The Calibrated Sigma Method:

An Efficient Remedy for Between-Group Differences in

Response Category Use on Likert Scales

Accepted for publication in International Journal of Research in Marketing

Bert Weijters, Hans Baumgartner, Maggie Geuens *

Bert Weijters is Assistant Professor, Department of Personnel Management, Work and Organizational Psychology, Ghent University, B-9000 Ghent, Belgium, T.: +32 9 264 62 96, Fax: +32 9 264 64 94, E-mail: bert.weijters@ugent.be.

Hans Baumgartner is the Smeal Professor of Marketing in the Smeal College of Business at The Pennsylvania State University, Department of Marketing, 482 Business Building, University Park, PA 16802, T. 814 863 3559, E-mail: hansbaumgartner@psu.edu.

Maggie Geuens is Professor of Marketing at Ghent University and Vlerick Business School, Tweekerkenstraat 2, B-9000 Ghent, Belgium, T. + 32 9 264 35 21 - Fax + 32 9 264 42 79, E-mail: Maggie.geuens@UGent.be.

Abstract

The authors propose a procedure, labeled the calibrated sigma method, which is designed to correct for between-group differences in endorsement likelihood of response categories that are unrelated to the content of the items. The method is especially useful in cross-cultural research where group differences may reflect variation in scale usage rather than substantive differences. However, the procedure is also relevant in other situations, for example, when different data collection modes or different experimental manipulations affect respondents' perception of the meaning of the scale labels. The calibrated sigma method uses information derived from heterogeneous control items (calibration items) to reweight the responses to substantive items in a group-specific way. The advantages of the calibrated sigma method are that it avoids the arbitrariness in the assignment of particular numerical values to response categories; that it is compatible with the linear model, which is used by most marketing researchers; and that it does not require the use of complex nonlinear models involving the estimation of many additional measurement model parameters. The authors validate the calibrated sigma method on a simulated cross-linguistic data set pertaining to 12 different languages; an empirical data set collected from respondents of the same nationality but from two different language groups; and an experimental data set consisting of responses to two different response scale formats. The findings demonstrate that the proposed procedure controls for artefactual scale use differences across groups but does not eliminate substantive differences. It is particularly efficient for marketing research agencies, panel providers and other marketing researchers who analyze surveys involving multiple language groups, different scale formats, multiple modes of data collection, or different manipulations affecting the meaning of the response category labels.

Key words: Response bias, language differences, survey methods, Likert items

1. Introduction

When researchers want to compare scores on variables of interest across groups or conditions, scale usage heterogeneity is an important source of concern. The term scale usage heterogeneity (also called differential scale usage) refers to systematic differences in how respondents in different groups use the response scale, which are unrelated to substantive differences on the variables studied. Scale usage heterogeneity is problematic because it may lead to artificial differences between groups or mask true differences.

Scale usage differences are often conceptualized as individual differences that should be assessed and controlled at the respondent level (Baumgartner & Steenkamp, 2001; Fischer, 2004; Rossi, Gilula, & Allenby, 2001). However, in certain situations scale usage heterogeneity may occur primarily at the group level, in which case it is more appropriate to model differential scale usage at the group level. For example, when Likert-type rating scales anchored by labels such as ‘strongly (dis)agree’ or ‘completely (dis)agree’ are used in different languages, the meaning of the response category labels may subtly but systematically vary across languages, which can lead to differences in scale usage at the group level (Skevington & Tucker, 1999; Smith, Mohler, Harkness, & Onodera, 2005; Szabo, Orley, & Saxena, 1997; Weijters, Geuens, & Baumgartner, 2013). Similarly, data collection modes or experimental manipulations may affect the perceived meaning of the category labels and thus induce scale usage heterogeneity (Jordan, Marcus, & Reeder, 1980; Weijters, Schillewaert, & Geuens, 2008). In these cases, different response distributions across groups are not due to item content, but occur because of the non-equivalence of response category meanings.

In an attempt to remedy this potential bias, we introduce a procedure labeled the calibrated sigma method, which is designed to eliminate the non-comparability of responses across groups (e.g., cultures, languages, modes of data collection, experimental conditions) at the group, rather than individual, level. Instead of assigning the same consecutive integers to the scale positions in all groups (e.g., in the case of a 5-point scale, 'strongly disagree' is usually coded as 1, 'disagree' as 2, 'neither agree nor disagree' as 3, 'agree' as 4, and 'strongly agree' as 5), the response categories are converted to numerical values in a group-specific way. Specifically, the numbers assigned to the response categories are based on the distribution of responses to an independent and heterogeneous set of control items, which serve no purpose other than assessing the content-free endorsement frequencies of the response categories in different groups (i.e., these calibration items are not used for substantive purposes). Thus, instead of arbitrarily assuming an equal-interval scale, the scale scores are chosen based on how the different groups respond to a set of content-free items, or at least items that share no obvious common content. For instance, 'strongly agree' might be coded as 5 in English, whereas 'tout à fait d'accord' is coded as 4.5 in French, corresponding to the different endorsement rates of the fifth option in response to the control items across the two languages.

After presenting an overview of previous approaches to dealing with scale usage differences at the individual level and a detailed description of the proposed procedure, we present three complementary studies in the current paper. In the first study, we use a simulated data set to illustrate how the proposed calibrated sigma method works, based on a comparison of traditionally coded and sigma coded responses simulated for twelve different languages, and we show how the new procedure can yield more valid results than the conventional procedure. Specifically, in contrast to the traditional procedure, the new procedure does not

indicate artificial group differences in case there are none while it does not wash out genuine differences. This study also demonstrates that testing for measurement invariance across groups will not identify scale usage differences when the bias is uniform across items. In the second study, we apply the calibrated sigma method to an empirical data set of respondents who share the same nationality (Belgian) but use different languages (Dutch and French), and we demonstrate that the new procedure leads to conclusions that differ from the conventional method but are consistent with the results of an analysis that corrects for response styles at the individual level. In particular, while the conventional method suggests that there might be a significant difference in the construct of interest between Dutch- and French-speaking respondents, the calibrated sigma method and the individual-level response style correction method both indicate that this difference is most likely caused by scale usage differences. In the third study, we illustrate the potential use of the proposed method in an experimental context in which survey responses are obtained with two alternatively labeled response scale formats to which respondents are randomly assigned. We demonstrate that calibrated sigma coding outperforms traditional coding and leads to results that are comparable to those obtained with more elaborate and involved individual-level response style correction methods.

2. Literature review

It is well-known in the survey literature that observed scores on variables of interest contain not only substantive but also non-content-related sources of variation. The term common method bias is often used to refer to the general problem of non-random variance in measures that is independent of content (Podsakoff, MacKenzie, & Podsakoff, 2012). In this paper we are specifically concerned with systematic differences in how respondents use the

response scale (scale usage heterogeneity). Usually, differential scale usage is conceptualized as a respondent-specific phenomenon, such that different respondents vary in their preference for certain scale positions. Two broad approaches to controlling for individual-level scale usage heterogeneity can be distinguished.

In the first approach, the items measuring the substantive constructs of interest are used to assess and correct for differences in scale usage, and the sources of differential scale usage are not identified in detail. A popular method exemplifying this approach is to standardize (or at least mean-center) the data within respondents. That is, a person's responses to the substantive items are converted into z-scores by subtracting from each response the respondent's mean response across all items and dividing by the standard deviation of the respondent's ratings (Fischer, 2004). This method acknowledges that there may be systematic differences in the level and spread of people's responses across items, but otherwise the sources of scale usage heterogeneity are left unexplored. Although the method is simple, there are three problems. First, the procedure assumes that the raw data contain interval information, even though ratings probably only yield ordinal data. Second, the within-person estimates of scale usage (means and standard deviations) may not be very reliable, particularly if they are based on few responses. Third and most importantly, the respondent-specific means and standard deviations are supposed to be "pure" measures of scale usage, but since they are based on the same items for which substantive analyses are to be conducted, it is likely that scale usage will be confounded with content. More sophisticated methods are available that correct for the first two problems. For example, Rossi, et al. (2001) proposed a Bayesian approach that properly handles the ordinal nature of the data and provides a more reliable individual-level correction for scale usage heterogeneity. However, the confounding of content

and scale usage cannot be circumvented when evidence about scale usage is derived from the substantive items themselves.

In the second approach, independent control items are used to get an estimate of differential scale usage, and usually specific reasons for scale usage heterogeneity are posited and assessed. Common mechanisms giving rise to individual differences in scale usage are net acquiescence (a preference for the agreement versus disagreement, or, more generally, the positive versus negative response options on the rating scale), extreme responding (a preference for the most extreme response categories on either side of the rating scale), and midpoint responding (a preference for the middle position on the rating scale). Collectively, these biases are referred to as response styles (Baumgartner & Steenkamp, 2001). For a 5-point response scale, measures of net acquiescence (NARS), extreme responding (ERS), and midpoint responding (MRS) for each respondent can be computed as follows:

$$NARS = [f(5)*2 + f(4)*1 - f(2)*1 - f(1)*2]/J$$

$$ERS = [f(1) + f(5)]/J$$

$$MRS = f(3)/J$$

where f (o) refers to the number of times that a respondent selects response option o across all control items (e.g., f(5) refers to the frequency of endorsement of the most positive response category), and J is the number of control items. Provided that the control items are heterogeneous in content (i.e., they do not share common content), these response style measures can be expected to be “pure” measures of scale usage, and if the control items have no content overlap with the substantive items, they can be used to correct for scale usage differences.

Specifically, one way to purify the original data is to regress the raw scores for each of the substantive items on the various response style measures. The residuals from this regression are the corrected scores purged of stylistic response tendencies, which can be used in subsequent analyses (Baumgartner & Steenkamp, 2001). We call this procedure residualization. A variation on this technique is to use the response style measures as covariates in substantive analyses of interest. A sophisticated version of this approach is the Representative Indicators Response Style Means And Covariance Structure (RIRSMACS) approach proposed by Weijters, et al. (2008). With this method, each substantive item is related to a comprehensive set of response styles that are measured by multiple items. That is, each response style factor is indicated by multiple measures of each response style, and each substantive item is specified as a function of NARS, ERS, and MRS factors. In their paper, Weijters, et al. (2008) applied the RIRSMACS model to a seven-point scale and used separate acquiescence and disacquiescence measures, but for the five-point scales used in our empirical studies, this would lead to identification and convergence problems due to collinearity, so we will use NARS instead. The major advantage of multiple measures is that the response styles can be assessed more reliably and that measurement error is not passed on to the residualized scores, which presumably have been corrected for extraneous influences. Two disadvantages of the approach are that a relatively large number of heterogeneous control items is needed to construct multiple indicators of each response style, and that many additional parameters are estimated.

The two approaches discussed above are designed to control scale usage heterogeneity at the respondent level. Such individual-level measures of scale usage heterogeneity can be used to conduct comparisons across groups or conditions, and individual-level response style

measures can serve to remove stylistic variance from substantive measures so that substantive comparisons can be based on corrected scores. However, scale usage differences need not always occur at the respondent level, and under certain circumstances it may be preferable to correct for differential scale usage at the group level. For instance, prior research has shown that translations of response category labels can result in labels whose meaning varies across different languages, and that as a consequence endorsement rates may differ simply because nonequivalent labels were used in different languages (Skevington & Tucker, 1999; Smith, et al., 2005; Szabo, et al., 1997; Weijters, et al., 2013). In particular, response category labels used in different languages may vary in intensity. For example, in a comparison of response category labels in the U.S. and Germany, Smith, et al. (2005) found subtle intensity differences between apparently equivalent labels such as 'definitely agree' and its German translation, 'stimme bestimmt zu' (i.e., 'definitely' is a stronger term in English than 'bestimmt' is in German). Furthermore, Weijters et al. (2013) recently demonstrated that a category label may be more idiomatic and thus more familiar in one language than another. Since more familiar labels are more readily endorsed by respondents, regardless of the substantive content of the items in question, this can lead to bias. For example, 'tout à fait d'accord' in French is more familiar than 'strongly agree' in English and this leads to greater endorsement of the endpoint categories in French than in English, regardless of content (Weijters, et al., 2013).

Thus, nonequivalent response category labels are a potential source of systematic bias. This problem is particularly serious since response category labels vary across languages, but often the same response options are used across all or at least many items of a questionnaire within a language (Rindfleisch, Malter, Ganesan, & Moorman, 2008). If response frequency

distributions differ for reasons unrelated to content, this causes cross-linguistic bias in parameter estimates (De Jong, Steenkamp, Fox, & Baumgartner, 2008). However, the bias occurs because the meaning of the category labels differs across languages (or nationalities) and is thus a group-level phenomenon. If this is the case, the correction for scale usage differences should occur at the group level, not at the level of individual respondents. The benefit of a group-level correction is that it may be possible to get more reliable and valid estimates of differential scale usage, and the correction is simpler.

We have focused on the example of differential scale usage across language groups due to nonequivalent response category labels to motivate the group-level correction. But the approach is more general and can be used whenever there are concerns that a group-level confound may invalidate comparisons across groups. First, response style research has shown that a substantial proportion of response style variation occurs at the country level (De Jong, et al., 2008; Van Rosmalen, Van Herk, & Groenen, 2010), so it is possible that even when optimally equivalent response category labels are used in cross-national research, respondents from different nationalities use response scales differently. Second, cross-group differences in scale usage are prevalent in contexts other than cross-national survey research. Examples include research contexts in which different scale formats are used across groups of respondents, studies in which different data collection methods are used for different groups, and experiments in which manipulations result in a different perception of scale category meanings across experimental conditions. The distinguishing features of the proposed approach are that the correction occurs at the group level, that it is based on independent control items, and that it does not require the identification of specific sources of scale usage differences.

3. Conceptual development

3.1. The forgotten alternative coding approach for Likert items

When introducing his popular attitude measurement scale, Likert (1932) initially considered two alternative ways of coding the data. In one approach, consecutive integers are assigned to the response categories. For example, for a five-point scale, integers of 1 to 5 are assigned to the five scale positions (1 = ‘strongly disagree’, 2 = ‘disagree’, 3 = ‘neither agree nor disagree’, 4 = ‘agree’, 5 = ‘strongly agree’). Although the assignment of consecutive integers is arbitrary and the implicit equal-interval assumption may not be justified, this method has become the norm for researchers using Likert-type response scales, and we thus refer to it as the traditional approach. With the other method (which Likert called the sigma method of scoring), each response category is assigned a value based on the proportion of respondents who selected a given scale position. Specifically, the endorsement frequencies of the different response categories are transformed into a z-score using the cumulative normal probability distribution (Likert, 1932, p. 22).

Likert specifically devised the sigma method in order to make responses to different response formats comparable (e.g., a graded multiple choice question and a strength of approval item). We follow Likert’s lead and propose that, in a situation in which the comparability of responses might be questionable because of the presence of scale usage heterogeneity across groups or conditions (e.g., when response category labels have different meanings for different groups due to different languages or experimental manipulations), using the more complex sigma method of coding will be more appropriate than using the simpler traditional coding system. However, we also extend Likert’s original sigma method in

two ways. First and foremost, we calculate the sigma values on the basis of the response proportions observed across a *heterogeneous* set of control items, where heterogeneous means that the items share no common substantive content. The control items are calibration items that are included in the questionnaire specifically for the purpose of assessing scale usage differences that are independent of content. The sigma values derived from the control items are then applied to the focal items (i.e., the items in the questionnaire that are of substantive interest, as opposed to the calibration items).[1] This avoids the confounding of content and scale usage that may arise when scale usage differences are inferred from the substantive items to be compared. Second, we use a different set of sigma values to code the response categories for different groups of respondents to correct for the fact that there may be scale usage differences across the different groups of respondents. In other words, we use the control items to get a "pure" measure of scale usage (because the control items presumably share no common substantive content) for each of the groups to be compared, and we then use the sigma values obtained from the control items in each group to weight the responses to the substantive items.

3.2. Estimating group-specific response frequency distributions based on specially-chosen control items

[1] Procedures similar to the Likert (1932) sigma method are also implemented in statistical packages such as SAS (referred to as Blom, Tukey, or van der Waerden normal scores based on ranks) and LISREL. As in Likert's approach, these transformations use the cumulative distribution of the responses to the substantive item of interest itself as the basis for the normalization function. This does not allow the analyst to correct for scale usage independently of content, which is the main characteristic of the calibrated sigma method proposed here. The calibrated sigma method aims to weight scores in a group-specific way without sacrificing the ability to make comparisons between groups.

Research on response styles has made it plain that it is necessary to clearly differentiate between responding based on content and responding based on factors unrelated to content (Baumgartner & Steenkamp, 2001). If the diagnosis of, and correction for, scale usage heterogeneity is based on the same items that are also used to conduct substantive comparisons on variables of interest, there is the danger of confounding scale usage variance and substantive variance. To avoid this problem, the preferred approach is to use independent control items (Podsakoff, et al., 2012). These control items should satisfy two requirements. First, they should be heterogeneous in content. If the items have little or nothing in common, then responses to these items can be treated as pure measures of scale usage. Second, there should be no overlap in content between the control items and the items for which substantive comparisons are to be conducted.

In prior research, the scale developed by Greenleaf (1992b) has sometimes been used for this purpose. This scale consists of items such as 'I am a homebody,' 'A college education is very important for success in today's world,' 'When I see a full ashtray or waste-basket, I want it emptied immediately', 'I eat more than I should', and 'No matter how fast our income goes up, we never seem to get ahead'. Although Greenleaf proposed these items to measure extreme responding, the heterogeneous content of the items makes them well-suited for quantifying differences in response patterns in general (Weijters, et al., 2013). We will use this scale in our empirical work and henceforth refer to the scale as the Response Pattern Scale (RPS). The use of a standard set of heterogeneous items to quantify the relative use of the various response categories optimizes both internal validity (no confounding of content and style) and external validity (the findings can be generalized to other items and content

domains). Furthermore, the use of a common scale makes our results comparable to those of other studies based on the same item set (Arce-Ferrer, 2006).

3.3. Correcting group-specific response biases

Researchers who have recognized the need to control for differential scale usage have generally applied a correction at the individual-respondent level. Although it is certainly possible that scale usage varies by respondent, it can be difficult to get valid and reliable individual-level estimates of scale usage, and models incorporating the possibility of scale usage heterogeneity are usually forced to make various simplifying assumptions (in order to identify the model) that may not be satisfied in practice. Furthermore, under certain circumstances, there are theoretical reasons to expect that differential scale usage occurs across groups of respondents (or conditions) rather than individual respondents, in which case it seems most appropriate to apply a correction at the group level. For example, if it is thought that the meaning of response category labels varies across different languages and that, as a consequence, different language groups differentially select certain scale positions, then one should correct for this difference at the language level.

Following this logic, we propose to code the response categories of all items in a group-specific way, based on the proportion of endorsements of each response category by a group of respondents answering the control items (e.g., the RPS). For example, if the response category ‘strongly disagree’ is used *less* often in one group than in another, regardless of what respondents are being asked, the choice of ‘strongly disagree’ in response to a substantive item should get a *more* extreme weight in the first group than in the second to eliminate the effect of differential scale usage and equate the response behavior of the two groups. Thus, instead of

arbitrarily coding observed responses by using consecutive integers, the responses are coded in an empirical way based on the frequency with which a group of respondents selects the response categories for a set of heterogeneous (content-free) control items. In this context, it is worth pointing out that Likert (1932) proposed the traditional coding approach based on evidence from a rudimentary experiment in a single-sample setting, which showed that the simpler procedure led to results that were very similar to those of the more complex sigma method. Although Likert himself cautioned that his findings were exploratory and needed to be replicated, subsequent researchers have simply continued this practice without much additional supportive evidence.

3.4. Formal model

Formally, we propose the following approach. For each group g, and for all response categories k (where the total number of response categories is K), the cumulative proportion of consecutive responses (P_{1g}, P_{2g}, …, P_{Kg}) is computed across the control items (e.g., the items in the RPS). Assuming approximate normality of the underlying distribution, the sigma value for category k in group g is obtained as,

$$\sigma_{k,g} = \Phi^{-1} [½ * (P_{k,g} + P_{k-1,g})],$$

where Φ^{-1} refers to the inverse cumulative normal distribution function, with $P_0 = 0$ and $P_K = 1$. Instead of coding the response categories for the substantive items with their rank number k, the observed responses get assigned the $\sigma_{k,g}$ value corresponding to the response category of the group in question. Note that the cumulative proportions are computed across the control

items, not the substantive items. Furthermore, the sigma values could, in principle, be computed based on a subsample of all respondents in a given group (so that not all respondents have to complete the control items, if survey completion time is at a premium), or a pretest with representative samples of respondents from each group could be conducted to get the sigma values. This makes the method potentially more cost-efficient, although it is crucial that the sigma values obtained from the subsamples or pretest samples be representative of the full samples for which substantive comparisons are to be conducted.

3.5. Illustrative example

In Table 1 we provide a brief worked example of how to compute calibrated sigma values based on hypothetical responses to the 16 five-point RPS items by two samples of respondents in two different groups.

Insert Table 1 about here

In step 1, the mean frequency with which each scale category is chosen across the 16 control items has to be computed. For example, in MS Excel one can use the '=COUNTIF()' function to compute the frequencies of each response category across the 16 RPS items and then calculate the means of the resulting variables. In step 2, the frequencies are recoded as proportions (in the example by dividing the mean frequencies by 16). In step 3, the cumulative proportions are calculated based on the proportions in step 2. In step 4, the midpoint of each category proportion is calculated. Finally, in step 5, these midpoint proportions per category are transformed into calibrated sigma codes (which correspond to standardized z-scores). For example, in MS Excel one can use the formula '=NORM.S.INV()'. The sigma values obtained for the two groups can then be used to recode the responses to the substantive items of

respondents from groups A and B, respectively (so in group A, for instance, a ‘strongly disagree’ response would be coded as -1.96).

4. Study 1: Recoding data to remedy scale usage differences across languages

To illustrate the newly proposed method, we demonstrate its use on a simulated data set. This allows us to be certain of the true underlying model, which is impossible with empirical data. The goals of the simulation study are to (1) illustrate a hypothetical model of how cross-language response bias may come about, (2) demonstrate the use of the calibrated sigma method and compare it to the traditional coding scheme for Likert items, and (3) show how language-specific scale usage may cause bias that is hard to detect with classic measurement invariance tests because these tests assume that the bias is non-uniform across items. If, as hypothesized, tests of measurement invariance cannot detect language-specific differential scale usage, the implication is that measurement invariance testing should be complemented with the calibrated sigma coding method.

4.1. Method

Data generation. Imagine a scenario where a company wants to compare trust in its brand across twelve language groups, using a three item scale. We simulate data for a sample of N=12,000, with 1000 respondents for each of the following languages: Dutch, French, English, German, Spanish, Polish, Slovakian, Hungarian, Romanian, Swedish, Italian, and Turkish. We assume one latent construct, $\xi_1 \sim N(0,1)$, measured by means of three indicators;

80 percent of the variance in the three indicators is explained by the latent variable, the remaining 20 percent is explained by unique factors $\delta_p \sim N(0,1)$, where p equals 1, 2 or 3.

We used the Monte Carlo facility in Mplus 7.3 to generate the data (using the default seed = 0). All observations are drawn from the same normal distributions irrespective of language group (i.e., one normal distribution for the latent construct ξ_1 and three unique factors δ_p). For each indicator variable, a continuous variable was created as the weighted sum of the latent construct and the indicator's unique term such that 80 percent of the total variance was due to the latent construct. Next, all continuous variables were categorized into seven response categories using the frequency distribution of the RPS items for each of the 12 languages observed in Study 3 of Weijters, et al. (2013). In other words, we created language-specific cutoff values (lower and upper boundaries for each response category) based on the cumulative proportions of the seven response categories across the RPS items observed in an actual study (i.e., although we use simulated data for the construct measures, the scale usage data for the 12 languages are based on actual empirical data). These cutoff values were then used to categorize the simulated continuous indicators: every value of the initially continuous variable was assigned to a response category if it fell between the response category's upper and lower boundary. The boundaries were computed as follows:

$$\text{Upper boundary}_{k,g} = \Phi^{-1}(P_{k,g}),$$

where k = response category 1 through 7, g = language 1 through 12, $P_{k,g}$ are the language-specific cumulative proportions on the RPS items for each response category, and Φ^{-1} is

the inverse cumulative normal distribution function (response category 7 does not have an upper boundary).

Internal consistency of the RPS items. The calibrated sigma method is based on the assumption that across-group differences in response category endorsement represent consistent variation in response patterns that are not specific to any one of the RPS items. To assess whether the RPS is a reliable group-level measure, we therefore computed Cronbach's alpha coefficients for the endorsement likelihoods of each response category, where the variables are the proportion of respondents who endorsed a given response category (e.g., strongly disagree, disagree, etc.) for the 16 RPS items and the unit of analysis is the language group. In other words, Cronbach's alpha measures how consistently the 12 language groups endorsed a given scale position (e.g., strongly disagree) across the 16 RPS items. Table 2 reports the results. All Cronbach's alpha coefficients exceed .80, which demonstrates consistent scale usage by different language groups across items that share little or no common substantive content. This conclusion aligns with earlier applications of the Greenleaf RPS to assess cross-national differences in response patterns (Clarke III, 2001).

Insert Table 2 about here

To summarize, we find highly consistent endorsements of each of the seven response options (e.g., similar proportions of 'strongly disagree' responses) across the 16 items in the RPS by different language groups. If the items shared common content and reflected the same or related constructs, there would be a substantive reason for this consistency in responding by different language groups. However, the 16 calibration items were purposely chosen to be free of common content, which makes it implausible that the response pattern consistency represents anything other than style (Greenleaf, 1992a, 1992b). In other words, systematic

differences across language groups in consistent responding to content-free items are most likely due to scale usage differences, not substantive differences. Whether cross-group differences in scale usage are the result of differential familiarity with the response category labels in different languages, culture-related response styles, communicational norms, or other confounds is immaterial and does not affect the effectiveness or validity of the calibrated sigma method.

Application of the calibrated sigma method. Next, the newly proposed method was applied and the initial integer values (i.e., 1 for strongly disagree, 2 for disagree, etc.) were recoded in accordance with the calibrated sigma method. The mapping functions for the twelve languages are shown in Table 3. For example, for all Dutch language respondents, 1 was recoded as -1.92, whereas for German language users, 1 was recoded as -1.66. This recoding reflects the fact that the first category (strongly disagree) was selected more often by German respondents regardless of item content, so that this category gets a less extreme weight for German respondents relative to Dutch respondents.

We ran the data generation procedure twice to simulate two alternative scenarios, labeled equal latent means vs. different latent means. In the first scenario, the latent factor means were equal across the 12 language groups. In the second scenario, 6 language groups (those with odd rank numbers in the data set and Table 3) were assigned a latent factor mean of -.25, while the remaining groups (those with even rank numbers) were assigned a latent factor mean of .25. The purpose of this manipulation was to demonstrate that, under the current data generation model, the calibrated sigma method will detect non-invariant latent means when the means actually differ across language groups (but not so when they do not differ), whereas traditional invariance testing will wrongly conclude that the latent means are

different when the latent means are actually the same but different language groups use the response scale differently.

Insert Table 3 about here

4.2. Findings

For each of the two scenarios (equal latent means vs. different latent means), we have two data sets for the observed variables: one based on traditional coding, the other based on calibrated sigma coding. For each of these four combinations (scenario by coding), we test a Means And Covariance Structure (MACS) model of the brand trust factor with three indicators and compare the findings. For all data sets, we test a sequence of nested models (Steenkamp & Baumgartner, 1998). Assessment of measurement invariance is best approached from a modeling perspective rather than a strict statistical hypothesis testing perspective, and the following indices of model fit are particularly well-suited in this context: TLI, CFI, RMSEA and an information criterion such as AIC or BIC (Little, 1997; Steenkamp & Baumgartner, 1998). BIC is used to select the optimal model, since it trades off closeness of fit with parsimony (i.e., the penalty for adding parameters increases with sample size); a lower BIC for a model relative to other models indicates a more optimal parameterization (Mulaik, 2009; Wicherts & Dolan, 2004).

We start from the unconstrained model (model A) where the indicators freely load on their underlying factor. Note that this model has zero degrees of freedom (and consequently perfect fit), making it easier to interpret the subsequent deterioration in fit. The next models test for metric invariance (i.e., equal factor loadings across groups; model B), scalar invariance

(i.e., equal intercepts in addition to equal loadings; model C), and factor mean invariance (model D), respectively. If latent means are compared across groups, both metric and scalar invariance have to be satisfied (Steenkamp & Baumgartner, 1988). The fit indices for these models are shown in Table 4 for both the traditional data and the calibrated sigma data in the equal versus different latent means scenarios.

Insert Table 4 about here

Let us first look at the equal latent means scenario. In the traditional data, the fit indices point toward model C (scalar invariance without invariance of latent means) as the preferred model (i.e., model C has the lowest BIC value, and the other fit indices support this conclusion). On the other hand, in the calibrated sigma data, the fit indices support the hypothesis of invariant latent means, which indeed corresponds to the known true situation. Figure 1 shows the estimated means and their corresponding 95% confidence intervals for the traditional data and the calibrated sigma data (groups are shown in order of their RPS mean). Clearly, for the traditional data the estimated means are biased in the direction that is to be expected given the scale usage differences observed across groups, with the lowest mean in the German language and the highest mean in the Turkish language (see Table 3). These differences are not present in the data based on the calibrated sigma method.

Figure 1 here

Next, consider the different latent means scenario. In this case, in accordance with the true situation, the hypothesis of invariant latent means is rejected in the data set based on calibrated sigma coding. As seen in Figure 2, the language groups with odd rank numbers have latent means that are about .5 higher than those with even rank numbers (groups are ordered by their true means, and – within the true mean ordering – by RPS mean). Although

the hypothesis of equal latent means is also rejected in the traditional data, the latent means do not properly reflect the .5 difference between odd and even rank numbers. The reason is that the estimated latent means are not only influenced by the true difference in latent means but also by scale usage differences between the language groups.

Figure 2 here.

4.3. Discussion Study 1

The simulation study demonstrates some important points. First, it provides a hypothetical model of how response bias across languages may arise. Second, for cases where this model holds, the simulation demonstrates the effectiveness of the calibrated sigma method, both when there are no genuine differences in the focal construct across groups and when there are genuine differences. Importantly, when the traditional coding method is used, the bias introduced by differential scale usage across language groups is not necessarily unmasked by measurement invariance tests. On the contrary, the sequence of model tests may give very reassuring results for the hypothesis of measurement invariance, while providing clear reason to reject the hypothesis of equality of latent means, even though the latent means are actually equal by design. By contrast, when measurement invariance testing is applied to data coded with the calibrated sigma method, the analysis does not lead to biased estimates of latent means and misleading comparisons of means across language groups.

5. Study 2: Empirical application of the calibrated sigma method in a dual-language setting

To illustrate the use of the calibrated sigma method in an empirical setting, we compare samples of Dutch- and French-speaking Belgian respondents on their self-reported Need for Predictability. Need for Predictability is a facet of Need for Closure and, as the name of the construct suggests, refers to a preference for having secure knowledge that implies trans-situational consistency (Kruglanski, et al., 1997; Webster & Kruglanski, 1994). Even though in non-simulated settings it is impossible to know the true value of a latent variable with certainty, there are theoretical reasons to expect that Dutch- and French-speaking Belgians do not differ substantially from each other in terms of their Need for Predictability. In particular, the scores on uncertainty avoidance, a construct closely related to Need for Predictability (Richter & Kruglanski, 2004), are 97 and 93, respectively, for (Dutch-speaking) Flanders and (French-speaking) Wallonia – as compared to 53 for the Netherlands and 86 for France (Hofstede, 2001).[2]

In addition, there are several 'hard' indicators that support a lack of difference between Flanders and Wallonia in terms of need for predictability and uncertainty avoidance. For example, it is well established that uncertainty avoidance is negatively related to national rates of innovation (Shane, 1993). Consistent with Belgium's high score on uncertainty avoidance, its TEA index (Total Entrepreneurial Activity, measured as the percentage of individuals of

[2] Note that only one of the three measures on which uncertainty avoidance is based is a Likert-type scale, as two of the items are rated on scales ranging from '2 years at the most' to 'until I retire' (for employment stability) and from 'I always feel this way' to 'I never feel this way' (for stress). Therefore, it does not seem likely that the scales used by Hofstede to construct the uncertainty avoidance scores may suffer from the same bias that we have identified.

the active population who are involved in the start-up of an enterprise younger than 3.5 years) was quite low in 2008, and the index values were similar for Flanders and Wallonia (Bosma & Levie, 2010; Sleuwaegen & Buysse, 2010). Furthermore, Hofstede (1991) presents data that uncertainty avoidance is negatively correlated with the adoption of new media, use of the internet, and teletext. Consistent with this evidence, the adoption of online shopping was rather slow in both Dutch- and French-speaking Belgium (Europe, 2013), presumably because this form of shopping deviates from the known and trusted brick-and-mortar shopping. It is therefore plausible that if a difference is observed between the two groups in terms of Need for Predictability, it can be mainly attributed to differential scale usage bias, especially when the bias is in the same direction as the response distribution differences observed for unrelated items. This bias may be remedied by applying the calibrated sigma method.

To summarize, the aims of this study are to investigate whether a difference in Need for Predictability is observed between Dutch- and French-speaking Belgians and whether this observed difference persists after applying the calibrated sigma method. In addition, we will compare the results obtained with the calibrated sigma method not only to the results based on traditionally coded data, but also to the results obtained with a more elaborate correction method that accounts for individual variation in multiple response styles: the Representative Indicators Response Style Means And Covariance Structure approach or RIRSMACS (Weijters, et al., 2008).

5.1. Method

Data were collected by means of paper-and-pencil questionnaires distributed door-to-door using a random walk procedure, which yielded 538 valid responses. We obtained 292

usable and complete questionnaires from Dutch-speaking respondents and 246 from French-speaking respondents. In the Dutch-speaking sample, 55.4% of the respondents were female and the average age was 40.74. In the French-speaking sample, 59.8% were female and the average age was 40.34 years.

Need for Predictability was measured by means of eight items, three of which are reverse-scored (see the Appendix of Kruglanski, Webster, and Klem (1993) for the items). To measure and remedy response scale usage differences, the RPS was also included in the questionnaire (Greenleaf, 1992b). All items were administered in a five-point Likert format.

5.2. Findings

We ran measurement invariance tests in three different ways: (1) using the traditional coding, (2) using the calibrated sigma coding (see Table 5), and (3) using traditionally coded data corrected by means of the RIRSMACS approach. In preparation for applying the RIRSMACS method, we first ran preliminary analyses that supported metric and scalar invariance for the response styles NARS, ERS and MRS. Furthermore, these analyses also indicated a clear and significant between-group difference in response style means, as reported in Table 6 and consistent with the results in Table 5: the French-speaking sample had higher NARS and MRS means, but a lower ERS mean. The stronger tendency to agree rather than disagree (NARS) suggests inflated mean estimates in this sample relative to the Dutch-speaking sample.

Insert Table 5 about here

Insert Table 6 about here

For the focal analysis, we specified a two-group MACS model, with language as the grouping variable, in which the eight items of Need for Predictability load on one factor, with freely correlated residual terms for the three reversed items (Marsh, 1996). The model fits the data relatively well. Table 7 gives an overview of the fit indices for the three alternative approaches. Table 6 reports latent mean estimates for the French-speaking sample relative to the Dutch-speaking sample in terms of Need for Predictability under the three different scenarios.

Insert Table 7 about here

In the traditionally coded data, the scalar invariance model is the preferred model and the hypothesis of invariance of factor means is rejected (in favor of scalar invariance without means invariance; see the BIC values in the last column of Table 7). Indeed, the French-speaking sample has a factor mean that is significantly higher than the mean in the Dutch-speaking sample (see Table 6). By contrast, if we apply sigma coding to the data (see Table 5), the invariant latent means model is the preferred model and the mean estimate in the French-speaking sample does no longer differ significantly from that in the Dutch-speaking sample (see Table 6). The RIRSMACS-corrected model yields invariance testing results that are similar to those of the calibrated sigma model (see the last column of Table 7 for the BIC values) and also indicates no significant mean difference between the two language groups (see Table 6). The sign of the difference even reverses, but we should note that the standard error of the mean estimate is also markedly larger than in the other two analyses; this indicates lower precision of the parameter estimate based on the RIRSMACS model (which may be due to the much higher number of parameters estimated in this model).

5.3. Discussion Study 2

Study 2 demonstrates how the calibrated sigma coding of data works in an empirical setting. The application leads to some relevant conclusions about the method. First, factor mean comparisons that show significant differences based on traditional coding may no longer be significant when using calibrated sigma coding. In the current situation, there are good reasons to believe that the observed mean difference in Need for Predictability is a method artifact. Specifically, in an independent heterogeneous set of items (the RPS scale), we can already detect differences between the two groups in terms of their pattern of responding to content-free items: the French-language data show more midpoint and fewer disagreement responses (see Table 5 and the response style means in Table 6). This difference in scale usage is consistent with the observed difference in latent means based on traditional coding. Moreover, theoretically, no mean difference is expected, because both groups are culturally similar in terms of uncertainty avoidance levels. Also note that the small difference in uncertainty avoidance (Hofstede, 2001) is in the direction opposite to that observed for Need for Predictability, whereas it is in the direction we would expect based on the response pattern differences between the two groups. When the scale usage difference across the two language groups is accounted for by using the calibrated sigma method, the factor mean difference becomes smaller and statistically non-significant. Finally, when applying an individual-level correction for differences in stylistic responding (i.e., the RIRSMACS method), the factor mean difference also becomes statistically non-significant.

When evaluating the parsimony-adjusted fit by means of BIC, a remarkable similarity becomes evident between the pattern shown by BIC across the models estimated in the simulation study (Study 1; see Table 4) and in the empirical study (Study 2; see Table 7). For the first three models (i.e., unconstrained, metric invariance, scalar invariance), the model fit

indices are nearly identical for the traditional and sigma coded data. However, for the model that additionally imposes factor mean invariance, the BIC goes up in the traditional method, whereas it continues to decrease for the calibrated sigma method. This means that measurement invariance tests do not necessarily provide evidence of differential method bias across groups when this bias is uniform across items, which is the case when differential scale usage is the cause of method bias (Little, 1997; Weijters, et al., 2008). However, in traditionally coded data the uniform bias does influence the results, as the latent mean estimate captures not only the central tendency of the substantive latent variable but also the central tendency in scale usage.

6. Study 3: Empirical application of the calibrated sigma method to a survey experiment

Although the most plausible explanation for the results in Study 2 is a scale usage difference between Dutch-speaking and French-speaking respondents, the findings are not conclusive since the "true" Need for Predictability of Flanders and Wallonia is unknown and respondents cannot be randomly assigned to a language or cultural group. Therefore, to complement Study 2, Study 3 offers a second empirical application of the calibrated sigma method, this time focusing on a situation in which measurement non-equivalence between two groups is the result of an experimental manipulation of the response scale format.

Specifically, all respondents completed the same scale, Susceptibility to Normative Influence or SNI (Bearden, Netemeyer, & Teel, 1989), but they were randomly assigned to two alternative scale formats, which were designed to elicit specific scale usage differences. In one format, only the endpoints were labeled: 'disagree' for the first category and 'agree' for the fifth category. In the other format, all five response categories were labeled: 'extremely

disagree,' 'disagree,' 'neither agree nor disagree,' 'agree,' and 'extremely agree'. The endpoint-labeled format was expected to lead to more endpoint responses than the fully labeled format for two reasons. First, if only the endpoints are labeled, they are comparatively more salient and less ambiguous than the other response categories, which should encourage greater endorsement of the endpoints (Weijters, Cabooter, & Schillewaert, 2010). Second, the labels of the endpoints in the endpoint-labeled condition are less intense and therefore more likely to be endorsed (de Langhe, Puntoni, Fernandes, & van Osselaer, 2011). The SNI scale generally has a mean below the midpoint (Bearden, et al., 1989), because consumers typically do not believe that their behaviors are influenced by others (Bearden, et al., 1989; Gopinath & Nyer, 2009). For scales that have a scale mean below the midpoint, response style theory predicts that increased endpoint responding will lead to a lower mean estimate, because extreme responding mostly leads to extreme negative rather than extreme positive responding (Baumgartner & Steenkamp, 2001). As a consequence, we expected that the SNI scale would show a lower mean in the endpoint-labeled condition than in the fully labeled condition. However, this difference does not reflect a content-based (substantive) difference since respondents were randomly assigned to conditions and SNI should be the same on average; any observed difference is simply a scale usage difference due to the use of different response scale formats. Hence, the intent of the study was to experimentally induce a scale usage difference and to demonstrate that the resulting bias can be corrected using the calibrated sigma method.

In summary, in the second empirical study we provide further evidence in support of the validity of the calibrated sigma approach, and we illustrate that the method can also be applied in experimental settings in which a manipulation causes a difference in the meaning of

scale categories and therefore different scale usage. Finally, we also show that the sigma values can be computed based on a random subsample and then applied to the complete sample.

6.1. Method

Data were collected from the Amazon Mechanical Turk panel. In the sample (N = 455), age ranged from 19 to 70 years (M = 37.1, SD = 11.9) and 46.4% of respondents were women. The questionnaire contained the 16 RPS items and the eight items measuring the normative factor of Susceptibility to Normative Influence or SNI (Bearden, et al., 1989). Example items are 'I rarely purchase the latest fashion styles until I am sure my friends approve of them,' and 'It is important that others like the products and brands I buy' (Cronbach's alpha = .94).

All 24 items were administered using five-point rating scales, but as explained previously, respondents were randomly assigned to two alternative scale formats, either a format in which only the endpoints were labeled ('disagree' for the first category and 'agree' for the fifth category), or a format in which all five response categories were labeled ('extremely disagree,' 'disagree,' 'neither agree nor disagree,' 'agree,' and 'extremely agree').

6.2. Findings

We coded the responses to the SNI items in different ways. With traditional coding, the response categories were coded with consecutive integers from one to five, irrespective of experimental condition. As expected, the traditionally scored scale mean was below the midpoint (i.e., the midpoint is three on a five-point scale) in both conditions, but the mean was

lower in the endpoint-labeled condition than in the fully labeled condition ($M_{\text{all labeled}}$ = 2.15, SE = .056; $M_{\text{endpoint-labeled}}$ = 1.91, SE = .059); this difference was statistically significant: t(453) = 2.953, p = .003). Thus, the endpoint-labeled condition had more extreme responses than the fully labeled condition, which leads to more extreme (lower) SNI scores in that condition. However, this is merely a scale usage difference, and once calibrated sigma scoring is used, which corrects for this scale usage difference, the difference goes away: With calibrated sigma coding, the categories were coded based on the response pattern observed in the RPS data, separately for each experimental response format condition, as shown in Table 8. As expected, the endpoint-labeled format led to more extreme responses. The higher endorsement rate of endpoints in the endpoint-labeled format is reflected in the sigma values for the endpoint categories, which are less extreme than the sigma values for the fully labeled condition. When the sigma values were used to code the responses in the two experimental conditions, the difference in SNI between the two groups was no longer significant: t(453) = -.047, p = .963.

To illustrate that the sigma values can also be computed based on a random subsample and then applied to the complete sample, we randomly sampled approximately half of the respondents and computed sigma values based on this random subsample (separately for each of the two conditions), as shown in the last two columns of Table 8. In the data based on the split-half calibrated sigma coding, an independent samples t-test also showed no significant difference between the two groups: t(453) = .854, p = .393.

Insert Table 8 about here

We also computed response style indicators consistent with RIRSMACS guidelines to provide additional evidence, based on another method, that it is important to take into account scale usage differences and that the calibrated sigma method leads to the right conclusion.

Three indicators each were computed for NARS, ERS and MRS. Preliminary analyses supported metric and scalar invariance for the response style factors, but showed a significant difference in response style means, as reported in Table 9. In line with expectations, the endpoint labeled group showed more extreme responses (and slightly more midpoint responses) at the expense of responses that express moderate (dis)agreement (see Table 8).

Insert Table 9 about here

We ran measurement invariance tests in four different ways: (1) using the traditional coding; (2) using the calibrated sigma coding; (3) using the split-half calibrated sigma coding; and (4) using traditionally coded data corrected by means of the RIRSMACS approach. Table 9 reports the factor mean estimates based on the alternative analyses, while Table 10 presents the model fit results.

Insert Table 10 about here

In the traditionally coded data, the scalar invariance model is the preferred model and the hypothesis of invariance of factor means is rejected. Compared to the group that used the fully labeled scale format (whose factor mean was set to zero), the group that used the endpoint-labeled scale format had a significantly lower factor mean. In contrast, if we apply sigma coding to the data, the invariant latent means model is the preferred model (see Table 10) and the mean estimates in the two experimental groups do not differ significantly (see Table 9). Applying the subsample-based sigma codes to the full sample yielded results that were equivalent to those using sigma codes computed from the full sample (see Tables 9 and 10). Furthermore, the RIRSMACS-corrected model yields similar conclusions: the model with invariant factor means is the preferred model (see Table 10) and the factor mean in the

endpoint-labeled condition is not significantly different from the factor mean in the fully labeled condition (see Table 9).

6.3. Discussion Study 3

The results of Study 3 offer additional support for the validity of the calibrated sigma method. Respondents were randomly assigned to two alternative scale formats, which resulted in a significant factor mean difference for the traditionally coded data. By contrast, when the data were coded using the calibrated sigma method (based on either the complete sample or a random subsample), the factor means were no longer significantly different. Since the respondents were randomly assigned to experimental conditions, there is no plausible explanation other than differential scale usage. Moreover, an individual-level response style correction (the RIRSMACS method) yielded results that were consistent with those based on the calibrated sigma method.

8. General discussion

We proposed a method that can be used to correct for group-level scale usage heterogeneity, that is, differences across groups in the endorsement likelihood of response categories that are unrelated to the content of the items. Such differences can be due to several aspects of the data collection that affect the perceived meaning of the response scale categories, including the use of a different scale format, the use of experimental manipulations, the use of a different language, and/or culturally driven response style differences. Although the practice of assigning consecutive integers to the response categories in Likert items is widely accepted,

it may be inappropriate in such settings. We therefore advocate that dedicated calibration items which are free of common substantive content be included in questionnaires, in order to capture the relative frequency with which the response categories are chosen by different groups of respondents.

The calibrated sigma method has several advantages. Most importantly, it avoids the arbitrariness in the assignment of particular numerical values to response categories. Furthermore, it is compatible with the linear model, which is used by most marketing researchers, and can be easily integrated into commonly used data analysis procedures (including measurement invariance testing based on confirmatory factor analysis). In addition, it does not require the use of complex nonlinear models involving the estimation of many additional measurement model parameters. On the contrary, there is no limitation to the number of Likert items and underlying latent factors that can be corrected with the calibrated sigma method, because the approach consists of a simple recoding operation, after which the common linear model (e.g., for factor analysis and/or regression) can be applied in the usual way.

When using traditional Likert coding, measurement invariance tests produced problematic results, because they incorrectly suggested a difference in factor means across language or experimental groups. Moreover, the invariance tests indicated that measurement invariance was established, thus encouraging misplaced confidence in the comparability of the data. Remarkably, this finding consistently occurred across a simulated cross-linguistic data set (Study 1), an empirical cross-linguistic dataset (Study 2), and data from a scale format experiment (Study 3). Our results show that calibrated sigma coding enhances the validity of measurement invariance tests. It is therefore a complement (not a substitute) for measurement

invariance testing. We specifically recommend that researchers, when working with grouped data that may be biased by differential scale usage, perform a sensitivity analysis by comparing findings based on traditional coding with findings based on calibrated sigma coding. In settings where there are true differences not due to differential scale usage, implementing the calibrated sigma coding method does not overcorrect (i.e., it does not make true differences disappear; see Study 1).

A requirement for the application of the calibrated sigma method is the availability of a set of items that are heterogeneous in content but homogeneous in form, particularly in terms of the labeling of the response categories. This may increase the costs of administering a survey. However, if data collection costs are an issue, the calibrated sigma codes can be based on data gathered in a pilot study (sampling from the same population of interest as in the main study), secondary data collected from a representative sample of respondents who completed a similar questionnaire, or a random subsample of respondents. To illustrate the last point, in Study 3 we showed how sigma values based on random subsamples (with sample sizes of approximately 110 per group) led to very similar results and the same substantive conclusions as when using the full sample (with sample sizes of approximately 220 per group). For market researchers working with cross-national consumer panels, it may be worthwhile to run a study to select and calibrate response option categories that can then be used in subsequent surveys, assuming that the pilot sample is representative of subsequent samples and that the data collection conditions are equivalent.

In the extant literature, several alternative methods have been proposed to deal with response bias across groups (e.g., nationalities, cultures, and languages). For a recent review of such methods, we refer to Baumgartner and Weijters (2015). What sets the calibrated sigma

method apart from most existing methods is that it involves a recoding operation at the group level. The disadvantage of this approach is that the method does not account for individual (within-group) differences in scale usage. The advantage is that it is particularly parsimonious and relatively easy to apply in cases where scale usage varies primarily between groups.

The most obvious domain of application of the method is in survey research involving multiple languages, where researchers are well aware of the issue of differential scale usage and where there is an active research tradition aiming to address this issue (de Langhe, et al., 2011; Weijters, et al., 2013; Weijters, Puntoni, & Baumgartner, in press). Here, the calibrated sigma method is especially useful when comparing Likert-type data in different languages, including samples that share the same nationality but use a different language, a common situation in marketing research (Holmqvist & Van Vaerenbergh, 2013; Van Vaerenbergh & Holmqvist, 2014; Weijters, et al., in press).

Despite a general lack of awareness of the matter, there are other domains where differential scale usage at the group level poses threats to validity. First, accumulating evidence shows that data collected by means of different modes of data collection may exhibit differential scale usage (Chang & Krosnick, 2009; Dillman, et al., 2009; Duffy, Smith, Terhanian, & Bremer, 2005; Fricker, Galesic, Tourangeau, & Yan, 2005; Roster, Rogers, Albaum, & Klein, 2004; Weijters, et al., 2008). For instance, it is not very surprising that telephone respondents who interact with an interviewer and verbally respond to auditory stimuli use scale points differently than online respondents who get to see a visual response scale in front of them to which they respond by clicking the appropriate response.

Second, the calibrated sigma method is applicable when comparing multi-category rating data obtained with different response scale formats. This may occur when comparing

secondary data from different surveys that use the same multi-item scale but use different response scale formats (Cabooter, Weijters, Geuens, & Vermeir, 2016).

Third, response scale usage can vary as a function of often manipulated variables such as self-regulatory focus or self-construal (Cabooter, Millet, Weijters, & Pandelaere, 2016; Lalwani, Shrum, & Chiu, 2009). Other experimental manipulations may also lead to a possibly unintended shift in response category usage (e.g., differences in ambient lighting, questionnaire readability, etc.). While it is not common to account for these differences, differential scale usage may present an alternative explanation for some results obtained in experimental priming studies. Although the issue is not top of mind at present, we believe it may gain importance in the future, given the increasing sophistication of measurement models in marketing research (Martínez-López, Gázquez-Abad, & Sousa, 2013).

Finally, the calibrated sigma method may also be applicable in other contexts where different groups of respondents tend to show different response patterns regardless of content, for instance, situations where native speakers versus non-native speakers may attach different meanings to the same response category labels used in the same language (de Langhe, et al., 2011; Harzing, 2006; Weijters, et al., in press).

The proposed approach assumes that the sigma values computed from the content-free control items (i.e., the items in the RPS scale) are valid indicators of differential scale usage and do not reflect substantive differences due to cultural or other group-specific characteristics. In other words, the control items should be free of common content within a given group (e.g., within a language group), but in addition they should only assess group-specific scale usage, not substantive differences between groups. Since the control items were chosen to be deliberately diverse in content, it seems highly unlikely that the consistency in responding

observed across the 16 RPS items in Study 1 (based on Cronbach's coefficient alpha, see Table 2) is due to substantive reasons. The more plausible explanation for this consistency is that different groups use the response scale in idiosyncratic ways. However, future research should investigate whether the RPS items (or other control items that researchers might use) are valid indicators of differential scale usage across even more culturally diverse groups.

It may be helpful to future researchers to have a set of response category labels and their corresponding calibrated sigma values available in different languages. As a first step, Table 3 provides the proportions and corresponding calibrated sigma values for the response category labels in the language groups used in Study 1. Opportunities for future research include studies that provide additional calibrated sigma values for other languages and/or other commonly used response category labels. The new method may also be relevant for other scale formats that do not assess agreement or disagreement (e.g., "very true" to "very false", "very much like me" to "not at all like me"), but it is important to point out that this would require the use of calibration items that use a similar format. Future research is necessary to further explore these possibilities.

Table 1
Proposed procedure for obtaining calibrated sigma values

			Response category				
Group	Step	Operation	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Group A	1	Mean frequency across 16 RPS items	0.800	3.200	6.400	3.200	2.400
	2	Mean proportion	0.050	0.200	0.400	0.200	0.150
	3	Cumulative proportion ($P_{k,g}$)	0.050	0.250	0.650	0.850	1.000
	4	[½ * ($P_{k,g}$ + $P_{k-1,g}$)]	0.025	0.150	0.450	0.750	0.925
	5	Sigma value	-1.960	-1.036	-0.126	0.674	1.440
Group B	1	Mean frequency across 16 RPS items	1.600	4.800	4.800	2.400	2.400
	2	Mean proportion	0.100	0.300	0.300	0.150	0.150
	3	Cumulative proportion ($P_{k,g}$)	0.100	0.400	0.700	0.850	1.000
	4	[½ * ($P_{k,g}$ + $P_{k-1,g}$)]	0.050	0.250	0.550	0.775	0.925
	5	Sigma value	-1.645	-0.674	0.126	0.755	1.440

Note: RPS = Response Pattern Scale (Greenleaf 1992b); k (1 to K) indexes response categories; g (1 to G) indexes groups or conditions.

Table 2
Internal consistency of the 16 RPS items at the group level in Study 1

	Category	Cronbach's alpha
1	Strongly disagree	0.84
2	Disagree	0.81
3	Slightly disagree	0.82
4	Neutral	0.93
5	Slightly agree	0.83
6	Agree	0.82
7	Strongly agree	0.89

Note: Table 2 displays the Cronbach's alpha coefficients for seven response categories in 12 language groups in Study 3 of Weijters, et al. (2013). At the respondent level, for each response category (1 to 7) and each RPS item (1 to 16), an indicator dummy variable was created signaling that the response category was selected (dummy = 1) or was not selected (dummy = 0). This resulted in 7 * 16 = 112 dummy variables at the respondent level. The data were then aggregated to the language-group level by averaging the dummies across respondents within each group. This resulted in a dataset where the language group is the unit of analysis (N = 12) and the 112 variables represent the proportion of respondents in a given language group who endorsed a given response category for a given RPS item. The latter dataset was used to compute seven Cronbach's alpha values (one for each response category proportion). For instance, the Cronbach's alpha for response category one ('strongly disagree' in English) quantifies the extent to which the 16 RPS proportions form an internally consistent scale of the likelihood of selecting response option one.

Table 3
Response distributions and sigma values for Study 1

		Category						
		1	2	3	4	5	6	7
Labels	Dutch	Volledig oneens	Oneens	Enigszins oneens	Neutraal	Enigszins eens	Eens	Volledig eens
	French	Pas du tout d'accord	Pas d'accord	Plutôt pas d'accord	Neutre	Plutôt d'accord	D'accord	Tout à fait d'accord
	English	Strongly disagree	Disagree	Slightly disagree	Neutral	Slightly agree	Agree	Strongly agree
	German	Überhaupt nicht einverstanden	Nicht einverstanden	Eher nicht einverstanden	Neutral	Einigermaßen einverstanden	Einverstanden	Vollkommen einverstanden
	Spanish	Muy en desacuerdo	En desacuerdo	Levemente en desacuerdo	Neutral	Algo de acuerdo	De acuerdo	Muy de acuerdo
	Polish	Zdecydowanie nie zgadzam się	Nie zgadzam się	Raczej się nie zgadzam	Obojętne	Raczej się zgadzam	Zgadzam się	Zdecydowanie zgadzam się
	Slovakian	Veľmi nesúhlasím	Nesúhlasím	Trochu nesúhlasím	Nezaujatý	Trochu súhlasím	Súhlasím	Veľmi súhlasím
	Hungarian	Egyáltalán nem értek egyet	Nem értek egyet	Nem teljesen értek egyet	Semleges	Valamennyire egyetértek	Egyetértek	Teljesen egyetértek
	Romanian	Nu sunt deloc de acord	Nu sunt de acord	Nu prea sunt de acord	Neutru	Sunt puţin de acord	Sunt de acord	Sunt complet de accord
	Swedish	Instämmer inte alls	Instämmer inte	Instämmer inte helt	Neutral	Instämmer något	Instämmer	Instämmer helt
	Italian	Non sono assolutamente d'accordo	Non sono d'accordo	Non sono in parte d'accordo	Indifferente	Sono in parte d'accordo	Sono d'accordo	Sono assolutamente d'accordo
	Turkish	Kesinlikle katılmıyorum	Katılmıyorum	Bazen katılmıyorum	Farketmez	Bazen katılıyorum	Katılıyorum	Kesinlikle katılıyorum

Table 3 (continued)

RPS %	Dutch	5.5%	12.9%	13.0%	23.0%	21.4%	17.7%	6.5%
	French	7.1%	8.0%	13.9%	20.1%	24.6%	14.9%	11.4%
	English	3.3%	8.8%	11.0%	21.9%	22.1%	22.6%	10.4%
	German	9.7%	12.0%	13.3%	22.4%	17.9%	16.9%	7.9%
	Spanish	3.7%	9.3%	10.8%	21.5%	22.4%	21.0%	11.5%
	Polish	6.3%	10.1%	13.6%	18.7%	21.2%	17.6%	12.6%
	Slovakian	4.7%	16.2%	11.4%	17.2%	21.5%	22.3%	6.6%
	Hungarian	6.7%	10.7%	12.5%	18.9%	20.0%	19.6%	11.6%
	Romanian	7.2%	9.7%	10.9%	17.4%	16.5%	21.9%	16.5%
	Swedish	7.6%	11.6%	11.4%	17.1%	20.4%	20.9%	11.0%
	Italian	5.2%	9.2%	9.4%	14.9%	25.7%	23.0%	12.6%
	Turkish	6.1%	11.0%	8.0%	9.5%	22.8%	23.3%	19.2%
Sigma	Dutch	-1.92	-1.18	-0.68	-0.18	0.39	1.02	1.85
	French	-1.81	-1.22	-0.77	-0.28	0.29	0.88	1.58
	English	-2.13	-1.43	-0.93	-0.41	0.15	0.78	1.63
	German	-1.66	-1.01	-0.57	-0.10	0.42	0.98	1.76
	Spanish	-2.09	-1.38	-0.90	-0.40	0.16	0.77	1.58
	Polish	-1.86	-1.21	-0.73	-0.27	0.23	0.79	1.53
	Slovakian	-1.98	-1.13	-0.62	-0.23	0.26	0.93	1.84
	Hungarian	-1.83	-1.17	-0.72	-0.27	0.22	0.79	1.57
	Romanian	-1.80	-1.17	-0.76	-0.35	0.09	0.60	1.39
	Swedish	-1.77	-1.11	-0.68	-0.28	0.20	0.79	1.60
	Italian	-1.94	-1.29	-0.87	-0.49	0.04	0.70	1.53
	Turkish	-1.87	-1.20	-0.80	-0.53	-0.10	0.50	1.30

Note: Endorsement proportions for the RPS in 12 languages were based on Weijters, et al. (2013).

Table 4
Model fit comparison for Study 1

Scenario	Coding	Model	χ^2	df	TLI	CFI	RMSEA	BIC
Equal latent means	Traditional	A. Unconstrained	.0	0	1.000	1.000	.000	117422.7
		B. Metric invariance	25.3	22	1.000	1.000	.012	117241.3
		C. Scalar invariance	51.2	44	1.000	1.000	.013	**117060.6**
		D. Means invariance	291.6	55	.990	.994	.066	117197.6
	Sigma	A. Unconstrained	.0	0	1.000	1.000	.000	73926.7
		B. Metric invariance	25.2	22	1.000	1.000	.012	73745.2
		C. Scalar invariance	47.0	44	1.000	1.000	.008	73560.4
		D. Means invariance	70.2	55	.999	1.000	.017	**73480.3**
Different latent means	Traditional	A. Unconstrained	.0	0	1.000	1.000	.000	116894.9
		B. Metric invariance	24.7	22	1.000	1.000	.011	116712.9
		C. Scalar invariance	49.4	44	1.000	1.000	.011	**116531.0**
		D. Means invariance	1009.8	55	.961	.974	.132	117388.0
	Sigma	A. Unconstrained	.0	0	1.000	1.000	.000	73665.7
		B. Metric invariance	24.1	22	1.000	1.000	.010	73483.2
		C. Scalar invariance	46.5	44	1.000	1.000	.008	**73299.0**
		D. Means invariance	954.0	55	.963	.976	.128	74103.1

Note: The lowest BIC value per data set is printed in boldface to indicate the preferred model.

Table 5
Response distributions and sigma values for Study 2

Language	Cat.	Label	Proportion	Sigma value
Dutch	1	Helemaal niet akkoord	.169	-1.374
	2	Eerder niet akkoord	.192	-.626
	3	Neutraal	.223	-.067
	4	Eerder akkoord	.231	.526
	5	Helemaal akkoord	.181	1.319
French	1	Pas du tout d'accord	.109	-1.603
	2	Pas vraiment d'accord	.162	-.877
	3	Neutre	.302	-.195
	4	Plutôt d'accord	.234	.498
	5	Tout à fait d'accord	.190	1.296

Table 6
Mean estimates for the French speaking sample in Study 2

		M	SE	t	p
Response styles	NARS	.962	.245	3.919	<.001
	ERS	-.287	.122	-2.359	.018
	MRS	.484	.095	5.111	<.001
Need for predictability	Traditional coding	.152	.055	2.769	.006
	Calibrated sigma coding	.062	.036	1.703	.088
	RIRSMACS corrected	-.082	.090	-.917	.359

Note: Table 6 reports the latent mean estimates for the French-speaking sample in Study 2. Since the mean for the Dutch-speaking group was constrained to zero, the t-tests and p-values can be used to evaluate the null hypothesis of mean equality.

Table 7
Model fit indices for Study 2

Coding	Model	χ^2	df	CFI	TLI	RMSEA	BIC
Traditional coding	A. Unconstrained	74.394	34	.962	.938	.066	13162.5
	B. Metric invariance	100.779	41	.944	.924	.074	13144.9
	C. Scalar invariance	122.188	48	.931	.919	.076	**13122.3**
	D. Means invariance	130.489	49	.924	.913	.079	13124.3
Calibrated sigma coding	A. Unconstrained	75.575	34	.961	.937	.067	9806.1
	B. Metric invariance	100.801	41	.945	.924	.074	9787.3
	C. Scalar invariance	118.750	48	.934	.924	.074	9761.2
	D. Means invariance	121.728	49	.933	.923	.074	**9757.9**
RIRSMACS	A. Unconstrained	309.988	172	.947	.917	.055	31592.6
	B. Metric invariance	337.196	179	.940	.908	.057	31575.8
	C. Scalar invariance	349.924	186	.937	.908	.057	31544.5
	D. Means invariance	350.985	187	.937	.909	.057	**31539.3**

Table 8
Different response scale formats get different sigma values (Study 3)

			Full sample		Random split half	
	Category	Category labels	RPS response %	Calibrated sigma value	RPS response %	Calibrated sigma value
Fully labeled	1	Extremely disagree	4.9%	-1.97	5.5%	-1.92
	2	Disagree	18.6%	-1.07	18.6%	-1.04
	3	Neither agree nor disagree	19.8%	-.43	20.4%	-.40
	4	Agree	40.7%	.35	40.5%	.38
	5	Extremely agree	16.0%	1.41	15.0%	1.44
Endpoint labeled	1	Disagree	10.0%	-1.64	9.5%	-1.67
	2		15.7%	-.92	16.0%	-.93
	3		23.5%	-.32	23.5%	-.32
	4		28.9%	.35	28.8%	.34
	5	Agree	21.8%	1.23	22.1%	1.22

Note: Sample sizes are $N_1 = 227$ and $N_2 = 228$ for the full sample and $N_1 = 114$, $N_2 = 118$ for the random split half sample in the fully labeled (N_1) and endpoint-labeled (N_2) conditions, respectively.

Table 9
Mean estimates for the fully labeled condition in Study 3

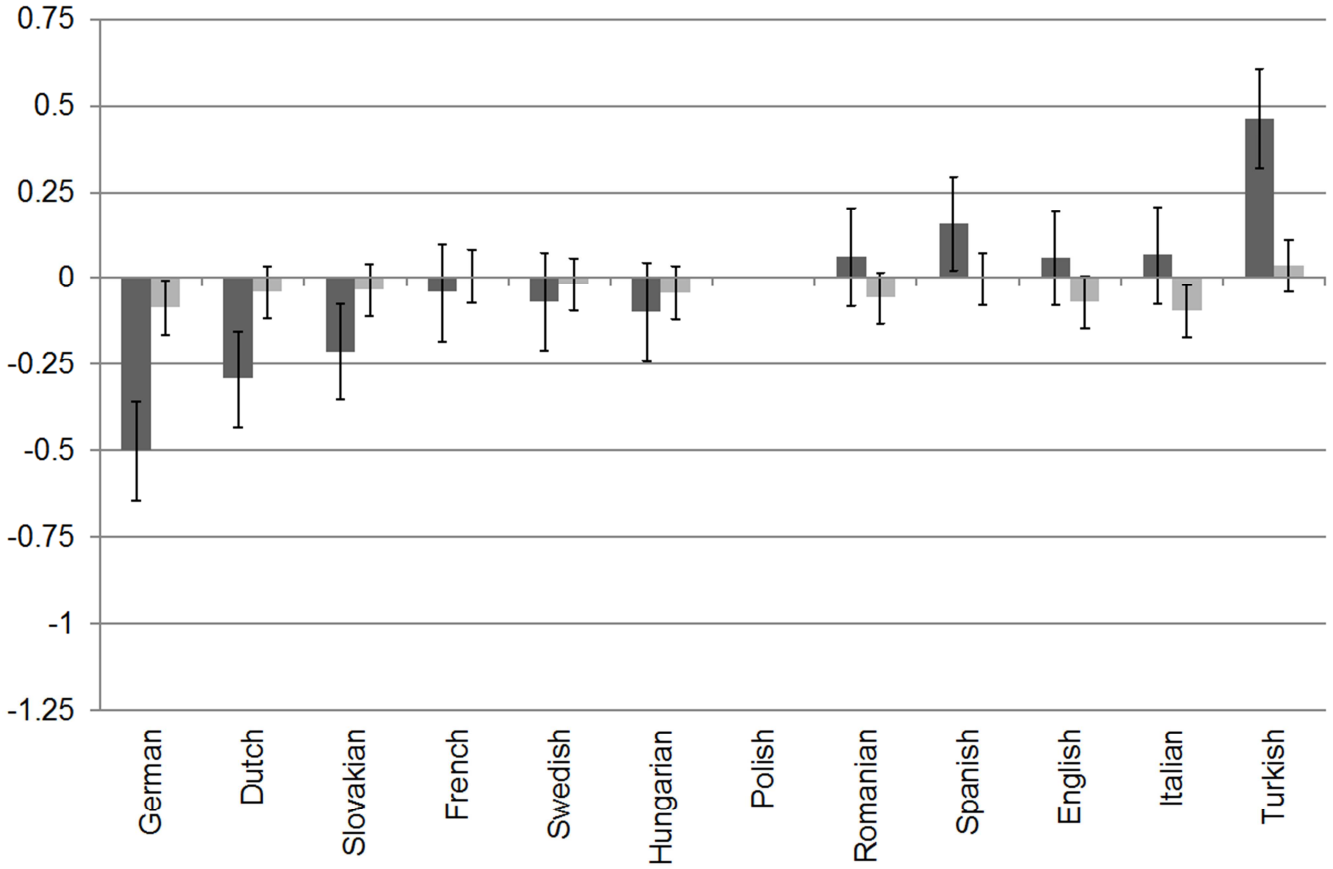
		M	SE	t	p
Response styles	NARS	-.070	.038	-1.818	.069
	ERS	.102	.019	5.322	< .001
	MRS	.036	.014	2.608	.009
Susceptibility to Normative Influence	Traditional coding	-.173	.064	-2.700	.007
	Calibrated sigma coding	.010	.045	.218	.828
	Random half calibrated sigma coding	-.030	.045	-.667	.505
	RIRSMACS corrected	-.074	.091	-.814	.416

Note: Table 9 reports the latent mean estimate for the fully labeled condition in Study 3. Since the mean for the endpoint-labeled group was constrained to zero, the t-test and p-value can be used to evaluate the null hypothesis of mean equality.

Table 10
Model fit indices for Study 3

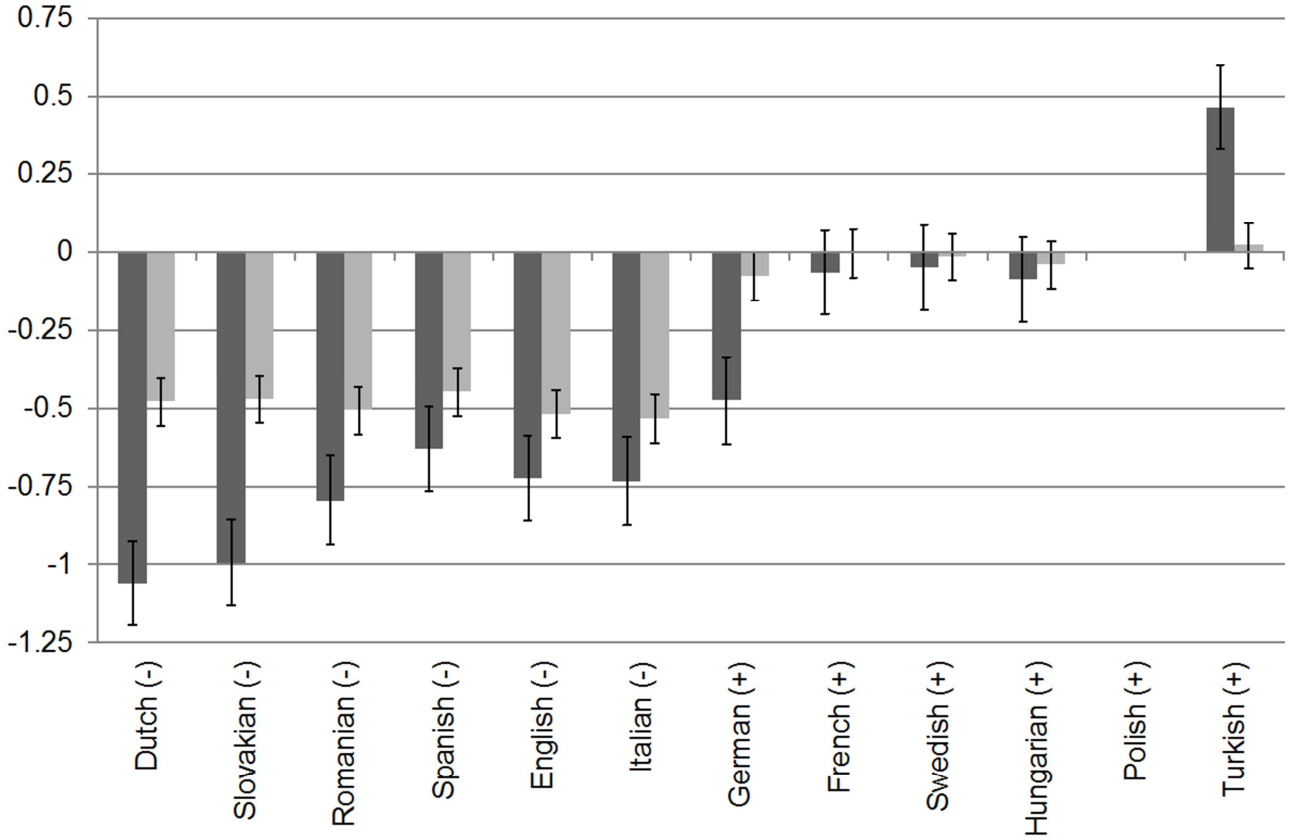
Approach	Model	χ^2	df	CFI	TLI	RMSEA	BIC
Traditional coding	A. Unconstrained	126.26	40	.971	.959	.097	8032.2
	B. Metric invariance	131.18	47	.971	.966	.089	7994.3
	C. Scalar invariance	139.75	54	.971	.970	.084	**7960.0**
	D. Means invariance	147.22	55	.969	.968	.086	7961.4
Calibrated sigma coding	A. Unconstrained	129.23	40	.970	.958	.099	5664.9
	B. Metric invariance	135.02	47	.971	.965	.091	5627.8
	C. Scalar invariance	139.68	54	.971	.970	.084	5589.6
	D. Means invariance	139.72	55	.972	.971	.082	**5583.6**
Split-half sigma coding	A. Unconstrained	128.67	40	.970	.959	.099	5676.5
	B. Metric invariance	134.29	47	.971	.965	.090	5639.3
	C. Scalar invariance	139.88	54	.971	.970	.084	5602.1
	D. Means invariance	140.32	55	.972	.971	.083	**5596.4**
RIRSMACS correction	A. Unconstrained	279.02	178	.975	.962	.050	9489.2
	B. Metric invariance	293.77	185	.973	.961	.051	9461.1
	C. Scalar invariance	299.47	192	.973	.962	.050	9424.0
	D. Means invariance	300.14	193	.974	.963	.049	**9418.5**

Figure 1: Corrected and uncorrected latent means (with 95% CI) when true means are equal across groups (Study 1)



Note for figure 1: Bars represent the estimated factor means with 95% CI's for traditionally coded (dark grey) vs. sigma coded (light grey) data. In the equal latent means scenario, the true latent means were equal across the twelve language groups. The Polish group served as the reference group and had the latent mean fixed to zero (with zero standard error). Language groups are shown in ascending order of RPS mean.

Figure 2: Corrected and uncorrected latent means (with 95% CI) when true means are different across groups (Study 1)



Note for figure 2: Bars represent the estimated factor means with 95% CI's for traditionally coded (dark grey) vs. sigma coded (light grey) data. In the different latent means scenario, the true latent means were -.25 for the language groups marked by a '-' (Dutch, English, Spanish, Slovakian, Romanian, and Italian) and .25 higher for the other groups marked with a '+'. The Polish group served as the reference group and had the latent mean fixed to zero (with zero standard error). Language groups are shown in ascending order of (1) true mean and (2) RPS mean within true mean condition.

References

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

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research, 38*(May), 143-156.

Baumgartner, H., & Weijters, B. (2015). Response Biases in Crosscultural Measurement. In S. Ng & A. Y. Lee (Eds.), *Handbook of Culture and Consumer Psychology* (pp. 370). Oxford Oxford University Press USA.

Bearden, W. O., Netemeyer, R. G., & Teel, J. E. (1989). Measurement of consumer susceptibility to interpersonal influence. *Journal of Consumer Research*, 473-481.

Bosma, N. S., & Levie, J. (2010). Global Entrepreneurship Monitor 2009 Executive Report.

Cabooter, E., Millet, K., Weijters, B., & Pandelaere, M. (2016). The 'I'in Extreme Responding. *Journal of Consumer Psychology, forthcoming*.

Cabooter, E., Weijters, B., Geuens, M., & Vermeir, I. (2016). Scale format effects on response option interpretation and use. *Journal of Business Research, 69*(7), 2574-2584.

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the internet comparing sample representativeness and response quality. *Public Opinion Quarterly, 73*(4), 641-678.

Clarke III, I. (2001). Extreme response style in cross-cultural research. *International Marketing Review, 18*(3), 301-324.

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

de Langhe, B., Puntoni, S., Fernandes, D., & van Osselaer, S. M. J. (2011). The Anchor Contraction Effect in International Marketing Research. *Journal of Marketing Research, 48*(2), 366-380.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research, 38*(1), 1-18.

Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2005). Comparing data from online and face-to-face surveys. *International journal of market research, 47*(6), 615.

Europe, E.-c. (2013). Europe B2C Ecommerce Report 2013. In: Brussels.

Fischer, R. (2004). Standardization to account for cross-cultural response bias - A classification of score adjustment procedures and review of research in JCCP. *Journal of Cross-Cultural Psychology, 35*(3), 263-282.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly, 69*(3), 370-392.

Gopinath, M., & Nyer, P. U. (2009). The effect of public commitment on resistance to persuasion: The influence of attitude certainty, issue importance, susceptibility to normative influence, preference for consistency and source proximity. *International Journal of Research in Marketing, 26*(1), 60-68.

Greenleaf, E. A. (1992a). Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles. *Journal of Marketing Research, 29*(2), 176-188.

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

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

Hofstede, G. H. (1991). *Cultures and organisations-software of the mind: intercultural cooperation and its importance for survival*: McGraw-Hill.

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*: Sage.

Holmqvist, J., & Van Vaerenbergh, Y. (2013). Perceived importance of native language use in service encounters. *The Service Industries Journal, 33*(15-16), 1659-1671.

Jordan, L. A., Marcus, A. C., & Reeder, L. G. (1980). Response styles in telephone and household interviewing: A field experiment. *Public Opinion Quarterly, 44*(2), 210-222.

Kruglanski, A. W., Atash, M., DeGrada, E., Mannetti, L., Pierro, A., & Webster, D. M. (1997). Psychological theory testing versus psychometric nay-saying: Comment on Neuberg et al.'s (1997) critique of the Need for Closure Scale.

Kruglanski, A. W., Webster, D. M., & Klem, A. (1993). Motivated resistance and openness to persuasion in the presence or absence of prior information. *Journal of Personality and Social Psychology, 65*(5), 861.

Lalwani, A. K., Shrum, L., & Chiu, C.-Y. (2009). Motivated response styles: the role of cultural values, regulatory focus, and self-consciousness in socially desirable responding. *Journal of Personality and Social Psychology, 96*(4), 870.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*(1), 53-76.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*(4), 810.

Martínez-López, F. J., Gázquez-Abad, J. C., & Sousa, C. M. (2013). Structural equation modelling in marketing and business research: Critical issues and practical recommendations. *European Journal of Marketing, 47*(1/2), 115-152.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*: CRC Press.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539-569.

Richter, L., & Kruglanski, A. W. (2004). Motivated closed mindedness and the emergence of culture. *The psychological foundations of culture*, 101-121.

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

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach. *Journal of the American Statistical Association, 96*(March), 20-31.

Roster, C. A., Rogers, R. D., Albaum, G., & Klein, D. (2004). A comparison of response characteristics from web and telephone surveys. *INTERNATIONAL JOURNAL OF MARKET RESEARCH., 46*, 359-374.

Shane, S. (1993). Cultural influences on national rates of innovation. *Journal of business venturing, 8*(1), 59-73.

Skevington, S. M., & Tucker, C. (1999). Designing response scales for cross‐cultural use in health care: Data from the development of the UK WHOQOL. *British journal of medical psychology*, *72*(1), 51-61.

Sleuwaegen, L., & Buysse, R. (2010). De contextuele determinanten van het ondernemerschap in Vlaanderen. In: Katholieke Universiteit Leuven.

Smith, T. W., Mohler, P. P., Harkness, J., & Onodera, N. (2005). Methods for assessing and calibrating response scales across countries and languages. *Comparative sociology, 4*(3), 365.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research, 25*(June), 78-90.

Szabo, S., Orley, J., & Saxena, S. (1997). An approach to response scale development for cross-cultural questionnaires. *European Psychologist, 2*(3), 270.

Van Rosmalen, J., Van Herk, H., & Groenen, P. J. F. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research, 47*(1), 157-172.

Van Vaerenbergh, Y., & Holmqvist, J. (2014). Examining the relationship between language divergence and word-of-mouth intentions. *Journal of Business Research, 67*(8), 1601-1608.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67*(6), 1049.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The Effect of Rating Scale Format on Response Styles: The Number of Response Categories and Response Category Labels. *International Journal of Research in Marketing, 27,* 236-247.

Weijters, B., Geuens, M., & Baumgartner, H. (2013). The Effect of Familiarity with the Response Category Labels on Item Response to Likert Scales. *Journal of Consumer Research, 40*(2), 368-381.

Weijters, B., Puntoni, S., & Baumgartner, H. (in press). Methodological issues in cross-linguistic and multilingual advertising research. *Journal of Advertising*.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36*(3), 409–422.

Wicherts, J. M., & Dolan, C. V. (2004). A cautionary note on the use of information fit indexes in covariance structure modeling with means. *Structural Equation Modeling, 11*(1), 45-50.