing Rating Scales for Effective Measurement in Surveys. In: Lyberg, L.E., Biemer, P., Collins, M., de Leeuw, E.D., Dippo, C., Schwarz, N., & Trewin, D. (Eds.). *Survey Measurement and Process Quality* (pp. 141-164). NY: Wiley-Interscience.

Krosnick, J.A., & Fabrigar, L.R. (2003). *Designing Questionnaires to Measure Attitudes.* New York: Oxford Univ. Press.

Kruglanski, A.W., & Freund, T. (1983). The freezing and unfreezing of lay-interferences: Effects on impressional primacy, ethnic stereotyping and numerical anchoring. *Journal of Experimental Social Psychology, 19,* 448-468.

Lalwani, A.K. (2009). The Distinct Influence of Cognitive Busyness and Need for Closure on Cultural Differences in Socially Desirable Responding. *Journal of Consumer Research, 36(2),* 305-316.

Lam, T.C.M., & Stevens, J.J. (1994). Effects of Content Polarization. Item Wording and Rating Scale Width on Rating Response. *Applied Measurement in Education, 7(2),* 141-159.

Lazovik, G.F., & Gibson, C.L. (1984). Effects of Verbally Labeled Anchor Points on the Distributional Parameters of Rating Measures. *Applied Psychological Measurement, 8(1)*, 49-57.

Lee, A.Y., Aaker, J.L., & Gardner, W.L. (2000). The Pleasures and Pains of Distinct Self-Construals: The Role of Interdependence in Regulatory Focus. *Journal of Personality and Social Psychology*, *78*, 1122-1134.

Lee, C., & Green, R.T. (1991). Cross-cultural examination of the Fishbein behavioral intentions model. *Journal of International Business Studies, 22,* 289-305.

Liberman, N., Idson, L.C., Camacho, C.J., & Higgins, T.E. (1999). Promotion and Prevention Choices Between Stability and Change. *Journal of Personality and Social Psychology, 77*, 1135-1145.

Light, C.S., Zax, M., & Gardiner, D.H. (1965). Relationship of age, sex and intelligence level to extreme response style. *Journal of Personality and Social Psychology*, *2,* 907-909.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 44-53.

Lindell, M.K., & Whitney, D.J. (2001). Accounting for common method variance in cross-sectional research designs. *Journal of Applied Psychology*, *86(1),* 114-121.

Little, T.D. (2000). On the Comparability of constructs in cross-cultural research: A critique of Cheung and Rensvold. *Journal of Cross-Cultural Psychology, 31*, 213-219.

Little, T.D., Slegers, D.W., & Card, N.A. (2006). A Non-Arbitrary Method of Identifying and Scaling Latent Variables in SEM and MACS models. *Structural Equation Modeling, 13(1),* 59-72.

Lockwood, P., Jordan, C.H., & Kunda, Z. (2002). Motivation by Positive or Negative Role Models: Regulatory Focus Determines Who Will Best Inspire Us. *Journal of Personality and Social Psychology, 83*, 854-864.

Lorr, M., & Wunderlich, R.A. (1980). Mood states and acquiescence. *Psychological Reports, 46(1),* 191-195.

Lozano, L.M., Garcia-Cueto, E., & Muñiz, J. (2008). Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales. *Methodology, 4(2),* 73-79.

MacGregor, D. (1993). Time pressure and Task adaptation: alternative perspectives on laboratory studies. In: Svenson, O., & Maule, A.J. (Eds.). *Time pressure and stress in human judgment and decision making* (pp. 27-40). New York: Plenum Press.

Macrae, C.N., Milne, A.B., & Bodenhausen, G.V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology, 66,* 37-47.

Marín, G., Gamba, R.J., & Marín, B.V. (1992). Extreme response style and Acquiescence among Hispanics: The role of acculturation and education. *Journal of Cross-Cultural Psychology, 23(4*), 498-509.

Marsh, H.W., & Parducci, A. (1978). Natural Anchoring at the Neutral Point of Category Rating Scales. *Journal of Experimental Social Psychology, 14,* 193-204.

Marsh, H.W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*, 810-819.

Marsh, L.C., & Cormier, D.R. (2002). *Spline Regression Models*. Quatitative Applications in the Social Sciences. Sage Publications, Thousand Oaks.

Matell, M.S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement, 31*, 657-674.

Matell, M.S., & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert scale items? Effects of testing time and scale properties. *Journal of Applied Psychology, 56(6)*, 506-509.

Maule, A.J., & Hockey, G.R.J. (1993). State, Stress, and Time Pressure. In: Svenson, O., & Maule, A.J. (Eds.). *Time pressure and stress in human judgment and decision making* (pp. 27-40). New York: Plenum Press.

Mayo, R., Schul, Y., & Burnstein, E. (2004). 'I am not guilty' vs 'I am innocent': Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology*, *40,* 433-449.

Mazaheri, M., & Theuns, P. (2009). Effects of Varying Response Formats on Self-ratings of Life-satisfaction. *Social Indicators Research, 90(3),* 381-395.

McArdle, J.J., & McDonald, R.P. (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 234-251.

McClendon, M.J. (1991). Acquiescence and Recency Response-Order Effects in Interview Surveys. *Sociological Methods and Research, 20(1),* 60-103.

McCrae, R.R., & Costa, P.T. Jr. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In: Wiggins, J.S. (Eds.). *The Five-factor model of personality: Theoretical perspectives* (pp. 51-87). New York: Guilford.

McCroskey, J.C., Prichard, S.V.O., & Arnold, W.E. (1986). Attitude Intensity and the neutral point on Semantic Differential scales. *Public Opinion Quarterly*, *31*, 642-645.

McGee, R.K. (1967). Response set in relation to personality: An orientation. In: Berg, I.A. (Ed.). *Response set in personality assessment* (pp. 1-31). Chicago: Aldine Publishing Company.

McGuire, W.J. (1969). The nature of attitudes and attitude change. In: Lindzey, G., & Aronson, E. (eds.). *Handbook of social psychology* (vol. 3 pp. 136-314). Reading: MA: Addison-Wesley.

Menezes, D., & Elbert, N.F. (1979). Alternative Semantic Scaling Formats for Measuring Store Image: An Evaluation. *Journal of Marketing Research*, *16*, 80-87.

Messick, S. (1968). Response Sets. In: Sills, D.L. (Ed.). *International Encyclopedia of the Social Sciences Vol 13* (pp. 492-496). New York: Macmillan.

Messick, S. (1991). Psychology and methodology of response styles. In: Snow, R.E., & Wiley, D.E. (Eds.). *Improving Inquiry in Social Science* (pp. 161-200). Hillsdalen, NJ: Erlbaum.

Mick, D.G. (1996). Are studies of dark side variables confounded by socially desirable responding? The case of materialism. *Journal of Consumer Research, 23(sept),* 106-119.

Mirowsky, J., & Ross, C.E. (1991). Eliminating Defense and Agreement Bias from Measures of the Sense of Control: A 2x2 Index. *Social Psychology Quarterly, 54(2)*, 127-145.

Moser, C.A., & Kalton, G. (1972). *Survey Methods in Social Investigation.* London: Heinemann.

Mourali, M., Böckenholt, U., & Laroche, M. (2007). Compromise and Attraction Effects under Prevention and Promotion Motivations. *Journal of Consumer Research, 34*, 234-247.

Mullen, M.R. (1995). Diagnosing Measurement Equivalence in Cross-National Research. *Journal of International Business studies, 26(3),* 573-596.

Muthén, L.K., & Muthén, B.O. (2007). *Mplus User's Guide. Fifth Edition*. Los Angeles, CA: Muthén & Muthén.

Naemi, B.D., Beal, D., & Payne, S.C. (2009). Personality Predictors of Extreme Response Style. *Journal of Personality, 77(1),* 261-282.

Narayan, S., & Krosnick, J.A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, 60(1),* 58-88.

Nowlis, S.M., Kahn, B.E., & Dhar, R. (2002). Coping with Ambivalence: the Effect of Removing a Neutral Option on Consumer Attitude and Preference Judgments. *Journal of Consumer Research, 29*, 319-334.

Nunnally, J.C. (1967). *Psychometric Theory*. NY: McGraw Hill.

O'Donnell, R.D., & Eggemeier, F.T. (1986). Workload assessment methodology. In: Boff, K.R., Kaufman, L., & Thomas, J.P. (Eds.). *Handbook of Perception and Human Performance, Vol II,* New York: Wiley.

O'Muircheartaigh, C., Gaskell, G., & Wright, D.B. (1995). Weighing Anchors: Verbal and Numerical Labels for Response Scales. *Journal of Official Statistics*, *11*, 295-307.

O'Muircheartaigh, C., Krosnick, J.A., & Helic, A. (2000). *Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data*. Presented at Annu. Meet. Am. Assoc. Public Opin. Res., Fort Lauderdale, FL.

O'Neill, H.W. (1967). Response Style Influence in Public Opinion Surveys. *Public Opinion Quarterly, 31*, 137-157.

Ones, D.S., Viswesvaran, C., & Reiss, A.D. (1996). Role of desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81(6),* 660-679.

Osgood, C.E. (1941). Ease of individual judgment-process in relation to polarization of attitudes in the culture. *Journal of Social Psychology, 14*, 403-418.

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

Oyserman, D., Coon, H.M., & Kemmelmeier, M. (2002). Rethinking Individualism and Collectivism: Evaluation of Theoretical Assumptions and Meta-Analysis. *American Psychological Association, 128*, 3-72.

Paas, F., & van Merriënboer, J.J.G. (1994a). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6,* 51-71.

Paas, F., Tuovinen, J.E., Tabbers, H., & Van Gerven, P.W.M. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist, 38(1)*, 63-71.

Parducci, A. (1965). Category Judgment: A range-Frequency Model. *Psychological Review, 72(6),* 407-418.

Parducci, A., & Wedell, D.H. (1986). The Category Effect With Rating Scales: Number of Categories, Number of Stimuli, and Method of Presentation. *Journal of Experimental Psychology: Human Perception and Performance, 12(4),* 496-516.

Paulhus, D.L., Graf, P., & Van Selst, M. (1989). Attentional Load Increases the Positivity of Self-Presentation. *Social Cognition, 7(4),* 389-400.

Paulhus, D.L. (1991). Measurement and control of response bias. In: Robinson, J.P., Shaver, P.R., & Wright, L.S. (Eds.). *Measures of Personality and Social Psychological attitudes.* (pp. 17-59). San Diego: Academic Press.

Paulhus, D.L., & Lim, D.T.K. (1994). Arousal and Evaluative Extremity in Social Judgements: A Dynamic Complexity Analysis. *European Journal of Social Psychology, 24*, 89-99.

Payne, J.W., Bettman, J.R., & Johnson, E.J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14,* 534-552.

Peabody, D. (1962). Two components in Bipolar Scales: Direction and Extremeness. *Psychological Review, 69(2)*, 65-73.

Peabody, P. (1966). Authoritarianism scales and response bias. *Psychological Bulletin, 65*, 11-23.

Pham, M.T. (1996). Cue Representation and Selection Effects of Arousal on Persuasion. *Journal of Consumer Research, 22(March),* 373-387.

Pham, M.T., & Avnet, T. (2004). Ideals and Oughts and the Reliance on Affect versus Substance in Persuasion. *Journal of Consumer Research, 30*, 503-518.

Pham, M.T., & Higgins, T.E. (2005). Promotion and Prevention in Consumer Decision Making: State of the Art and Theoretical Propositions. In: Ratneshwar, S., & Mick, D. G. (Eds.). *Inside Consumption Consumer motives, goals and desires* (pp. 1-60). London/New York: Routledge.

Podsakoff, P.M., MacKenzie, S.B., & Podsakoff, N.P. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology*, *88*, 879-903.

Posavac, S.S., Sanbonmatsu, D.M., & Fazio, R.H. (1997). Considering the Best Choice: Effects of the Salience and Accessibility of Alternatives on Attitude Decision-Consistency. *Journal of Personality and Social Psychology, 72,* 253-475.

Posavac, S.S., Herzenstein, M., & Sanbonmatsu, D.M. (2003). The Role of Decision Importance and the Salience of Alternatives in Determining the Consistency between Consumer's Attitudes and Decisions. *Marketing Letters, 14,* 47-57.

Presser, S., & Schuman, H. (1980). The Measurement of A Middle Position in Attitude Surveys. *Public Opinion Quarterly, 44,* 70-78.

Preston, C.C., & Colman, A.M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power and respondent preferences. *Acta Psychologica, 104*, 1-15.

Rindfleisch, A., Malter, A.J., Ganesan, S., & Moorman, C. (2008). Cross-sectional versus longitudinal survey research: Concepts, findings and guidelines. *Journal of Marketing Research, 45(3)*, 261-279.

Rindskopf, D. (2003). Mixture or Homogeneous? Comment on Bauer and Curran. *Psychological Methods, 8(3),* 384-393.

Riordan, C.M., & Vandenberg, R.J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in a equivalent manner? *Journal of Management, 20,* 643-671.

Rohrmann, B. (2003). Verbal Qualifiers for Rating Scales: Sociolinguistic considerations and psychometric data. Working paper.

Rorer, L.G. (1965). The great response-style myth. *Psychological Bulletin, 63*, 129-156.

Ross, C.E., & Mirowsky, J. (1984). Socially - Desirable response and Acquiescence in a Cross-Cultural survey of Mental Health. *Journal of Health and Social Behavior*, *25*, 189-197.

Rossi, P.E., Gilula, Z., & Allenby, G.M. (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.

Rozin, P., & Royzman, E.B. (2001). Negativity Bias, Negativity Dominance and Contagion. *Personality and Social Psychology Review, 5(4*), 296-320.

Sanbonmatsu, D.M., & Fazio, R.H. (1990). The Role of Attitudes in Memory-Based Decision Making. *Journal of Personality and Social Psychology, 59(4),* 614-622.

Sawyer, A.G., Parker, W.M., & Sendak, P.E. (1979). The Role of Laboratory Experiments to test marketing strategies. *Journal of Marketing, 43(3),* 60-67.

Schaeffer, N.C., & Barker, K. (1995). Alternative methods of presenting bi-polar scales in telephone interviews: 1 to 7 vs. -3 to +3 and neutral vs. ambivalent. Presented at Annual Meeting Am. Assoc. Public. Opin. Res., Fort Lauderdale, FL.

Schaeffer, N.C., & Presser, S. (2003). The Science of Asking Questions. *Annual Review of Sociology, 29,* 65-88.

Schaeffer, E.M., Krosnick, J.A., Langer, G.E., & Merkle, D.M. (2005). Comparing the quality of data obtained by minimally balanced and fully balanced attitude questions. *Public Opinion Quarterly, 69(3),* 417-428.

Schimmack, U., & Colcombe, S. (2002). Theory of affective reactions to ambivalent Situations (TARAS): A cognitive account of negativity dominance. Working paper.

Schmitt, N., & Stults, D.M. (1985). Factors defined by negatively defined keyed items: The result of careless respondents. *Applied Psychological Measurement*, *9*, 367-373.

Schuman, H., & Presser, S. (1980). The Measurement of a Middle Position in Attitude Surveys. *Public Opinion Quarterly, 44(1)*, 70-85.

Schuman, H., & Presser, S. (1996). Questions and answers in attitude surveys: Experiments in question form, wording & content. New York: Academic.

Schwarz, N., Knauper, B., Hippler, H.J., Noelle-Neumann, E., & Clark, L. (1991). Rating Scales: Numeric Values May Change the Meaning of Scale Labels. *Public Opinion Quarterly, 55(4),* 570-582.

Schwarz, N., & Hippler, H.J. (1995). The numeric values of rating scales: A comparison of their impact in mail surveys and telephone interviews. *International Journal of Public opinion Research, 7,* 72-74.

Schwarz, N. (1999). Self-Reports: How the questions shape the answers. *American Psychologist, 54,* 93.

Sharma, S., & Weathers, D. (2003). Assessing generalizability of scales used in cross-national research. *International Journal of Research in Marketing, 20(1),* 287-295.

Shulman, A. (1973). A Comparison of Two Scales on Extremity Response Bias. *Public Opinion Quarterly, 37*, 407-412.

Simon, H.A. (1981). *The Science of the Artificial.* Boston: MIT press.

Simonson, I. (1989). Choice Based on Reasons: The Case of Attraction and Compromise Effect. *Journal of Consumer Research, 16,* 158-174.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and Commission in Judgment and Choice. *Journal of Experimental Social Psychology*, *27*, 76-105.

Srinivasan, V., & Basu, A.K. (1989). The metric quality of ordered categorical data. *Marketing Science, 8(3),* 205-230.

Steenkamp, J.B., & Burgess, S.M. (2002). Optimum Stimulation Level and Exploratory Consumer Behavior in an Emerging Consumer Market. *International Journal of Research in Marketing, 19,* 131-150.

Stening, B.W., & Everett, J.E. (1984). Response Styles in a Cross-Cultural Managerial Study. *Journal of Social Psychology, 122*, 151-156.

Strizhakova, Y., Coulter, R.A., & Price, L.L. (2008). The meanings of branded products: A cross-national scale development and meaning assessment. *International Journal of Research in Marketing, 25(2),* 82-93.

Suri, R., & Monroe, K.B. (2003). The Effects of Time Constraints on Consumers' Judgments of Prices and Products. *Journal of Consumer Research, 30*, 92-104.

Svenson, O., & Edland, A. (1987). Change of preferences under time pressure: Choices and Judgments. *Scandinavian Journal of Psychology, 29(4),* 322-330.

Svenson, O., & Benson, L. III (1993). Framing and Time Pressure in Decision Making. In: Svenson, O., & Maule, A.J. (Eds.). *Time pressure and stress in human judgment and decision making* (pp. 133-144). New York: Plenum Press.

Svenson, O., Edland, A., & Slovic, P. (1990). Choices and Judgments of incompletely described decision alternatives under time pressure. *Acta Psychologic, 75(2),* 153-169.

Swain, S.D., Weathers, D., & Niedrich, R.W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45(Feb),* 116-131.

Tourangeau, R., Rips, L. J., & Rasinski, K. (1984). *The Psychology of Survey Response*. NY: Cambridge University Press.

Tourangeau, R., & Rasinski, K.A. (1988). Cognitive Processes underlying the Context Effects in attitude measurement. *Psychology Bulletin*, *103*, 299-314.

Tourangeau, R., Rasinski, K., & D'Andrade, R. (1991). Attitude structure and belief accessibility. *Journal of Experimental Social Psychology, 27*, 48-75.

Tourangeau, R., Rips, L.J., & Rasinski, K. (2000). *The Psychology of Survey Response*. NY: Cambridge University Press.

Tourangeau, R., Couper, M.P., & Conrad, F. (2004). Spacing, Position and Order : Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly, 68 (3),* 368-393.

Tourangeau, R., Couper, M.P., & Conrad, F. (2007). Colors, Labels and Interpretive Heuristics for Response Scales. *Public Opinion Quarterly*, *71*, 91-112.

Trott, M.D., & Jackson, D.N. (1967). An Experimental analysis of Acquiescence. *Journal of Experimental Research in Personality, 2(4),* 278-288.

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

van Herk, H., Poortinga,Y.H., & Verhallen, T.M.M. (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.

Van Rosmalen, J., van Herk, H., & Groenen, P.J.F. (2010). Identifying response styles: A Latent-Class Bilinear Multinomial Logit Model. *Journal of Marketing Research, 47(1),* 157-172.

Velez, P., & Ashworth, S.D. (2007). The Impact of Item Readability on the Endorsement of the Midpoint Response in Surveys. *Survey Research Method, 1(2),* 69-74.

Watkins, D., & Cheung, S. (1995). Culture, gender and response bias: An analysis of responses to the Self-description questionnaire. *Journal of Cross-Cultural Psychology, 26,* 490-504.

Weathers, D., Sharma, S., & Niedrich, R.W. (2005). The impact of the number of scale points, dispositional factors and the status quo decision heuristic on scale reliability and response accuracy. *Journal of Business Research, 58*, 1516-1524.

Weems, G.H., & Onwuegbuzie, A.J. (2001). The Impact of Midpoint Responses and Reverse Coding on Survey Data. *Measurement and Evaluation in Counseling and Development, 34,* 166-176.

Wegner, B., Faulbaum, F., & Maag, G. (1982). Die Wirkung von Antwortvorgaben bei Kategorialskalen. *ZUMA-Nachrichten*, *10*, 3-20.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36(3),* 409–422.

Weijters, B., Geuens, M., & Schillewaert, N. (2009). The Proximity Effect: The Role of Interitem Distance on Reverse-Item Bias. *International Journal of Research in Marketing*, *26(1),* 2-12.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The Stability of Individual Response Styles. *Psychological Methods*, *15(1)*, 96-110.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27(3),* forthcoming.

Weijters, B., Geuens, M., & Schillewaert, N. (in press). The individual consistency of acquiescence and extreme response style in self-report questionnaires. Forthcoming in: *Applied Psychological Measurement*. Advance online publication doi: 10.1177/0146621609338593.

Weijters, B., Geuens, M., & Baumgartner, H. (2010). *The Effect of Language on Response Distributions in Likert Data.* Unpublished Working paper.

Weng, L.-J. (2004). Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability. *Educational and Psychological Measurement, 64(6),* 956-972.

Wildt, A.R., & Mazis, M.B. (1978). Determinants of Scale Response: Label versus Position. *Journal of Marketing Research, 15(2),* 261-267.

Winkler, J.D., Kanouse, D.E., & Ware, J.E. (1982). Controlling for Acquiescence Response set in scale development. *Journal of Applied Psychology*, *67*, 555-561.

Wong, N., Rindfleisch, A., & Burroughs, J.E. (2003). Do Reverse-Worded Items Confound Measures in Cross-Cultural Consumer Research? The Case of the Material Values Scale. *Journal of Consumer Research, 30(1),* 72-91.

Wright, P. (1975). Consumer Choice Strategies: Simplifying Vs. Optimizing, *Journal of Marketing Research, 12*, 60-67.

Yates, J.F., Lee, J.W., Bush, J.G.G. (1997). General Knowledge Overconfidence: Cross-National Variations, Response Styles and Reality, *Organizational Behavior and Human Decision Processes, 70 (2),* 87-94.

Zakay, D. (1990). The Impact of Time Perception Processes on Decision Making under Time Stress. In: Svenson, O., & Maule, A.J. (Eds.). *Time pressure and stress in human judgment and decision making* (pp. 59-72). New York: Plenum.

Zax, M., & Takahashi, S. (1967). Cultural influences on response style: Comparisons of Japanese and American college students. *The Journal of Social Psychology, 71,* 3-10.