Sociological Methodology
2014, Vol. 44(1) 369–399

DOI: 10.1177/0081175013516114
http://sm.sagepub.com

7

THE EFFECT OF LABELING AND NUMBERING OF RESPONSE SCALES ON THE LIKELIHOOD OF RESPONSE BIAS

Guy Moors*
Natalia D. Kieruj*
Jeroen K. Vermunt*

Abstract

Extreme response style (ERS) and acquiescence response style (ARS) are among the most encountered problems in attitudinal research. The authors investigate whether the response bias caused by these response styles varies with following three aspects of question format: full versus end labeling, numbering answer categories, and bipolar versus agreement response scales. A questionnaire was distributed to a random sample of 5,351 respondents from the Longitudinal Internet Studies for the Social Sciences household panel, of which a subsample was assigned to one of five conditions. The authors apply a latent class factor model that allows for diagnosing and correcting for ERS and ARS simultaneously. The results show clearly that both response styles are present in the data set, but ARS is less pronounced than ERS. With regard to format effects, the authors find that end labeling evokes more ERS than full labeling and that bipolar scales evoke more ERS than

*Tilburg University, Tilburg, The Netherlands

Corresponding Author:
Guy Moors, Tilburg University, Department of Methodology and Statistics, PO Box 90153, 5000LE Tilburg, The Netherlands
Email: guy.moors@uvt.nl

agreement style scales. With full labeling, ERS opposes opting for middle response categories, whereas end labeling distinguishes ERS from all other response categories. ARS did not significantly differ depending on test conditions.

Keywords

acquiescence response style, extreme response style, latent class analysis, full labeling, end labeling, measuring attitudes

1. INTRODUCTION

A survey researcher's ultimate dream is to develop unbiased measurements of opinions and attitudes. However, measurement error is hard to avoid, and when measurement error is not random, it is of great concern to any survey researcher. Response bias is a well-known source of nonrandom error, and Likert-type rating scales have been shown to be prone to all kinds of biases (Chan 1991; Greenleaf 1992; Kieruj and Moors 2010; Smith 1967). In this paper, we focus on the question of whether certain aspects of scale format—more specifically, the verbal and numerical labeling of the answer categories—affect a respondent's likelihood of providing biased responses.

Response bias is defined as response style whenever a person responds systematically to questionnaire items on some basis other than what the items were specifically designed to measure (Paulhus 1991). In this study, we focus on two commonly discussed response style behaviors in attitude research: (1) extreme response style (ERS) and (2) acquiescence response style (ARS). ERS is the tendency to choose only the extreme endpoints of the scale (Hurley 1998), and ARS is the tendency to agree rather than disagree with items regardless of item content (Van Herk, Poortinga, and Verhallen 2004). These response styles can be particularly problematic for comparative research; when left unevaluated, cultural differences may be misinterpreted as substantive differences in the construct being examined (Johnson et al. 2005). Developing scales that are not affected or are much less affected by response styles thus becomes important. Hui and Triandis (1989), for instance, found that cultural variations in ERS use were apparent when 5-point scales were used, but such variations vanished when 10-point scales were administered.

The process of constructing a rating scale is not as straightforward as it may first appear. There are several choices a researcher must make

when designing a rating scale. Deciding on the number of answer categories, for instance, is such an issue (Krosnick and Fabrigar 1997; Preston and Colman 2000; Symonds 1924). Similar problems arise with other aspects of rating scales, such as numbering and labeling of answer categories. A common distinction that is made when labeling occurs is that of ''full labeling'' and ''end labeling.'' In full labeling, all answer categories are verbally labeled (e.g., a five-point scale would consist of the labels ''completely disagree,''''disagree,''''do not disagree or agree,''''agree,'' and ''completely agree''), whereas in end labeling, only the end categories are labeled (e.g., ''completely disagree'' and ''completely agree''). We are interested in the question of whether the use of end labeling rather than full labeling evokes the use of ERS and ARS. Another topic of interest involves the issue of bipolar versus agreement scales and its influence on response behavior. These scales differ in their numbering of response categories, with full labeling presenting both negative and positive values, whereas end labeling presents only positive values. Finally, it seems to be common practice to attach numbers to response categories alongside the category labels. The question asked regarding this topic is whether presenting respondents with extra anchors in the form of numbers will yield different degrees of ERS and ARS.

We have organized this paper as follows. First, we present an overview of previous findings regarding the effect of scale format on data quality (i.e., reliability, validity, and response bias). Second, we discuss our research questions in more detail. Third, we briefly introduce the latent class model used in our analyses. Fourth, we investigate whether ERS and ARS are affected by full labeling versus end labeling, bipolar versus agreement scales, and the presence of numeric values of answer categories. Finally, we present our conclusions.

2. LITERATURE REVIEW: THE EFFECT OF SCALE FORMAT ON DATA QUALITY

In this research, we use a split-ballot design to study three interrelated topics regarding the labeling and numbering of attitude scales and their influence on the likelihood of response bias. Response bias refers to the issue of measurement validity in the sense that we question to what extent the relationship between indicators and latent content variables is biased by latent variables other than those intended. Deciding on

whether and how to label and/or number a response scale is a task faced by every survey research practitioner. Hence, whether the choices that have been made have consequences regarding response bias is of scientific as well as societal relevance. The first topic deals with full labeling versus end labeling of scales; the second topic revolves around the issue of numerical values and whether to use them to accompany the answer categories; the third topic deals with the comparison of agreement versus bipolar response scales. In the overview that follows, we discuss each of these topics on the basis of findings and perspectives from the literature. What unifies these studies across topics are two complementary theoretical propositions:

1. Survey question formats may increase response burden depending on how cognitively demanding they are. Originally coined by Simon (1955), the concept of ''satisficing'' has been used by Krosnick (1991), among others, to indicate that when response burden increases, respondents are more likely to satisfice rather than to optimize their responses. Consequently, response bias will increase.
2. In line with the principle of nonredundancy (Grice 1989), it is expected that respondents tend to look for cues on how to respond to survey questions in their attempts to give adequate answers. Consequently, they tend to assign meaning to all incentives given in the question format. Reflecting the satisficing principle, it is also expected that the less demanding the ''cue-looking'' task is, the less vulnerable a scale format is to response bias.

2.1. *Full Labeling versus End Labeling*

A considerable number of studies have been devoted to the issue of labeling all points or only the endpoints of a rating scale. Proponents of full labeling have argued that such labeling provides more information to respondents about how to interpret the scale (Johnson et al. 2005; Weng 2004). For this reason, the response load should be less burdensome in the case of full labeling, possibly leading to more accurate responses. In accordance with this reasoning, Dickinson and Zellinger (1980) showed that respondents prefer fully labeled scales to scales with end labeling. Furthermore, Arce-Ferrer (2006) showed that only one-fifth of respondents could correctly fill out the verbal center labels of an end-labeled scale, supporting the idea that respondents need help with interpreting

categories. In favor of end labeling, Krosnick and Fabrigar (1997) argued that numbered end-labeled scales may be less cognitively demanding than fully labeled scales because the former scales are more precise and easier to hold in memory. At the same time, other researchers argue that fully labeled scales show higher validity than scales with end labeling (Coromina and Coenders 2006; Krosnick and Berent 1993; Peters and McCormick 1966). This is contradicted by Andrews (1984), who found that validity was lower if full labeling rather than end labeling was used.

There have also been a limited number of studies focusing on the effect of end labeling versus full labeling on response style behavior. For example, Weijters, Cabooter, and Schillewaert (2010) found that fully labeled scales evoke more ARS and less ERS than scales that have end labeling. The latter finding points out that in the case of a fully labeled scale, the center categories become more salient to respondents than they are in scales in which only the end categories are labeled. In contrast, a study by Lau (2007) showed no significant effect of end labeling versus full labeling on ERS.

2.2. *Using Numerical Values to Accompany Answer Categories*

Whether the absence or presence of numerical labels affects data quality is a topic that has not yet been extensively studied, which may be because it is difficult to imagine how such absence or presence might affect response behavior. However, studies from different lines of research do show that alterations in the use of numbers can affect response behavior. For example, reversing the numerical values of a response scale (Krebs and Hoffmeyer-Zlotnik 2010) or making the verbal labels incompatible with the numerical labels (Hartley and Betts 2010; Lam and Kolic 2008; Rammstedt and Krebs 2007) are found to produce variations in response patterns. Because the results in these studies were at least partially dependent on the use of numerical values, the issue of whether to assign numerical values to category labels should probably not be dismissed without a closer look either.

Krosnick and Fabrigar (1997) argued that it is not usual for people to express their opinions in a numerical manner in daily life, and it may therefore not be a natural way for respondents to express themselves. Tourangeau, Couper, and Conrad (2007) found that rating scales with only verbal end labels and no numerical labels as opposed to scales that were fully labeled or numbered were prone to cues such as giving the

endpoints of the scale differing colors. This effect was entirely eliminated if labels for all categories were used (even if they were only numerical labels). These findings suggest that if no verbal or numerical labels are used, respondents become more susceptible to hints and thus more inclined to use other heuristics such as response style behavior to arrive at acceptable answers.

2.3. *Agreement Scales versus Bipolar Scales*

Agreement scales typically portray the gradual presence of a certain trait or the agreement with a certain position. For example, a scale consisting of seven answering categories uses numerical values that run from 1 to 7 (or 0 to 6), with category 1 representing disagreement and category 7 representing agreement. Selecting the lowest value on an agreement scale implies the absence of a trait or absence of agreement with a proposition. On the other hand, in the case of bipolar scales, a 7-point scale would have numerical values running from –3 to +3, with the lowest category not only implying the absence of a trait but also the exact opposite of the given trait. Several studies have shown that using bipolar scales instead of agreement scales can alter answering tendencies of respondents. For example, Schwarz et al. (1991) found that respondents who received a bipolar scale to rate the question ‘‘How successful would you say you have been in life?’’ used the lower categories considerably less often than respondents who received the agreement scale. They argue that respondents in the bipolar treatment interpret the lowest end label as the presence of failures, whereas the respondents in the agreement treatment interpret this same answering category as the absence of outstanding achievements. Other studies carried out by Schwarz and colleagues have yielded similar results (Schwarz 1999; Schwarz and Hippler 1995). The numerical values, the form, and probably other aspects of rating scales may appear as merely formal features to the survey constructer. What the literature review has shown is that such aspects of the scale may function as clues about how to go about answering questions by respondents. Response bias is then the outcome.

3. DEVELOPING THE RESEARCH QUESTION

In this study, we focus on ERS—the tendency to choose the endpoints of a scale—and ARS—the tendency to agree with questions—and we

attempt to establish whether these types of response styles are affected by certain format issues. Given the previous findings in this area of research, we were able to formulate some hypotheses regarding the effect of response format on response styles. First, because labeling only the ends of a scale makes the end categories more salient and clearer than the center categories, we expect respondents to be more inclined to use ERS when presented with an end-labeled scale than when presented with a fully labeled scale. Second, we expect respondents to make more use of ARS if the meaning of answering categories is less clear, that is, if only the endpoints are labeled and no numerical labels are used. Third, we expect that the type of numbering of end-labeled scales will affect the likelihood of ARS. Bipolar scales make use of both negative numbers, indicating levels of disagreement, and positive numbers, indicating agreement. Respondents will be less likely to use the answering categories in the lower half with this format compared with agreement scales that use only positive integers.

4. DATA, DESIGN, AND METHOD

4.1. *Participants*

Our split-ballot experiment was implemented in the Longitudinal Internet Studies for the Social Sciences (LISS) Web panel of CentERdata (http://www.lissdata.nl/lissdata/Home), which is a Dutch household panel that was initiated in 2008 and consists of 8,044 participants. We would like to stress that the quality of the sampling strategy matches the high standards set in regular face-to-face surveys. In contrast with voluntary Internet panels, the LISS panel includes households that were recruited using a random-sampling design. Participants who did not have personal computers and/or Internet access were granted Internet facilities so that these participants would not automatically be excluded from the panel.

Our attitudinal scales were fielded in February 2009 and filled out by 5,351 respondents, a response rate of 65 percent (American Association for Public Opinion Research code RR6). The sample was 46.1 percent male and 53.9 percent female. Ages ranged from 16 to 95 years, with a mean age of 47. A subsample of 3,266 respondents filled out the questionnaires with the five test conditions used in this research. A split-ballot experiment was adopted to ensure that experimental groups differed only in treatment. Although unlikely, we checked whether

differential nonresponse might have distorted the comparability of experimental groups. No significant differences in age, gender, education, and marital status between groups were found.

It is worth mentioning that the design of the LISS study reduces rather than emphasizes the risk for satisficing response behavior. Satisficing increases with the length of a questionnaire (response burden) and when respondents are less familiar with the survey context. LISS respondents had been familiarized with answering survey questions on several occasions prior to answering our set of questions. Furthermore, only short questionnaires were used in this Web survey, and as a consequence, fatigue or loss of interest was less likely to occur than with long questionnaires.

4.2. *Questionnaire*

This research requires the use of balanced sets of items. The minimum requirement to measure ARS is that at least one scale be partially balanced (Billiet and McClendon 2000), because it can only be said that respondents exert ARS if they agree with both positively and negatively worded items regardless of item content. Balanced scales are hard to find, presumably because they are difficult to operationalize in many situations. We have selected four items from two scales measuring attitudes toward environmental issues (α values ranging from .707 to .762) and attitudes toward risky driving (α values ranging from .740 to .766), as shown in Table 1. The items from the environment scale were adopted from a revised New Ecological Paradigm Scale by Dunlop et al. (2000). The driving scale was based on four items from a ''risky drivers'' attitude scale (Yilmaz and Çelik 2006). Both of these scales use an agree-disagree format. Although it has been argued that this format is susceptible to ARS, we maintained the format in this study not to arouse ARS but mainly because the format is still overwhelmingly used in today's survey practice. Furthermore, because our study varied in the way these labels were used, we were able to provide additional insights on the issue.

Each set of four items provided a fully balanced set, meaning that we included as many positively worded items as negatively worded items in the scales. Preliminary analyses revealed that the last item from the risky driving scale needed to be omitted, because virtually all respondents chose the sixth or seventh answering category. Our scales were

Table 1. Scales and Corresponding Items

1	(a) Humans are severely abusing the environment (–).
	(b) The balance of nature is strong enough to cope with the impacts of modern industry (+).
	(c) The so-called ‘‘ecological crisis’’ facing humankind has been greatly exaggerated (+).
	(d) The balance of nature is very delicate and easily upset (–).
2	(a) Safe drivers can exceed posted speed limits (+).
	(b) Driving above posted speed limits is not a problem if the conditions are proper (+).
	(c) Even if you have good driving skills, speeding is not OK (–).
	(d) It is always risky to drive after drinking alcohol (–).

Note: The original questions were translated from the Dutch by the authors.

positioned at the end of a larger questionnaire that took respondents about 20 minutes to fill out. This questionnaire was electronically sent to the panel members in February 2009 and was accessible for a one-month period. Three reminders were sent during this time.

4.3. *Design*

In the setup of this study, respondents were randomly assigned to five formats that varied in the use of labeling and numbering of response scales to the same set of questions. Each response scale included seven ordered answer categories that differed in the following ways:

Format 1: full labeling with numerical values
Format 2: full labeling without numerical values
Format 3: end labeling with numerical values
Format 4: end labeling without numerical values
Format 5: end labeling with bipolar numerical values

The fully labeled scales were labeled ‘‘totally disagree,’’‘‘disagree,’’‘‘disagree somewhat,’’‘‘neither disagree nor agree,’’‘‘agree somewhat,’’‘‘agree,’’ and ‘‘totally agree,’’ whereas the end-labeled scales were labeled only ‘‘totally disagree’’ and ‘‘totally agree’’ at the ends. Numerical values ran from –3 to +3 in the bipolar numbered scale treatment and from 1 to 7 in the agreement numbered treatments with numerical values. The starting value of 1 was chosen rather than 0 to avoid respondents misinterpreting the latter as identifying the ‘‘absence’’ of a

value on a scale. In the bipolar numbered scale, the value of 0 most clearly identified the neutral position. The end labeling with numerical values treatment (format 3) had about twice as many respondents assigned to it, which was done to anticipate future research. Other aspects of question format were held constant across test conditions following the standard ruling of the LISS procedure to which the respondents were accustomed, that is, no explicit ''don't know'' option and excluding the possibility of revising previously given responses. We did not want to depart from this procedure to avoid arousing suspicion regarding our experiment.

4.4. *Model*

We used a latent class confirmatory factor model originally proposed by Moors (2003) and extended by Morren, Vermunt, and Gelissen (2011) to detect and control for ERS. The first of these models suffered from a lack of parsimony, because all effects of the latent variables on response variables were defined as nonmonotone, resulting in $C - 1$ parameters per response variable, where C is the number of response categories. The extended model demonstrated that the complexity of the original model could be reduced by defining a monotone relationship between the latent content variables and the response variables and a nonmonotone relationship in the case of ERS. In this research, we further extended the model by simultaneously estimating ERS as well as ARS. Modeling ARS was possible by imposing equal-sign monotone effects on all response variables so that the prevalence of effects on items was equal in both positively and negatively worded items. The resulting model was a restricted multinomial logit model that can be written as a linear model for the logit of responding in category $c + 1$ instead of c, as follows:

$$\log \frac{P\left(Y_{ij} = c+1 | F1_i,\, F2_i,\, ERS_i,\, ARS_i\right)}{P\left(Y_{ij} = c | F1_i,\, F2_i,\, ERS_i,\, ARS_i\right)} = \left(\beta_{0jc+1} - \beta_{0jc}\right)$$
$$+ \beta_{1j} F1_i + \beta_{2j} F2_i + \left(\beta_{3c+1} - \beta_{3c}\right) ERS_i + \beta_4 ARS_i,$$

in which Y_{ij} denotes the response of individual i to rating item j, $F1$ and $F2$ denote the latent content factors, and ERS and ARS denote the latent response styles. This model shows how the parameters relate to the adjacent-category logits. The parameters β_{1j}, β_{2j}, and β_4 are effects on

the adjacent-category logits, and they define the monotone relationship between *F*1, *F*2, *ARS*, and *Y*. The term $(\beta_{3c+1} - \beta_{3c})$ defines the nonmonotone relationship of *ERS* with *Y* and implies the estimates of $C - 1$ β-parameters, where *C* is the number of response categories.

In this research, the latent class content factors referred to the two ''environment'' and ''risky driving'' attitude scales, and items were allowed to load only on their corresponding attitudinal factors. In the case of the ERS and ARS, all items loaded on these style factors because all such items were supposed to be affected by response bias. Content factors were allowed to correlate with one another, but style factors were not. This way, we were able to filter out the influences of response styles on attitudinal dimensions.

The latent class factor approach was particularly chosen because this method allows for estimating separate effects of a latent ''response style'' factor on each response category of the observed response items. As such, respondents' preferences for certain answer categories might show up. In this research, ERS was the response pattern that emerged. In the case of the two content factors and ARS, we simplified the model by imposing ordinal restrictions resulting in a single effect estimate per item. All models were estimated using the software program Latent GOLD 4.5 (http://www.statisticalinnovations.com), developed by Vermunt and Magidson (2005).

At this point, there may be concern that our model conflates substantive responses with response styles since it resembles the ''unmeasured latent method construct'' approach, which Richardson, Simmering, and Sturman (2009) advised against using because it works only when one can be sure that the bias is present in the data. We agree that estimating a response bias with a latent method factor can be dangerous in the sense that it might capture content information about the concepts being measured. To avoid this, it should be taken into account that a latent response style factor can be interpreted as a style factor only if the response pattern is not consistent with the content that is measured (Billiet and McClendon 2000). Hence, ARS can be unequivocally diagnosed only if respondents tend to agree with both negatively and positively worded items measuring the same concept, a situation achieved in this research by imposing the positive effects of ARS on all items.

As far as ERS is concerned, the following features of our model reduce the likelihood of confounding substantive responses with ERS: (1) ERS is uncorrelated with the content factors, (2) ERS is measured

as a single latent class factor influencing responses to sets of items that differ in substantive meaning, (3) ERS is the outcome of an exploratory search on which response categories are preferred systematically more (or less) than other categories independent of content, and (4) including ERS decreases the distance between extreme responders and endpoint avoiders without necessarily changing their relative positions on the content dimensions. Additional evidence that the applied strategy does not conflate substantive responses with response styles is presented in the Appendix, in which we demonstrate that relative positions on the content factors change only slightly when ERS is taken into account, making the relative distance between ''avoiders of extremes'' versus ''endpoint responders'' somewhat more narrow without completely eliminating it. Furthermore, in previous research (Kieruj and Moors 2013), it is demonstrated that an ERS factor defined by the latent class factor model correlated with an ERS index calculated as the sum of extreme responses in a larger set of uncorrelated items. The latter index accommodates Greenleaf's (1992) procedure to define a contentless measure of ERS. Correlations ranged from .371 to .493, which is fairly high because these questions were administered at other waves in the LISS panel. Weijters et al. (2010) adopted Greenleaf's procedure of defining a contentless measure of ERS by counting extreme scores, and they calculated a similar index to measure ARS. The problem with these kinds of indices is that they are deterministic rather than model based. There are two benefits to our model-based approach: (1) Model fit comparisons allow us to research whether including response style factors improve model fit, hence evaluating whether they did affect the measurement of substantive scales, and (2) the method partitions the responses on items into a part affected by content (true score) and a part affected by style (response bias).

4.5. *Model Specifications*

In the previous section, we elaborated on the model used in this research by defining the general model. The empirical analyses implied further model specifications that are specified in this section. As a general rule model specification implies model fit comparisons. In latent class analysis, decisions on model selection are based on log-likelihood (LL) estimates and information criteria. In this research, we make use of the Bayesian information criterion (BIC), which simultaneously estimates

Table 2. Model Fit Comparisons

Model	LC Factors Included	Npar	LL	ΔLL	BIC
1.1	No[a]	42	–38,675		77,690
1.2	Content	54	–35,322	3,353	71,082
1.3.1	Content + ARS	57	–35,235	3,441	70,930
1.3.2	Content + ERS	62	–34,010	4,665	68,522
1.4	Content + ERS + ARS	65	–33,962	4,713	68,449

Note: Results are from the pooled data set. ARS = acquiescence response style; BIC = Bayesian information criterion; ERS = extreme response style; LC = latent class; LL = log-likelihood; Npar = number of parameters.

[a]Reference LL value for ΔLL comparisons.

the fit of the model alongside its parsimony (number of parameters relative to the other models it is compared with) and partly compensates for sample size. The lower the BIC value, the better the balance between fit (LL) and complexity (number of parameters, Npar).

The basic model refers to a single sample, whereas our split-ballot approach involves five samples parallel to the five test conditions. As a way of screening the data, we first ran separate analyses on each of the five samples, but pooling the data and adopting a multiple-group comparison approach, in which the five conditions define the group variable, is the more solid way of testing our hypotheses. If it makes no difference which of the five response scale formats is used, then the measurement model would be the same in each treatment, and the group variable would have no effect on the latent class factors. By estimating alternative models in which effects of the group variable on the measurement part of the model are included and by comparing the model fit, we can decide on the effect of the five scale formats on response bias. How this works will become clear when we provide details on the alternative models we compared.

Prior to estimating whether test conditions affect the occurrence of response style biases, we needed to be sure that adding ERS and ARS to the model was really needed. For that purpose, we compared a reference model (model 1.1 in Table 2) that did not include latent factors (the one-class model) with four other models. First, a model with content factors and no style factors (model 1.2) was compared with model 1.1, with Table 2 showing that adding the content factors is a major improvement in terms of BIC and ΔLL. Second, the reference model

Table 3. BIC Values of Models with Varying Equality Restrictions

Model	Equality Restrictions	Npar	LL	BIC(LL)
2.1	No restrictions	72	–34,603	69,789
2.2	Restrictions on style factors	65	–33,962	68,449
2.3	Restrictions on all factors	60	–36,572	73,629

Note: Results are from the pooled data set. BIC = Bayesian information criterion; LL = log-likelihood; Npar = number of parameters.

was compared with a model that adds an ARS factor (model 1.3.1) and a model that adds an ERS factor (model 1.3.2) to the content factors. As can be seen, adding the ERS factor leads to a substantially bigger improvement in terms of BIC and ΔLL than adding the ARS factor. For that reason, it can be concluded that ERS constitutes a more important response style factor than ARS. Finally, we showed that in model 1.4, the BIC and ΔLL improve even more if both style factors are included in the model. The results presented in Table 2 make use of the pooled data set, but similar results were found when separate analyses were conducted for each treatment. Given that the model that includes both ERS and ARS was found to be the better fitting model in each separate treatment, we proceed with model 1.4. Our first conclusion drawn from the latter finding is that none of the tested response scale formats are immune from response biases.

Having selected a starting model in the first step, the next question is whether we can further simplify the model by imposing equality constraints on the effect of the latent variables on the items. After all, the starting model is still complex even with imposing ordinal restriction on the relationship of the content factors and ARS factor with the response items. In the case of ERS, we would have 7 (number of items) times 6 (7 − 1 response categories) parameter estimates in our measurement model. Fixing effects to be equal on all items would dramatically reduce the number of parameters to interpret. In Table 3, we compare model 2.1, in which these effects are set equal in all latent class factors, with model 2.2, in which this equality constraint is applied only to the style factors. Model 2.3 includes no such equality constraints. The results show that a model with equality restrictions on the style factors is the most appropriate model according to the BIC. We choose this model, which implies equal effects of ERS and ARS across all items, as our

starting model. In addition, we favor this model because conceptually it allies with those who argue that ERS should occur consistently across different concepts, independent of content (Greenleaf 1992). A similar reasoning can be adopted in the case of ARS. Of course, we can argue that items may evoke different levels of response style bias, but empirically, the model corresponding to this reasoning did not improve in terms of BIC compared with the model assuming equal effects for ERS and ARS (model 2.3 in Table 3).

Whether the effect of ERS and ARS is different depending on response scale format is tested in the step that follows, in which we adopt a multiple group comparison approach using the pooled data set. Pooling the data of the split-ballot experiment was feasible because all respondents did answer the same questions on a seven-point scale; only the labeling and numbering of the categories differed across groups. In the pooled data set, we assigned all respondents to a group variable, to indicate the different treatment they had received. Including this group variable in the selected model can be done at different levels. When an effect of the group variable on the latent class factors is included, we can determine whether the test conditions (i.e., differences in labeling and numbering of response categories) lead to differences in distribution of the latent class factors (structural model). The direct effects of the group variable on particular items indicate that response format influences responses to specific items independent of the latent variables defined in the model. This might be interpreted as item-specific response scale effects (measurement model). More interesting with respect to the research questions asked is whether the grouping variable interacts with the latent class factors in explaining responses to the question items. In particular, we are interested in whether the effect of ERS and/or ARS on response items depends on test conditions.

As can be seen in Table 4, we start with the most complex model of the five models presented, model 3.1, which includes the direct effects of the group on all latent factors (structural model), the direct effects of the group variable on the items, and the interaction effect of the group variable with the latent class factors on the items. This complex model is then compared with models in which particular effects are omitted, hence decreasing complexity. In model 3.5, no effect of the group variable is included, suggesting a fully homogeneous measurement model with no impact of response scale format whatsoever.

Table 4. The Effect of Test Conditions (Group Variable) on the Measurement of LC Factors

Model	Model Specification	Npar	LL	BIC(LL)
3.1	F1 + F2 + ERS + ARS + group + F1 × group + F2 × group + ERS × group + ARS × group	165	–33,675	68,685
3.2	F1 + F2 + ERS + ARS + F1 × group + F2 × group + ERS × group + ARS × group	137	–33,701	68,511
3.3	F1 + F2 + ERS + ARS + ERS × group + ARS × group	109	–33,731	68,345
3.4	F1 + F2 + ERS + ARS + ERS × group	105	–33,734	68,318
3.5	F1 + F2 + ERS + ARS	81	–33,878	68,412

Note: Structural part of all models includes group effects on all LC factors. ARS = acquiescence response style; BIC = Bayesian information criterion; ERS = extreme response style; F = factor; LC = latent class; LL = log-likelihood; Npar = number of parameters.

The complex model 3.1 has a considerably higher BIC value than the simpler model 3.5. We estimated several models that fall between the heterogeneous model 3.1 and the homogeneous model 3.5. Starting from model 3.1, we omitted the direct effect of the group variable on items (model 3.2). The model showed improvement over model 3.1 in terms of the BIC, but it was still less appropriate than model 3.5. Model 3.3 excluded the interaction terms of the content factors, which all proved to be nonsignificant in the previous model (F1 × group and F2 × group, $p > .10$). The model showed further improvement, with BIC values lower than the first model as well as the last. Note that model 3.3 directly relates to the research questions asked because it checks whether the effect of ERS and ARS on the items differs according to the test conditions. At the same time, the fact that the BIC value of model 3.3 is lower than that of model 3.2 implies that the effect of content factors on the response items does not depend on the response format of scales. By looking more closely at the estimates of model 3.3, we can further simplify the model by dropping the ARS × group interaction, which is not significant. This is confirmed in model 3.4. The ERS × group interaction, on the contrary, cannot be dropped because the model fit deteriorated, as can be seen by comparing model 3.4 with model 3.5. Hence, the most appropriate model in terms of the BIC includes direct effects of the group variable (i.e., the effects of response formats on the latent variables) and a group-specific ERS effect on the

Table 5. The Effect of ERS and ARS on the Response Items (Logit Coefficients)

Response Style		β	*SE*
ERS	rc1	5.222	0.328
	rc2	–1.025	0.138
	rc3	–2.650	0.194
	rc4	–2.293	0.274
	rc5	–2.466	0.166
	rc6	–1.085	0.121
	rc7	4.298	0.227
ARS		1.091	0.121

Note: There are equal effect parameters on all items. rc = response category.

item responses (i.e. model 3.4).[1] The interpretation of the effect parameters in this model is the subject of the next section.

5. RESULTS

The final selected model (model 3.4) indicates the presence of ERS and ARS in all treatments, the effect of test conditions on response styles, and test-specific ERS effects on item responses.

5.1. *The Effect of ERS and ARS on Item Responses*

Table 5 shows the logit effect (β values) of ERS and ARS on the response items (from the selected model 3.4), which were both significant ($p < .001$). Recall that we fixed these effects to be equal in all items. Separate effects of ERS on each response category were estimated, and the results show exactly the pattern we expected to emerge in the case of ERS (i.e., high positive values for the end categories with negative values for the categories lying in between). In fact, labeling this pattern ERS is the only possibility because the method as such allows only revealing response scale point preferences among respondents independent of the content of items.

Table 5 also shows the significant effect of the ARS factor on the item responses. (Note that in the case of ARS, we obtain only one effect parameter given its ordinal effect on items [$p < .001$].) However, as previously reported in Table 2, model fit did substantially increase by including ERS but only marginally by adding ARS. Furthermore, by

Table 6. Group (Question Format) Effects on the Latent Class Extreme Response Style Factor (Logit Coefficients)

Treatment	β	*SE*
End labeling + numbers	0.363	0.116
End labeling + no numbers	0.720	0.151
Full labeling + numbers	–1.218	0.183
Full labeling + no numbers	–0.941	0.151
Bipolar scale	1.124	0.163

transforming the logit (β) parameters to their odds ratios, we can calculate the change in log-odds of item responses when comparing meaningful categories of the LC-style factors. In the case of ARS, the odds ratio for $c + 1$ versus c equals 2.977 (exp [1.091]), which means that the likelihood of ARS almost triples when moving from the lowest to the highest class of ARS. With ERS, two comparisons can be made between the odds ratios of the two extreme response categories and their adjacent categories. The odds of the lowest relative to the odds of its adjacent category is 516.461 (exp [5.222 + 1.025]), whereas comparing the two highest categories gives a value of 217.674 (exp [4.298 + 1.085]). Given the rather weak effect of ARS and the fact that the effect of ARS on the response items did not depend on test conditions (the group variable), we have to conclude that the ARS latent factor does seem to capture some kind of ''acquiescence noise'' but is of lesser substantive importance. Inevitably, this conclusion holds only to the items asked in this particular research. On the other hand, ERS is prominently present.

5.2. *The Effect of Scale Format (Numbering and Labeling) on Response Styles*

In model 3.4, the nominal group variable, indicating the five test conditions, showed a significant effect only on ERS ($p < .001$), but not on the other latent factors. This indicates that the prevalence of ERS depends on numbering and/or labeling of response scales because this is what defined the test conditions. As can be seen in Table 6, the bipolar scale has the highest positive effect parameter, indicating that bipolar scales seem to evoke more ERS than agreement-style scales. Also, the end-labeling treatments show positive effects, whereas the full-labeling

treatments reveal negative effects, indicating that end labeling evokes more ERS than full labeling does.

5.3. *Test-specific ERS Effects on Item Responses*

The selected model (3.4) includes an overall effect of ERS on the response items, complemented with group-specific relative deviations from this overall effect (the ERS $\times$ group interaction effect for which the midpoint was set as the reference category). In Figure 1, we have added these group-specific deviations to the overall effect of ERS on response items to ease comparisons. The midpoint of the response scale defines the reference category, for which the value is set to zero.

The overall effect is strongly present in all treatments ($p < 0.001$), but there are some group-specific deviations as well (ERS $\times$ group effect significant at $p < 0.001$). The method-specific ERS effects on the response items relate to the estimated effects of the two categories adjacent to the extreme responses. When full labeling is used, the estimates of these ''agree'' and ''disagree'' categories are closer to the values of the other intermediate response categories than to the values of the endpoints. With end labeling, the adjacent ''agree/disagree'' categories fall much more in between the extreme and the middle categories. Hence, with end-labeling, the opposite of extreme response preference is defined by preferences for the midpoint categories; whereas in the case of full labeling, the style factor should be interpreted as contrasting extreme response scale preference versus a preference for either category in between the extreme ones. Regardless of these method-specific ERS effects on response categories, the overall effect of ERS on response items is overwhelming.

5.4. *The Implications for Sociological Survey Research Practice*[2]

The principal finding of this research indicates that it is unlikely that ERS can be avoided by carefully designing the response format of agreement scales. From a sociological point of view, we can ask whether this issue is of true concern. Before drawing the main conclusions from this study and discussing its contribution, we highlight some additional and complementary results that underscore the usefulness of our findings. We can do this by comparing results from models with and without response style factors, including the sociodemographic covariates age

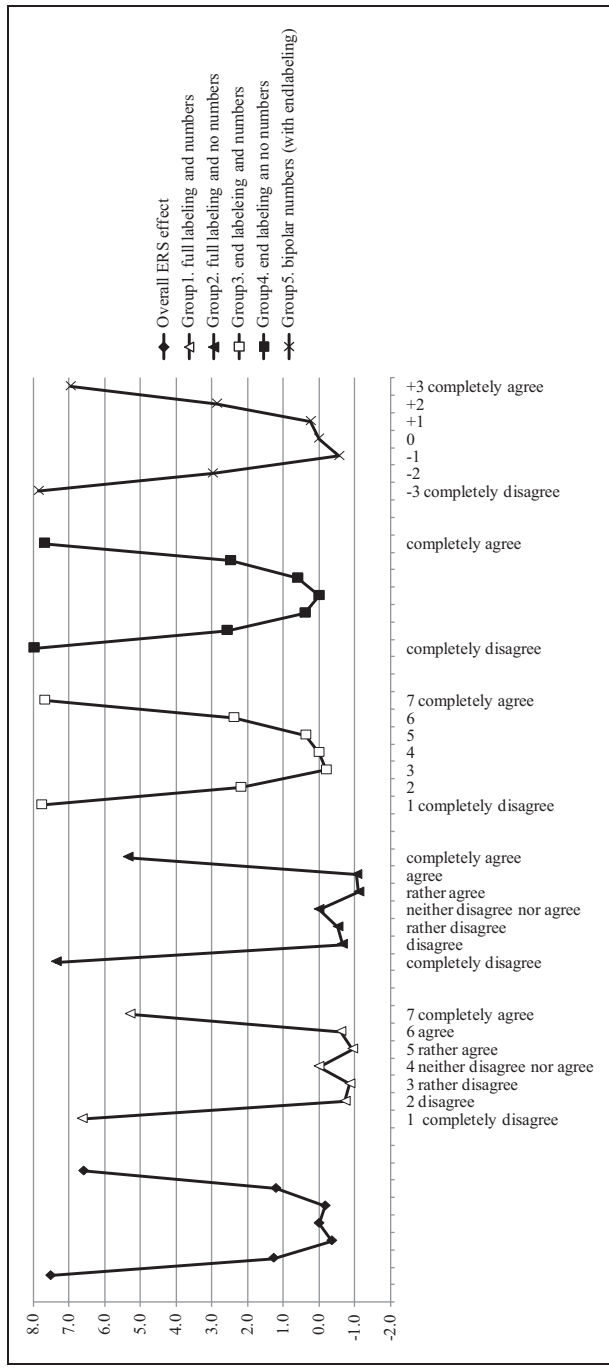

Figure 1. Overall and test-specific effects of extreme response style on response items.

group, education, and gender. Our argument is that the findings of this study are of concern to survey research practitioners when (1) significant improvement in measurement of latent variables has been achieved, (2) response styles correlate with sociodemographic characteristics, and/or (3) the relationship of covariates with content latent factors differs according to whether response style behavior is taken into account.

The improvement in model fit of the measurement model when including response styles was discussed in previous sections. Two complementary results not yet presented are that the effect of the latent content factors on both positively and negatively worded items are more similar in magnitude than when ERS and ARS are not included. This indicates that the balanced nature of the items is better articulated in the extended measurement model including response styles. Furthermore, the association between the two content factors also decreases when adding ERS and ARS. Hence, part of the relationship between content factors was spurious with regard to response style behavior.

None of the covariates are significantly related to ARS, confirming our previous interpretation of ARS as a kind of noise factor rather than a substantive measure of ARS in this data set. On the other hand, ERS is significantly and positively related to age. Both the ''environmentalism'' and ''safe drivers'' scales indicate that the issue must be relativized; hence, the higher the score, the less concern is expressed with environment and safe driving. Relativizing ''environmentalism'' correlates significantly with education and age groups. The higher the educational level, the less relativism is expressed. This effect is more articulated in the model including ERS and ARS compared with the model without these response styles. The relationship of age groups with the environmentalism scale follows a U-shaped form, with the age groups in the ranges 35 to 44 and 45 to 54 years revealing the highest levels of concern and the oldest (65 and older) and youngest (24 and younger) the least concern. Only minor differences are observed in this pattern comparing models with and without response styles. Relativizing ''safe driving'' is significantly related to gender and age. Men reveal fewer safe driving concerns, a finding that is somewhat more articulated when response styles are included. Relativizing ''safe driving'' decreases with age. Only the youngest group (aged 15 to 24) deviates from this pattern by having a lower level of relativizing than the 25 to 34 age group. The contrast between the oldest and youngest

age groups in ‘‘safe driving’’ is somewhat more pronounced in the model that includes response styles.

In summary, these complementary analyses demonstrate that it is worthwhile researching the impact of response style behavior in measuring attitudes as well as in estimating the effect of covariates on response styles and content factors. Metaphorically speaking, if preventing response style bias is (too) difficult, diagnosing and controlling its detrimental effects is valuable.

6. DISCUSSION

We set out to investigate whether certain aspects of question format (i.e., variations in labeling and numbering of response categories) would influence the use of ERS and ARS. Using a latent class model, we found a strong presence of ERS across all treatments. ARS was present as well, although less convincingly than ERS even when fully labeled agree-disagree scales were used. The latter might come as a surprise because agree-disagree formats of response scales are regarded as very vulnerable to ARS. We can think of several reasons why we found less evidence of ARS than ERS. First, we have to acknowledge that by presenting a balanced set of items, including both positively and negatively worded items, a kind of ‘‘preventive’’ check for ARS is implemented by design, which is not the case for ERS. Including a balanced set is necessary to be able to distinguish ARS from content-related response patterns. Unless respondents are careless in reading questions, balanced sets make respondents more aware of the fact that they should answer consistently across questions. Given that the LISS panel members can be considered as trained respondents, the likelihood of careless responses is rather small. Building on this thought, it might very well be that factors other than question format evoke ARS.

Long, exhaustive questionnaires in face-to-face interviews, for instance, might induce ARS to a greater extent. In addition, we should equally acknowledge that finding ARS in a balanced set of items by definition implies nonconsistent responses, whereas ERS can be perfectly in accordance with the content of the questions asked. Always ‘‘totally agreeing’’ or ‘‘totally disagreeing’’ with items instead of just ‘‘agreeing’’ or ‘‘disagreeing,’’ as an extreme responder would do, is less of a mistake than ‘‘agreeing’’ with an issue that a responder should have disagreed with, as might happen with an acquiescent responder.

ERS was strongly present in each treatment. Hence, question format in the form of labeling and numbering could not prevent the occurrence of this response style. However, it was also found that the amount and type of ERS used by respondents did, to some extent, differ across treatments. In line with our hypothesis, end labeling evoked more ERS than full labeling, which we expected because end labeling draws attention to the two extreme categories, which are thus clearer in meaning to respondents than the categories in between. In the case of full labeling, all categories are more or less equally clear to respondents, which means that no preference for certain categories is facilitated simply by labeling one category and not the other. In addition, as we expected, bipolar scales turned out to evoke more ERS than agreement scales. This suggests that bipolar scales (e.g., running from –3 to +3) may be harder to use than agreement-style scales. Furthermore, in daily life, people are much more accustomed to grade things using positive values only (with zero indicating a truly bad score) rather than giving negative values. As such, offering negative response values may be confusing.

Apart from the effect of response scale format on the amount of ERS used by respondents, we also found variations in the shape of ERS across formats. Variations were found in the contrast made between the extreme answering categories and the adjacent categories. Full labeling resulted in contrasting extreme category preference against any other preference, whereas with end labeling extreme responding is opposed by midscale preferences. Nevertheless, the most significant finding of our study is that ERS was consistently and strongly present in each treatment regardless of format issues. Therefore, we suspect that ERS is a kind of personal style that particular respondents exhibit when answering questions. This is in line with a previous study that showed that ERS is, for the most part, a stable trait that holds across different questionnaires and time (Kieruj and Moors 2013). As a result, our study seems to indicate that ERS cannot be prevented by adjusting question formats so that they will not trigger ERS in respondents. Instead of preventing the occurrence of ERS, then, it becomes necessary to dispose of a way to correct for ERS in measurement models. The latent class confirmatory factor model presented in our study serves this purpose. Of course, we do not exclude the possibility that there might be a question format that is largely unaffected by ERS. This research merely indicated that variations in numbering and labeling did not make a difference.

There were also some unforeseen results, such as the fact that we were not able to draw firm conclusions regarding ARS, because it was less prominently present in this research than reported by other researchers (Billiet and Davidov 2008; Billiet and McClendon 2000) using similar questionnaires. Nevertheless, this research found evidence that ERS influences the responses to attitudinal questions regardless of which type of labeling or numbering of response scales is used. Variations in ERS effect depending on question format were also present, but not in such a way that it could prevent the use of ERS. For survey practitioners, this implies that they have to content themselves with curing ERS bias after data are collected.

Every study has its limitations, and we thus faced the inevitable limitation that choices had to be made on which scales to include in our experiment. This research was part of a larger project that involved the use of four balanced sets of items measuring four different concepts. Two of these four sets were used to vary the length of response scales. The two scales presented in this research focused on the impact of labeling and numbering of scales on response behavior. The four selected scales were derived from the literature; no attempt was made to develop new balanced scales. The obvious limitation of the design was that we could not generalize our findings to other scales. We were capable only of demonstrating variations in response behavior within the selected sets of items. A minor limitation was that our study was restricted to ERS and ARS as response styles biasing measurement. Issues such as social desirability might influence the quality of the measurement as well. However, the measurement of ERS and ARS as defined in our model was unlikely to be affected by social desirability. ARS was measured as agreement with both positively and negatively worded items regarding a topic whereas social desirability would force respondents toward a particular direction on a content scale. ERS in our models contrasted respondents that tend to choose the two extreme values of the scale with respondents tending to avoid these. No clear difference in effect of ERS on the five nonextreme categories was observed. If social desirability were in play, it would be included in the content latent class factors of the current model. We had no scale to measure social desirability to check whether this was the case.

Every study raises new questions. First, the results indicated that ARS was much less prominent than expected from reading the literature. This suggests that the impact of response scale formats on ARS is

smaller than other features of survey design, such as length of interview or survey mode (e.g., Web vs. face-to-face). This does not necessarily contradict findings in previous research that indicated that ARS is stable and consistent over a four-year period (Billiet and Davidov 2008). Stability and consistency in measurement are regarded as indicating an intrinsic characteristic of the respondent. To investigate stability and consistency in measurement, however, identical survey methods must be used. Stability and consistency in ARS might then reflect consistency in the survey mode and context. We definitely need further research on this matter. Our study suggests that it might be possible to find an optimal survey design in which the occurrence of ARS is minimized even with the use of agree-disagree formats. This is especially important because the majority of attitude scales do not use balanced sets of items that exclude the possibility of filtering out acquiescent response behavior.

Second, ERS was omnipresent in this study regardless of variations in labeling and numbering that were used. Another way of looking at extreme responders (and their counterimage, extreme avoiders) is that they have higher likelihoods of undifferentiated responses, a process also known as straight-lining or ticking the same category. One avenue then might be to think of designs that encourage differentiation in responses. Rating scales such as the ones used in this research aim at estimating direction (disagree vs. agree, negative vs. positive) alongside the intensity of the attitude (levels of agreement). Measuring direction separately from intensity might reduce nondifferentiation and ERS, but it will likely be at the cost of increased respondent's burden. Such a burden, however, increases the risk for satisficing, and hence response bias might be merely redirected to other types of bias, such as nonresponse or ARS.

In the end, we think that finding ways of reducing response bias and knowing whether it is inevitable or not is particularly important in today's survey research practice, which involves the comparison of groups that might exhibit different levels of vulnerability to response style behavior. Variations in labeling and numbering did have differential effects on response bias, but not to the extent that it neutralized its negative effect on measurement. The method used allowed for correction, but—if possible—preventing the bias from occurring is preferable over curing its undesired effect.

APPENDIX

How the Latent Class Approach Untangles Genuinely Held Attitudes from Response Style Patterns

Whenever a specified model distinguishes among content and response style factors, we should be confident that the method does not conflate substantive responses with response patterns that reflect styles. In this research, ERS and ARS are modeled alongside two content factors. The following general features of the model help avoid conflation of substantive responses with style: (1) Style factors are uncorrelated with the content factors, and (2) style factors load on all items from different content-related factors. As far as ARS is concerned, an additional statistical requirement is that a positive effect sign of ARS on both positive and negative items needs to be imposed. Regarding ERS, it is required that separate effects on each answer category should be modeled. Style factors should be included only if model fit improves. Conceptually, we have argued, imposing equality constraints of the effects of ERS and ARS on all items adds to the argument that systematically responding to items independent of the content reveals a response style.

There is little reason to believe that when respondents tend to agree with both positively and negatively worded items at the same time, such a pattern would not indicate acquiescence. Results regarding ERS also indicated that respondents high on this latent class factor tend to choose the endpoints of the scale more often than respondents who are low on ERS and thus avoid the use of endpoints. This ERS factor is defined as independent from the content factors in the model and for that reason captures preferences for the endpoints of a scale independent from the content. If ERS were not present in the data, it would not show up. Footprints of ERS can be seen when inspecting the residuals in the cross-tabulation of two items, as illustrated in Table A1.

Table A1 presents the adjusted standardized residuals comparing the observed frequencies with the expected frequencies under independence. If the two variables were associated only because of the presence of a substantive underlying factor, the adjusted standardized residuals should decrease when moving away from the main diagonal. In this table, however, the residuals increase toward the corner, indicating that part of the association is the result of some respondents' preference for the endpoints of the scale.

Table A1. Two-way Frequency Table of (1a) ''Abusing Environment'' (Rows) by (1b) ''Balance of Nature'' (Columns): Adjusted Standardized Residuals

Response Category	1	2	3	4	5	6	7	Total *n*
1	1.1	–1.3	–1.2	–1.2	0.1	3.6	5.9	15
2	–2.3	–2.8	–0.8	–1.6	3.9	9.6	–0.7	59
3	–4.1	–4.8	–0.7	2.8	6.3	3.4	0.1	131
4	–6.2	–9.7	0.2	14.3	1.9	0.1	1.0	358
5	–12.0	–10.4	11.8	4.4	5.4	–0.4	–1.7	968
6	–6.3	16.7	–2.4	–6.6	–4.0	–2.3	–2.6	1,132
7	29.5	3.1	–10.2	–9.3	–7.5	–2.5	3.6	603
Total *n*	360	924	863	591	398	107	23	3,266

Table A2. Two-way Frequency Table of the ''Safe Driving'' LC Factor Classification (Modal Assignment) with and without Controlling for ERS and ARS

	''Safe Driving'' Controlling for ERS and ARS			
''Safe driving'' (uncorrected model) classes	1	2	3	Total
1	842	40	0	882
2	371	637	65	1,073
3	0	168	1,143	1,311
Total	1,213	845	1,208	3,266

The impact of including ERS on the measurement of the latent content factors is illustrated in Table A2. As is usually done in latent class analysis, we used modal assignment to classify respondents into one of the three ordered categories of the latent variables. Table A2 presents the two-way table of class assignments for the ''safe driving'' factor on the basis of an analysis with and without response style factors. Given that the latent class ERS factor estimates the probability of giving an ''avoidance of extremes'' versus a ''preference for extremes'' response, the logical consequence is that some respondents move from one level to the adjacent level of the latent content factor when ERS is taken into account. Overall, Spearman's correlation in Table A2 is high at .87. Hence, relative positions on the latent content factor change slightly when response styles are taken into account. This is what we would

expect, because ''avoiders of extremes'' do not necessarily agree more or disagree less than the ''endpoint responders.'' Depending on how systematic these response preferences occur, their relative position might change.

Acknowledgments

We gratefully acknowledge the comments and suggestions made by reviewers. In this article, use is made of data from the LISS panel of CentERdata, Tilburg.

Funding

This research was supported by a grant from the Netherlands Organization for Scientific Research (NOW), grant number 400-06-052.

Notes

1. The method also requires choosing the number of equidistant category levels of the latent factors. Using the pooled data set, we ran the basic model with two, three, four, and five equidistant categories and compared the BIC values. We found that the fit improved considerably if three instead of two equidistant levels were used. Using four and five levels led to a slightly better model fit, but computational time increased immensely over the use of three levels. Furthermore, no substantive differences in results were found if we increased the number of factor levels. Therefore, we decided on using three equidistant levels in all other analyses. Standard procedure is to define category values between 0 and 1, which in this research have been recentered across the middle category (i.e., –0.5, 0, and +0.5).
2. A table with the results discussed in this section is available on request from the corresponding author.

References

Andrews, Frank M. 1984. ''Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach.'' *Public Opinion Quarterly* 48:409–42.

Arce-Ferrer, Alvaro J. 2006. ''An Investigation into the Factors Influencing Extreme-response Style: Improving Meaning of Translated and Culturally Adapted Rating Scales.'' *Educational and Psychological Measurement* 66:374–92.

Billiet, Jaak B. and Eldad Davidov. 2008. ''Testing the Stability of an Acquiescence Style Factor behind Two Interrelated Substantive Variables in a Panel Design.'' *Sociological Methods and Research* 36:542–62.

Billiet, Jaak B. and McKee J. McClendon. 2000. ''Modelling Acquiescence in Measurement Models for Two Balanced Sets of Items.'' *Structural Equation Modelling* 7:608–28.

Chan, Jason C. 1991. ''Response-order Effect in Likert-type Scales.'' *Educational and Psychological Measurement* 51:531–40.

Coromina, Lluís and Germà Coenders. 2006. ‘‘Reliability and Validity of Egocentered Network Data Collected Via Web. A Meta-analysis of Multilevel Multitrait Multimethod Studies.’’ *Social Networks* 28:209–31.

Dickinson, Terry L. and Peter M. Zellinger. 1980. ‘‘A Comparison of the Behaviorally Anchored Rating and Mixed Standard Scale Formats.’’ *Journal of Applied Psychology* 65:147–54.

Dunlop, Riley E., Kent D. Van Liere, Angela G. Mertig, and Robert E. Jones. 2000. ‘‘Measuring Endorsement of the New Ecological Paradigm: A Revised NEP Scale.’’ *Journal of Social Issues* 56:425–42.

Greenleaf, Eric A. 1992. ‘‘Measuring Extreme Response Style.’’ *Public Opinion Quarterly* 56:328–51.

Grice, Paul. 1989. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Hartley, James and Lucy R. Betts. 2010. ‘‘Four Layouts and a Finding: The Effects of Changes in the Order of the Verbal Labels and Numerical Values on Likert-type Scales.’’ *International Journal of Social Research Methodology* 13:17–27.

Hui, C. Harry and Harry C. Triandis. 1989. ‘‘Effects of Culture and Response Format on Extreme Response Style.’’ *Journal of Cross-cultural Psychology* 20:296–309.

Hurley, John R. 1998. ‘‘Timidity as a Response Style to Psychological Questionnaires.’’ *Journal of Psychology* 132:202–10.

Johnson, Timothy, Patrick Kulesa, Young I. Cho, and Sharon Shavitt. 2005. ‘‘The Relation between Culture and Response Styles. Evidence from 19 Countries.’’ *Journal of Cross-cultural Psychology* 36:264–77.

Kieruj, Natalia D. and Guy Moors. 2010. ‘‘Variations in Response Style Behavior by Response Scale Format in Attitude Research.’’ *International Journal of Public Opinion Research* 22:320–42.

Kieruj, Natalia D. and Guy Moors. 2013. ‘‘Response Style Behavior: Question Format Dependent or Personal Style?’’ *Quality and Quantity* 47:193–211.

Krebs, Dagmar and Juergen H. P. Hoffmeyer-Zlotnik. 2010. ‘‘Positive First or Negative First? Effects of the Order of Answering Categories on Response Behavior.’’ *Methodology* 6:118–27.

Krosnick, Jon A. 1991. ‘‘Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys.’’ *Applied Cognitive Psychology* 5:213–36.

Krosnick, Jon A. and Matthew K. Berent. 1993. ‘‘Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format.’’ *American Journal of Political Science* 37:941–64.

Krosnick, Jon A. and Leandre R. Fabrigar. 1997. ‘‘Designing Rating Scales for Effective Measurement in Surveys.’’ Pp. 141–64 in *Survey Measurement and Process Quality*, edited by Lars E. Lyberg, Paul Biemer, Martin Collins, Edith D. de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. New York: John Wiley.

Lam, Tony C. M. and Mary Kolic. 2008. ‘‘Effects of Semantic Incompatibility on Rating Response.’’ *Applied Psychological Measurement* 32:248–60.

Lau, Michael Y. 2007. ‘‘Extreme Response Style: An Empirical Investigation of the Effects of Scale Response Format Fatigue.’’ PhD dissertation, University of Notre Dame, Notre Dame, IN.

Moors, Guy. 2003. ‘‘Diagnosing Response Style Behaviour by Means of a Latent Class Factor Approach. Socio-demographic Correlates of Gender Role Attitudes and Perceptions of Ethnic Discrimination Re-examined.’’ *Quality and Quantity* 37:277–302.

Morren, Meike, Jeroen Vermunt, and John Gelissen. 2011. ‘‘Dealing with Extreme Response Style in Cross-cultural Research: A Restricted Latent Class Factor Analysis Approach.’’ Pp. 13–47 in *Sociological Methodology*, Vol. 41, edited by Tim Futing Liao. Hoboken, NJ: Wiley-Blackwell.

Paulhus, Del L. 1991. ‘‘Measurement and Control of Response Bias.’’ Pp. 17–59 in *Measures of Personality and Social Psychological Attitudes*, edited by John Paul Robinson, Philip R. Shaver, and Lawrence S. Wright. San Diego, CA: Academic Press.

Peters, David L. and Ernest J. McCormick. 1966. ‘‘Comparative Reliability of Numerically Anchored Versus Job-task Anchored Rating Scales.’’ *Journal of Applied Psychology* 50:92–96.

Preston, Carolyn C. and Andrew M. Colman. 2000. ‘‘Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences.’’ *Acta Psychologica* 104:1–15.

Rammstedt, Beatrice and Dagmar Krebs. 2007. ‘‘Does Response Scale Format Affect the Answering of Personality Scales? Assessing the Big Five Dimensions of Personality with Different Response Scales in a Dependent Sample.’’ *European Journal of Psychological Assessment* 23:32–38.

Richardson, Hettie A., Marcia J. Simmering, and Michael C. Sturman. 2009. ‘‘A Tale of Three Perspectives: Examining Post-hoc Statistical Techniques for Detection and Correction of Common Method Variance.’’ *Organisational Research Methods* 12:762–800.

Schwarz, Norbert. 1999. ‘‘How the Questions Shape the Answers.’’ *American Psychologist* 54:93–105.

Schwarz, Norbert and Hans J. Hippler. 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, Norbert, Bärbel Knäuper, Hans J. Hippler, Elisabeth Noelle-Neumann, and Leslie Clark. 1991. ‘‘Rating Scales. Numeric Values May Change the Meaning of Scale Labels.’’ *Public Opinion Quarterly* 55:570–82.

Simon, Herbert A. 1955. ‘‘A Behavioral Model of Rational Choice.’’ *Quarterly Journal of Economics* 59:99–118.

Smith, David H. 1967. ‘‘Correcting for Social Desirability Response Sets in Opinion-attitude Survey Research.’’ *Public Opinion Quarterly* 31:87–94.

Symonds, Percival M. 1924. ‘‘On the Loss of Reliability in Rating Scales Due to Coarseness of the Scale.’’ *Journal of Experimental Psychology* 7:456–61.

Tourangeau, Roger, Mick P. Couper, and Frederick Conrad. 2007. ‘‘Color, Labels, and Interpretive Heuristics for Response Scales.’’ *Public Opinion Quarterly* 71:91–112.

Van Herk, Hester, Ype H. Poortinga, and Theo M. M. Verhallen. 2004. ‘‘Response Styles in Rating Scales. Evidence of Method Bias in Data from Six Countries.’’ *Journal of Cross-cultural Psychology* 35:346–60.

Vermunt, Jeroen and Jay Magidson. 2005. *Latent GOLD 4.0 User's Guide*. Belmont, MA: Statistical Innovations.

Weijters, Bert, Elke Cabooter, and Niels Schillewaert. 2010. ''The Effect of Rating Scale Format on Response Styles: The Number of Response Categories and Response Category Labels.'' *International Journal of Research Marketing* 27:236–47.

Weng, Li-Jen. 2004. ''Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-retest Reliability.'' *Educational and Psychological Measurement* 64:956–72.

Yilmaz, Veysel and H. Eray Çelik. 2006. ''Risky Driving Attitudes and Self-reported Traffic Violations among Turkish Drivers: The Case of Eskişehir.'' *Doğuş Üniversitesi Dergisi* 7:127–38.

Author Biographies

Guy Moors is an assistant professor in the Department of Methodology and Statistics at Tilburg University in the Netherlands. His research is on survey methodology, cross-cultural comparative research, social demography, and applied latent variables modeling. He has published work on these topics in methodological, sociological, demographic, and public opinion journals as well as in books.

Natalia D. Kieruj is a researcher in the survey research department at CentERdata, an independent research institute located at Tilburg University. In 2012, she received her PhD on a study titled ''Question Format and Response Style Behavior in Attitude Research.'' In her research, she applies latent class confirmatory factor analysis to correct for response biases in attitude research. Currently, she carries out Web-survey research in both the CentERpanel and the LISS panel at CentERdata.

Jeroen K. Vermunt is a professor in the Department of Methodology and Statistics at Tilburg University. His research is on methodologies of social, behavioral, and biomedical research, with a special focus on latent variable models and techniques for the analysis of categorical, multilevel, and longitudinal data sets. He has widely published on these topics in statistical and methodological journals and has also coauthored many articles in applied journals in which these methods are used to solve practical research problems. He is the codeveloper (with Jay Magidson) of the Latent GOLD software package. In 2005, he was awarded the Leo Goodman Award by the Methodology Section of the American Sociological Association.