THE EFFECT OF LABELING AND NUMBERING OF RESPONSE SCALES ON THE LIKELIHOOD OF RESPONSE BIAS (°)

Guy Moors

Tilburg University, FSW-MTO, Room P.1.110. P.O. Box 90153, 5000 LE, Tilburg, The Netherlands. E-mail: guy.moors@uvt.nl.

Natalia D. Kieruj

CentERdata, Room T428. P.O. Box 90153, 5000 LE, Tilburg, The Netherlands. E-mail: n.d.kieruj@uvt.nl.

Jeroen K. Vermunt

Tilburg University, FSW-MTO, Room P.1.114. P.O. Box 90153, 5000 LE, Tilburg, The Netherlands. E-mail: j.k.vermunt@uvt.nl.

In this paper use is made of data from the LISS panel of CentERdata.

(°) accepted for publication in Sociological Methodology

ABSTRACT

Extreme response style (ERS) and acquiescence response style (ARS) are among the most encountered problems in attitudinal research. We investigate whether response bias caused by these response styles vary with three aspects of question format, namely full versus end labeling, numbering answering categories and bipolar versus agreement response scales. A questionnaire was distributed to a random sample of 5351 respondents from the LISS household panel which was assigned to one of five treatments with differing scale formats. We apply a latent class factor model that allows for diagnosing and correcting for ERS and ARS simultaneously.

Results show clearly that both response styles are present in our dataset, but ARS is less pronounced than ERS. With regard to format effects, it is found that end labeling evokes more ERS than full labeling, and that bipolar scales evoke more ERS than 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.

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 non-random error and Likert type rating scales have shown to be prone to all kinds of biases (Chan 1991; Greenleaf 1992; Kieruj and Moors 2010; Smith 1967). In this paper, we take interest in the question if certain aspects of scale format, more specifically the verbal and numerical labeling of the answering 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, namely extreme response style (ERS) and 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).

The process of constructing a rating scale is not as straightforward as it may seem. There are several choices a researcher has to make when designing a rating scale. Deciding on the number of answering 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, like numbering and labeling of answering categories. A common distinction that is made when it comes to labeling is that of 'full labeling' and 'end labeling'. In the former case, all answering categories are verbally labeled (e.g. a 5-point scale would consist of the labels 'completely disagree', 'disagree', 'do not disagree or agree', 'agree' and 'completely agree') whereas, in the latter case only the end categories are labeled (e.g. completely disagree and completely agree). We are interested in the question if the use of end labeling rather than full labeling evokes the use of ERS and ARS. Also, the issue of bipolar versus agreement scales and its influence on response behavior is a topic of interest. These scales differ in their numbering of response categories with the former presenting both negative and positive values, whereas the latter only presents positive values. Finally, it seems to be common practice to attach numbers to response categories alongside the category labels. The question asked regarding to this topic is whether presenting respondents with extra anchors in the form of numbers, will yield different amounts of ERS and ARS.

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

Literature review: the effect of scale format on data quality

In this research, we will 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. The latter 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 other latent variables than the intended. Deciding on whether and how to label and/or number the response scale is a task of every survey research practitioner. Hence whether these choices made have consequences regarding response bias are of scientific as well as societal relevance. The first topic deals with full versus end labeling of scales and the second topic revolves around the issue of numerical values and whether or not to use them to accompany the answering categories. In addition, the third topic deals with the comparison of agreement versus bipolar response scales. In the following overview, each of these topics is discussed from findings and perspectives from the literature. What unifies these studies across topics are the following complementary theoretical propositions:

(a) Survey question formats may increase response burden depending on how cognitively demanding they are. Original coined by Simon (1955) the concept of 'satisficing' has been used by Krosnick (1991) – among others – to indicate that when response burden increases a respondent is more likely to satisfice rather than to optimize his responses. By consequence response bias will increase.

(b) In line with the principal of nonredundancy (Grice, 1989) it is expected that respondents tend to look for cues on how to respond to survey questions in their attempt to give adequate answers. As such they tend to assign meaning to all incentives given in the question format. Similar to the

satisficing principal it is expected that the less demanding the 'cue-looking' task is, the less vulnerable a scale format is to response bias.

Full versus end labeling

A considerable amount of studies has been devoted to the issue of labeling all or just the endpoints of a rating scale. In favor of full labeling, it has been argued that they provide more information to respondents about how to interpret the scale (Johnson, Kulesa, Cho and Shavitt 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 since the former is more precise and easier to hold in memory. At the same time, it is argued 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 instead of end labeling was used.

There have also been a limited amount of studies that focused on the effect of end 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 is explained by pointing out that in the case of a fully labeled scale, the center

categories become more salient to respondents than when only the end categories are labeled. A study by Lau (2007) on the contrary, showed no significant effect of end versus full labeling on ERS.

Using numerical values to accompany answering 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 due to the fact that it is difficult to imagine how the 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. Since results in these studies were at least partially dependent on the use of numerical values, the issue whether or not 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 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 like giving the endpoints of the scale differing colors. This effect was entirely eliminated if labels for all categories were used (even if they were just 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 like response style behavior to arrive at satisfying answers.

Bipolar versus agreement 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, Knäuper, Hippler, Noelle-Neumann and Clark (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.

Developing the research question

In this study, we focus on ERS – the tendency to choose the end-points of a scale – and ARS – the tendency to agree with questions – and attempt to establish if 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, since 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, i.e. if only the end points are labeled ánd 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 to agreement scales that only use positive integers.

Data, design and method

Participants

Our split-ballot experiment was implemented in the LISS web panel of CentERdata, which is a Dutch household panel consisting of 8044 participants and was initiated in 2008 (http://www.lissdata.nl/lissdata/Home). We like to underscore that the quality of the sampling strategy matches with high standards set in regular face-to-face surveys. Different from voluntary internet panels, this household panel includes households that were recruited using a random sampling design. Participants who did not have a personal computer and/or internet access received this facility so that these participants were not automatically excluded from participation in the panel.

Our attitudinal scales were fielded in February 2009 and filled out by 5351 respondents leading to a response rate of 65% (AAPOR RR6). The sample was 46.1% male and 53.9% female. Ages ranged from 16 to 95 years of age with a mean age of 47. The purpose of a split-ballot experiment is to achieve that experimental groups only differ in treatment. Although unlikely, we checked whether differential non response 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 of satisficing response behavior. Satisficing increases with length of the questionnaire (response burden) and when respondents are less familiar with the survey context. LISS-respondents have been familiarized with answering survey questions on several occasions prior to answering our set of questions. Furthermore, only short questionnaires are used in this web

survey and by consequence fatigue or loss of interest are less likely to occur than with long questionnaires.

Questionnaire

In this research, we need to use of balanced sets of items. The minimum requirement to measure ARS is that at least one scale is partially balanced (Billiet and McClendon 2000) since 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 towards environmental issues (α's ranging from .707 to .762) and attitudes towards risky driving (α's ranging from .740 to .766) (Appendix A). The items from the environment scale were adopted from a revised NEP scale by Dunlop, van Liere, Mertig and Jones (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, since our study varies in the way these labels are used we are 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 since virtually all respondents chose the sixth or seventh answering category. Our scales were positioned at the very 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 during one month. Three reminders were sent during this period of time.

Design

In the setup of this study, respondents were randomly assigned to five treatments that varied in the use of labeling and numbering of response scales to the same set of questions with each 7 ordered answering categories in the following way:

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; and

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 only labeled ‘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 is 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 0-value most clearly identifies 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 on 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, i.e. 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.

Method

We employ 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 parsimoniousness since all effects of the latent variables on response variables were defined as non-monotone resulting in *C-1* parameters per response variable with *C* being 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 non-monotone relationship in the case of ERS. In this research we further extend 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 is 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:

$$\log \frac{P(Y_{ij} = c+1 \mid F1_i, F2_i, ERS_i, ARS_i)}{P(Y_{ij} = c \mid F1_i, F2_i, ERS_i, ARS_i)} =$$

$$(\beta_{0jc+1} - \beta_{0jc}) + \beta_{1j} F1_i + \beta_{2j} F2_i + (\beta_{3c+1} - \beta_{3c}) ERS_i + \beta_4 ARS_i$$

in which Y_{ij} denotes the response of individual *i* to rating item *j*; F1 and F2 the latent content factors; and ERS and ARS 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 define the monotone relationship between F1, F2, ARS and Y. The term $(\beta_{3c+1} - \beta_{3c})$ defines the non-monotone relationship of ERS with Y and implies the estimates of $C - 1$ β-parameters, with C being the number of response categories.

In this research, the latent class content factors refer to the two 'environment' and 'risky driving' attitude scales, and items are only allowed to load on their corresponding attitudinal factor. In the case of the ERS and ARS, all items load on these style factors since all items are supposed to be affected by response bias. Content factors were allowed to correlate among each other, but style factors not. This way, we are 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 preferences for certain response 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, a reader might be concerned that our model conflates substantive responses with response styles. Our model resembles the 'unmeasured latent method construct' approach of which Richardson, Simmering and Sturman (2009) advise against using it since it only works when one is 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 one aims to measure. To avoid this, it should be taken into account that a latent response style factor can only be interpreted as a style factor if the response

pattern is not consistent with the content that is measured (Billiet and McClendon 2000). Hence, ARS can only be unequivocally diagnosed if respondents tend to agree with both negatively and positively worded items measuring the same concept. This is achieved in this research by imposing 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: (a) ERS is uncorrelated with the content factors; (b) ERS is measured as a single LC factor influencing responses to sets of items that differ in substantive meaning; (c) 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 (d) including ERS decreases the distance between extreme responders and endpoint avoiders without necessarily changing their relative position on the content dimensions. Additional evidence that the applied strategy does not conflate substantive responses with response styles is presented in appendix B. In this appendix we demonstrate that relative positions on the content factors slightly change when ERS is taken into account making the relative distance between ‘avoiders of extremes’ versus ‘endpoint responders’ somewhat more narrow without completely vanishing it. Furthermore, in a 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 procedure (1992) to define a contentless measure of ERS. Correlations ranged from .371 to .493, which is fairly high since these questions were administered at other waves in the LISS panel.Weijters, Cabooter, and Schillewaert (2010) adopted Greenleaf’s procedure (1992) of defining a contentless measure of ERS by counting extreme scores and calculated a similar index to measure ARS. The problem with these kind of indices is that they are deterministic and not model based. The benefits of our

model based approach are that: (a) model fit comparisons allow to research whether including response style factors improve model fit, hence evaluating whether they did affect the measurement of substantive scale; and (b) that it partitions the responses on items into a part affected by content (true score) and a part affected by style (response bias).

Model specifications

In the previous section, we elaborated on the method used in this research by defining the basic 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 is based on log-likelihood (LL) estimates and information criteria. In this research, we make use of BIC (Bayesian Information Criterion) which simultaneously estimates the fit of the model alongside its parsimoniousness (number of parameters relative to the other models it is compared to) and partly compensates for sample size. The lower the BIC value the better the balance between fit (=LL) and complexity (=Npar).

The basic model refers to a single sample, whereas our split-ballot involves five samples parallel to the five test conditions. As a way of screening the data, we have first run 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 made no difference which of the five different 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 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 Table1) 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 to model 1.1 and Table 1 shows that adding the content factors is a major improvement in terms of BIC and ΔLL. Second, the reference model is compared to 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, it is shown that in model 1.4 BIC and ΔLL improve even more if both style factors are included in the model. Results presented in table 1 make use of the pooled dataset, 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. The first conclusion we can draw from the latter finding is that none of the tested response scale formats are immune from response biases.

Insert table 1 about here

Having selected a starting model in the first step, the next question was whether we could 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 2, we compare a 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 only applied to the style factors. Model 2.3) includes no such equality constraints. Results show that a model with equality restrictions on the style factors is the most appropriate model according to 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 since 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, one could argue 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 to the model assuming equal effects for ERS and ARS (model 2.3 in Table 2).

Insert table 2 about here

Whether the effect of ERS and ARS is different depending on response scale format is tested in the following step, in which we adopt a multiple group comparison approach using the pooled dataset. Pooling the data of the split ballot experiment was feasible since all respondents did answer the same questions on a 7-point scale, only the labeling and numbering of the categories differed across groups. In the pooled dataset, 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, it is tested 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). 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.

Insert table 3 around here

As it can be seen in Table 3, we started with the most complex model 3.1 that included the direct effects of 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 to models in which particular effects are omitted. The final model 3.6 defines a model in which no effect of the group variable is included, suggesting a fully homogeneous measurement model with no impact of response scale format whatsoever.

The complex model 3.1 has a considerably higher BIC value than the simpler model 3.6. We estimated several models that are in-between the heterogeneous model 3.1 and the

homogeneous model 3.6. Starting from model 3.1 we omitted the direct effect of the group variable on items (model 3.2). The model improved over model 3.1 in terms of BIC but was still less appropriate than model 3.6. Model 3.3 excluded the interaction terms of the content factors which all proved to be non-significant in the previous model (F1*group and F2*group, $p > .1$). The model further improved with BIC values lower than the first as well as the last model. Note that model 3.3 directly relates to the research questions asked since it checks whether the effect of ERS and ARS on the items is different depending on the test conditions. At the same time, the lower BIC value of model 3.3 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 having a closer look at the estimates of model 3.3, we could further simplify the model by dropping the ARS*group interaction, which was not significant. This is confirmed in model 3.4. The ERS*group interaction, on the contrary, could not be dropped since then model fit deteriorated; as it can be seen in comparison of models 3.4 with 3.5. Hence, the most appropriate model in terms of 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 item responses[1]. The interpretation of the effect parameters in this model is subject to the next section.

[1] The method also requires choosing the number of equidistant category levels of the latent factors. Using the pooled dataset, we ran the basic model with 2, 3, 4 and 5 equidistant categories and compared the BIC values. We found that the fit improved considerably if 3 instead of 2 equidistant levels were used. Using 4 and 5 levels lead to a slightly better model fit, but computational time increased immensely over the use of 3 levels. Furthermore, no substantive differences in results were found with increasing the number of factor levels. Therefore we decided on using 3 equidistant levels in all other analyses. Standard procedure is to define

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.

The effect of ERS and ARS on item responses.

Table 4 shows the logit effect (beta's) of ERS and ARS on the response items (from the final model 3.4) which were both significant ($p < 0.001$). 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 since the method as such only allows revealing response scale point preferences among respondents independent of the content of items.

Insert table 4 about here

Table 4 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) ($p < 0.001$). However, as previously reported (table1), model fit did substantially increase by including ERS but only marginally by adding ARS. Furthermore, by transforming the logit (beta)

---

category values between 0 and 1, which in this research have been re-centered across the middle category, i.e. -0.5, 0 and +0.5.

parameters to its odds ratios one 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 only holds to the items asked in this particular research. ERS, on the other hand, is prominently present.

The effect of scale format (i.e. numbering and labeling) on response styles.

In the final model 3.4, the nominal group variable – indicating the five test conditions –only showed a significant effect 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 since this defined the test conditions. As it can be seen in Table 5, 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.

Test-specific ERS effects on item responses.

The final 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*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 0.

Insert figure 1 about here

The overall effect is strongly present in all treatments ($p < 0.001$), but there is some group specific deviations as well (ERS*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.

Discussion

We set out to investigate if 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 as ERS even when fully labeled agree-disagree scales were used. The latter might come as a surprise since 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 other factors than question format evoke ARS. Long exhaustive questionnaires in face to face interviews, for instance, might induce ARS to a greater extent. Secondly, we should equally acknowledge that finding ARS in a balanced set of items by definition implies non-consistent 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 whereas it should have been 'disagreeing' – 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 differ across treatments, to some extent. 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 and 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 the respondent so 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 by using positive values only (with '0' indicating a truly bad score) rather than giving negative values. As such, the 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 versus any other preference, whereas with end labeling extreme responding is opposed by mid scale 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 the present 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 like the fact that we were not able to draw firm conclusions regarding ARS since it was less prominently present in this research than reported by other researchers using similar questionnaires. Nevertheless, this research found evidence that ERS influences the responses to attitudinal questions regardless which type of labeling or numbering of response scales is used. Question format specific ERS effects are 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 is collected.

Every study has its limitations. An inevitable limitation is that choices were 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 literature; no attempt was made to develop new balanced scales. The obvious limitation of the design is that we cannot generalize our findings to other scales. We were only capable of demonstrating variations in response behavior within the selected sets of items. A minor limitation is 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 is unlikely to be affected by social desirability. ARS is measured as agreement with both positively and negatively worded items regarding a topic whereas social desirability would force respondents towards a particular direction on a content scale. ERS in our models contrasts 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 non-extreme categories was observed. If social desirability was in play it would be included in the content latent class factors of the current model. We have 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 present 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 versus face-to-face). This does not necessarily contradict findings in previous research that indicated that ARS is stable and consistent over a 4-year period of time (Billiet and Davidov, 2008). Stability and consistency in measurement is regarded as indicating an intrinsic characteristic of the respondent. To investigate stability and consistency in measurement, however, it is required that identical survey methods are used. Stability and consistency in ARS might then reflect consistency in the survey mode and context. We definitely need further research on this matter. This 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 since the majority of attitude scales do not use balanced sets of items which exclude the possibility of filtering out acquiescent response behavior. Second, ERS was omnipresent in this study regardless of variation in labeling and

numbering that is used. Another way of looking at extreme responders (and their counterimage extremes avoiders) is that have higher likelihoods of undifferentiated responses. As with ARS we need additional research on whether survey modes impact response styles. A second avenue might be to think of designs that encourage differentiation in responses. Rating scales like the ones used in this research aim at estimating direction (disagree versus agree; negative versus positive) alongside the intensity of the attitude (levels of agreement). Disentangling might reduce non-differentiation and ERS but likely at the cost of increased respondent's burden.

In the end we think that finding ways of reducing response bias and knowing whether it is inevitable or not is highly important in today's survey research practice that involves the comparisons 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 to correct for it, but – if possible – preventing the bias to occur is preferable over curing its undesired effect.

Funding

This work was supported by a grant from the Dutch Science Foundation [grant number 400-06-052].

REFERENCES

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

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* 66:374-392.

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* 36:542-562.

Billiet, Jaak B., and McKee J. McClendon. 2000. “Modelling acquiescence in measurement models for Two balanced sets of items.” *Structural Equation Modelling* 7:608-628.

Chan, Jason C. 1991. “Response-order effect in Likert-type scales.” *Educational and Psychological Measurement* 51:531-540.

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* 28:209-231.

Dickinson, Terry L., and Peter M. Zellinger. 1980. “A comparison of the behaviorally anchored rating and mixed standard scale formats.” *Journal of Applied Psychology* 65:147-154.

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* 56:425-442.

Greenleaf, Eric A. 1992. “Measuring extreme response style.” *Public Opinion Quarterly* 56:328-351.

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.

Hurley, John R. 1998. “Timidity as a response style to psychological questionnaires.” *The Journal of Psychology* 132:202-210.

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* 36:264-277.

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* 22:320-342.

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* 6:118-127.

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* 37:941-964.

Krosnick, Jon A., and Leandre R. Fabrigar. 1997. “Designing rating scales for effective measurement in surveys.” In: *Survey Measurement and Process Quality*, eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 141- 164. New York: Wiley.

Lam, Tony C. M., and Mary Kolic. 2008. “Effects of semantic incompatibility on rating response.” *Applied Psychological measurement* 32:248-260.

Lau, Michael Y. 2007. “Extreme response style: An empirical investigation of the effects of scale response format and fatigue.” [dissertation].

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 & 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.” *Sociological Methodology*

Paulhus, Del L. 1991. “Measurement and control of response bias.” In: *Measures of personality and social psychological attitudes*, eds. J. P. Robinson, P. R. Shaver and L. S. Wright, 17-59. San Diego: 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, Carla E. Grayson, and Bärbel Knäuper. 1998. “Formal features of rating scales and the interpretation of question meaning.” *International Journal of Public Opinion Research* 10:177-183.

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* 55:570-582.

Simon, Herbert A. 1956. “Rational choice and the structure of the environment.” *Psychological Review* 63:129-138.

Smith, David H. 1967. “Correcting for social desirability response sets in opinion-attitude survey research.” *The 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* 7:456-461.

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 styls in rating scales. Evidence of method bias in data from sic countries.” *Journal of Cross-cultural Psychology* 35:346-360.

Vermunt, Jeroen, and Jay Magidson. 2005. “LATENT GOLD 4.0 user’s guide”. Belmont, Massachusetts: Statistical Innovations Inc.

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* 27:236-247.

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* 64:956-972.

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* 7:127-138.

Table 1. Model fit comparisons

Model #	LC Factors included	Npar	LL	Δ LL	BIC
1.1	No (°)	42	-38675		77690
1.2	Content	54	-35322	3353	71082
1.3.1	Content + ARS	57	-35235	3441	70930
1.3.2	Content + ERS	62	-34010	4665	68522
1.4	Content + ERS + ARS	65	-33962	4713	68449

Results are from the pooled dataset

(°) Reference *LL* value for *ΔLL* comparisons

Table 2. BIC values of models with varying equality restrictions

Model #	Equality restrictions	Npar	LL	BIC(LL)
2.1	No restrictions	72	-34603	69789
2.2	Restrictions on style factors	65	-33962	68449
2.3	Restrictions on all factors	60	-36572	73629

Results are from the pooled dataset

Table 3. The effect of test conditions (group variable) on the measurement of LC factors.

Model #		Npar	LL	BIC(LL)
Model 3.1	F1 + F2 + ERS + ARS + group + F1*group + F2*group + ERS*group + ARS*group	165	-33675	68685
Model 3.2	F1 + F2 + ERS + ARS + F1*group + F2*group + ERS*group + ARS*group	137	-33701	68511
Model 3.3	F1 + F2 + ERS + ARS + ERS*group + ARS*group	109	-33731	68345
Model 3.4	F1 + F2 + ERS + ARS + ERS*group	105	-33734	68318
Model 3.5	F1 + F2 + ERS + ARS	81	-33878	68412

note: structural part of all models includes group effects on all LC factors

Table 4. Effect of ERS and ARS on the response items (logit coefficients)

Response style		Beta	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

rc = response category

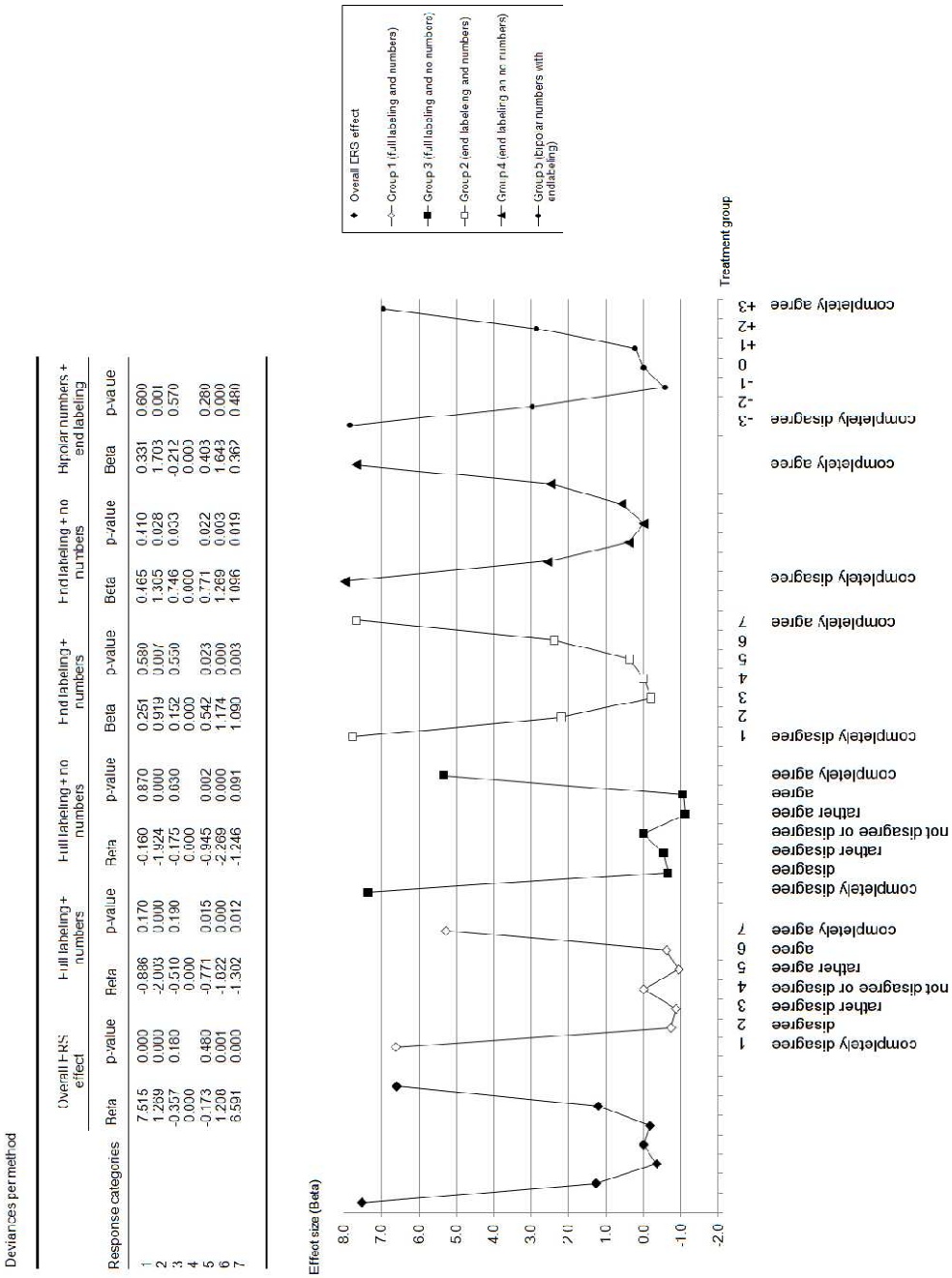
note: equal effect parameters on all items

Table 5. Group (question format) effects on the latent class ERS factor (logit coefficients).

Treatment	Beta	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

Figure caption

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



Appendix A: Item wording (translated from Dutch)

1a) Humans are severely abusing the environment (-).

1b) The balance of nature is strong enough to cope with the impacts of modern industry (+).

1c) The so-called ‘ecological crisis’ facing humankind has been greatly exaggerated (+).

1d) The balance of nature is very delicate and easily upset (-).

2a) Safe drivers can exceed the speed limits (+).

2b) There is no problem to drive above the speed limits if the conditions are proper (+).

2c) Even if you have good driving skills, this does not mean that speeding is OK (-).

2d) It is always risky to drive after drinking alcohol (-) (removed from the scale).

Appendix B. How the latent class approach untangles genuinely held attitudes from response style patterns.

Whenever a model is defined that distinguishes among content and response style factors one should be confident that the method does not conflate substantive responses with response patterns reflecting styles. In this research ERS and ARS are modeled alongside two content factors. The following general features of the model contribute to avoiding conflation of substantive responses with style: (a) style factors are uncorrelated with the content factors; and (b) 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 positively and negatively items needs to be imposed. Regarding ERS, it is required that separate effects on each answer category should be modeled. Style factors should only be included 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, that 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 was 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 is illustrated in the following table B1.

Insert table B1 here

Table B1 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 if one moves away from the main diagonal. In this table however, the residuals increase towards the corner, indicating that part of the association is the result of some respondents' preference for the endpoints of the scale.

The impact of including ERS on the measurement of the latent content factors is illustrated in table B2.

Insert table B2 here

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 B2 presents the two-way table of class assignments for the 'save driving' factor based on 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, the spearman correlation in the table B2 is at a high 0.87 level. Hence, relative positions on the latent content factor slightly change when response styles are taken into account. This is what one would expect since '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.

Table B1. Two-way frequency table of (1a) ‘abusing environment’ by (1b) ‘balance of nature’ (adjusted standardized residuals).

Response categories	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	1132
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	3266

Table B2. Two-way frequency table of the ‘Save driving’ LC factor classification (modal assignment) with and without controlling for ERS and ARS.

	‘Save driving’ controlling for ERS and ARS			
‘Save driving’ (uncorrected model)				
Classes	1	2	3	Total
1	842	40	0	882
2	371	637	65	1073
3	0	168	1143	1311
Total	1213	845	1208	3266