British Journal of Mathematical and Statistical Psychology (2019), 72, 401–425

The British Psychological Society

www.wileyonlinelibrary.com

Special Issue Paper

Clustering preference data in the presence of response-style bias

Mariko Takagishi[1]*, Michel van de Velden[2] and Hiroshi Yadohisa[3]

[1]Graduate School of Engineering Science, Osaka University, Japan
[2]Econometric Institute, Erasmus University Rotterdam, The Netherlands
[3]Facluty of Culture and Information Science, Doshisha University, Japan

Preference data, such as Likert scale data, are often obtained in questionnaire-based surveys. Clustering respondents based on survey items is useful for discovering latent structures. However, cluster analysis of preference data may be affected by response styles, that is, a respondent's systematic response tendencies irrespective of the item content. For example, some respondents may tend to select ratings at the ends of the scale, which is called an 'extreme response style'. A cluster of respondents with an extreme response style can be mistakenly identified as a content-based cluster. To address this problem, we propose a novel method of clustering respondents based on their indicated preferences for a set of items while correcting for response-style bias. We first introduce a new framework to detect, and correct for, response styles by generalizing the definition of response styles used in constrained dual scaling. We then simultaneously correct for response styles and perform a cluster analysis based on the corrected preference data. A simulation study shows that the proposed method yields better clustering accuracy than the existing methods do. We apply the method to empirical data from four different countries concerning social values.

1. Introduction

In cluster analysis, respondents are allocated to groups of similar observations (MacQueen, 1967). In many applications, respondents are clustered based on their preferences with respect to a set of items. To measure such preferences, questionnaires are frequently used in which respondents indicate their preference using a rating scale, such as a Likert scale, where respondents make selections from a set of predetermined preference categories. Clustering respondents relative to their answers may be useful to identify latent clustering structures.

Questionnaire-based preference data may be affected by so-called 'response styles'. A response style can be defined as a systematic response tendency irrespective of item content (Baumgartner & Steenkamp, 2001). Examples of response styles include an extreme response style, a tendency to select only categories at the ends of the scale; and a midpoint response style, a tendency to only select the middle of the scale. In this paper, we refer to data in which observations are affected by response styles as 'response-style-biased data'.

*Correspondence should be addressed to Mariko Takagishi, 1-3, Machikaneyama, Toyonaka, Osaka 560-8531, Japan (email: m.takagishi0728@gmail.com).

DOI:10.1111/bmsp.12170

Response styles are related to various factors, including culture (Cheung & Rensvold, 2000; Meisenberg & Williams, 2008), education (Meisenberg & Williams, 2008), gender (Austin, Deary, & Egan, 2006; Weijters, Geuens, & Schillewaert, 2010), and age (Stukovsý, Palat, & Sedlakova, 1982). In cross-cultural surveys, typically several of the above-mentioned factors are present and response-style bias is considered particularly significant (Baumgartner & Steenkamp, 2001). Moors (2012) and Cheung and Rensvold (2000) showed that response styles can lead to incorrect conclusions. Biases due to response styles can be considered as 'systematic error', rather than 'random error' (Baumgartner & Steenkamp, 2001). Therefore, to perform a meaningful data analysis, such systematic errors must be considered.

In practice, if data contain response-style bias, cluster analysis may yield clusters of respondents with similar response styles (response-style-based clusters), rather than clusters with similar item preferences (content-based clusters). For example, assume that in a survey one group of respondents tends to select midpoint categories, while another group tends to favour endpoint categories, regardless of their preferences. Applying cluster analysis to the resulting data may extract clusters of respondents who have selected midpoint and endpoint categories. However, these clusters only reflect their response styles and any content-based structure in the data remains undetected.

Several methods have been proposed to detect and control for response-style bias (Javaras & Ripley, 2007; Johnson, 2003). Especially in the item response theory (IRT) framework, B¨ockenholt and Meiser (2017) reviewed two types of IRT models designed to handle response styles: threshold-based models such as polytomous Rasch models and their mixture extensions (von Davier & Yamamoto, 2007; Rost, 1991), and an item response (IR) tree model (Böckenholt, 2012, 2017), which can be used to distinguish the effects of the judgement processes associated with content and response style. Plieninger and Meiser (2014) also validated several IR tree methods using an empirical data set. In other IRT-related research involving response styles, IRT and mixture IRT models have further been applied to correct for response style by adjusting parameters representing the response styles (Austin *et al.*, 2006; Bolt & Johnson, 2009; Meiser & Machunsky, 2008; Morren, Gelissen, & Vermunt, 2012). Typically, however, these methods only allow for the presence of relatively few (e.g., two) types of response styles. In addition, the estimation of such IRT models generally requires relatively large sample sizes (e.g., Finch & French, 2012, p. 177).

An alternative approach for handling and correcting for response styles was proposed by Rosmalen, Van Herk, and Groenen (2010). The primary objective of their latent-class bilinear multinomial logit model, however, was to investigate how response style and item content (and background variables, if relevant) affect responses in a low-dimensional space.

Furthermore, Schoonees, Velden, and Groenen (2015) proposed constrained dual scaling (CDS), which was designed to detect several, typically more than two, response styles and, compared to other studies, focuses more on correcting the response-style bias. While other probabilistic models control for response styles by adjusting parameters related to the probabilities for selecting specific ratings, in CDS the correction is done by transforming the original value. In this paper, we consider the application of *k*-means cluster analysis to CDS-corrected data and refer to this as 'CDS tandem analysis'.

Constrained dual scaling is an extension of dual scaling for preference data (Nishisato, 1980), which involves dimension reduction. Specifically, Schoonees *et al.* (2015) formulated a CDS approach yielding parameters that can be interpreted as response

styles. To estimate the parameters in CDS, dimension reduction is applied. In particular, a one-dimensional solution is required to estimate the response styles. However, the use of dimension reduction implies a loss of respondent-specific information that may complicate the retrieval of accurate content-based clusters. In other words, CDS can remove respondents' differences that may be useful for content-based clustering.

To address these problems, we propose a new method for correcting and clustering response-style-biased data. Throughout this paper, we refer to our new method as CCRS. To achieve our objective, we first focus on correction of response styles, and introduce a framework to detect, and correct for, response styles by generalizing the definition of response styles used in CDS. In this way, we obtain a new correction method that does not require dimension reduction and that includes CDS as a special case. Next, we consider content-based clustering of the corrected data. However, rather than performing these steps sequentially, we propose to simultaneously correct for respondent-specific response styles and apply content-based clustering to the corrected data. By this simultaneous approach, we avoid a potential problem associated with the CDS tandem analysis, where the response style correction removes information relevant to the content-based cluster analysis. Note that, although in this paper we only consider content-based clustering, our new correction method can be used in combination with other data analysis methods as well.

The remainder of this paper is organized as follows. In Section 2 we introduce a novel framework to identify and correct for response style, by formalizing the concept of response functions. In Section 3 we describe the proposed correction method within the framework introduced in Section 2 and then introduce the new CCRS method. In Section 4 we briefly describe CDS and show that the CDS correction method can be considered a special case in the proposed framework. Finally, the performance of the proposed method is evaluated and compared to those of existing methods in a simulation study and an empirical example in Sections 5 and 6, respectively.

2. Formalizing response functions

To describe the proposed methodology, a new framework is first introduced to formalize the concept of a 'response function'. Response styles and corrected values are defined more rigorously here than in previous studies by van de Velden (2008) and Schoonees *et al.* (2015). This framework can be used more generally when dealing with preference data possibly contaminated by response-style effects. The relationship between our framework and CDS is elaborated on in Section 4.

2.1. Category boundaries in preference data

Response-style problems occur when the interpretation of the preference categories differs for different respondents. For example, with five-point scale data, if a respondent has an acquiescence response style, that is, a tendency to select mostly the highest categories regardless of item content, the third category indicates a low preference of the respondent for that item, even though that category is the midpoint of the scale.

To express this formally, let $x_{ij} \in \{1, \ldots, q\}$ denote the q-point scale preference data provided by the ith respondent for the jth item ($i = 1, \ldots, n;\ j = 1, \ldots, m$). Suppose the observed preference data x_{ij} are related to the true preference data $x^*_{ij} \in \mathbb{R}$ as follows:

$$x_{ij} = \sum_{\ell=1}^{q} \ell I\{\eta_{i(\ell-1)} < x_{ij}^* \leq \eta_{i\ell}\},$$

where $I\{\cdot\}$ is an indicator function and $\eta_{i\ell}(\ell = 0, \ldots, q)$ are respondent-specific boundaries. We refer to the set of boundaries b_ℓ $(\ell = 0, \ldots, q)$, which are equal for all respondents and are spaced equally, as reference boundaries. In this paper, we consider a bounded interval, that is, $\eta_{i0} = b_0 = L$ and $\eta_{iq} = b_q = U$.

With this notation, 'response-style-biased data' are data for which the true preferences x_{ij}^* are categorized based on equally spaced reference boundaries b_ℓ even though each respondent has respondent-specific boundaries $\eta_{i\ell}$. This process is illustrated in Figure 1.

In Figure 1, respondent i has true preference x_{ij}^* and boundaries $\eta_{i\ell}(\ell = 1, \ldots, q)$ as shown on the upper scale. The aim is to 'estimate' x_{ij}^* from x_{ij}. In this example, the observed preference is $x_{ij} = 3$. If we ignore the possibility that each respondent has different boundaries and simply assume that the reference boundaries are used as shown on the lower scale in Figure 1, a rough estimation of x_{ij}^*, say $\hat{x}_{ij}^*$, would be far from the true one, x_{ij}^*. This indicates that, depending on the unobservable respondent-specific boundaries, we obtain a bias from the true x_{ij}^*.

2.2. Response functions

To correct for response-style-biased data, we define a response function as follows.

Definition 2.1. *(Response function). Suppose reference boundaries b_ℓ $(\ell = 1, \ldots, q-1)$ and respondent-specific boundaries $\eta_{i\ell}$ $(\ell = 1, \ldots, q-1)$ are given. Let both boundaries be monotonically increasing for ℓ. Then*

$$\phi_i : b_\ell \longmapsto \eta_{i\ell} \quad (\ell = 1, \ldots, q-1)$$

is defined as the response function for respondent i.

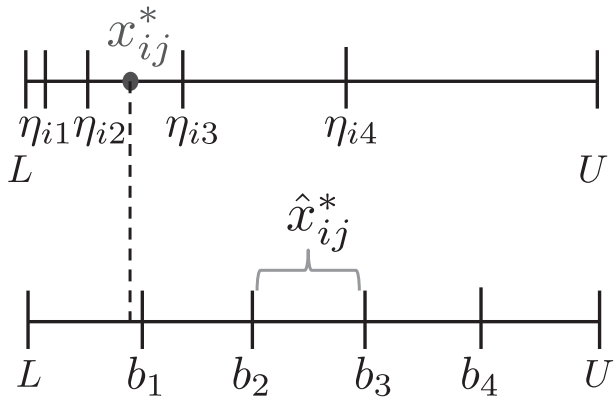

Figure 1. Response-style bias. On the upper scale [L, U] $\in \mathbb{R}$, respondent-specific boundaries are shown, while on the lower scale (equal-spaced) reference boundaries are shown. x_{ij}^* indicates the true preference, and $x_{ij}^* \in$ (b2, b3] is the estimation of x_{ij}^* on a scale with reference boundaries b_ℓ, when $x_{ij} = 3$ is obtained. The set of $\eta_{i\ell}$ $(\ell = 1, \ldots, q-1)$ on the upper scale represents a response style in which the fourth and fifth categories are more likely to be selected. [Colour figure can be viewed at wileyonlinelibrary.com]

From this definition, it follows that ϕ_i is a monotonically increasing function. In addition, we assume that the response function is continuous. For later purposes, it is useful to specifically define the response function corresponding to the absence of a response style.

Definition 2.2. *(No response style). If* $\eta_{i\ell} = b_\ell$ *(*$\ell = 1, \ldots, q-1$*), we say that respondent i has no response style.*

If ϕ_i is known for all respondents, we can use it to correct response-style-biased data, and to interpret respondents' response styles.

Definition 2.3. *(Correcting preference data using the response functions). Given q-point scale preference data* x_{ij} *with reference boundaries* $b_1, \ldots, b_{q-1}$*, and response functions* ϕ_i*, when* $x_{ij} = \ell$*, the corrected value of* x_{ij} *is*

$$y_{ij} = \phi_i(\tau(\ell)), \quad \text{where } \tau(\ell) \in (b_{\ell-1}, b_\ell].$$

This definition indicates that the corrected value of x_{ij} is defined as the product of the transformation of some value between $b_{\ell-1}$ and b_ℓ, $\tau(\ell)$, according to ϕ_i. In this paper, as in CDS, we fix $\tau(\ell) = (b_\ell + b_{\ell-1})/2$.

Figure 2 illustrates how a response function can be used to correct for response-style bias. Suppose that we want to know x_{ij}^* when the observed rating is $x_{ij} = 6$ on a seven-point scale. In this case, the argument of ϕ_i can be any value in the interval $(b_5, b_6]$. Following Definition 2.3, we use the midpoint of the interval, and call it the representative value of category 6. If we set $b_\ell = \ell$ $(\ell = 1, \ldots, q-1)$, 5.5 (i.e., the point on the horizontal axis in Figure 2) will be the argument of ϕ_i. Assuming that the true response function is

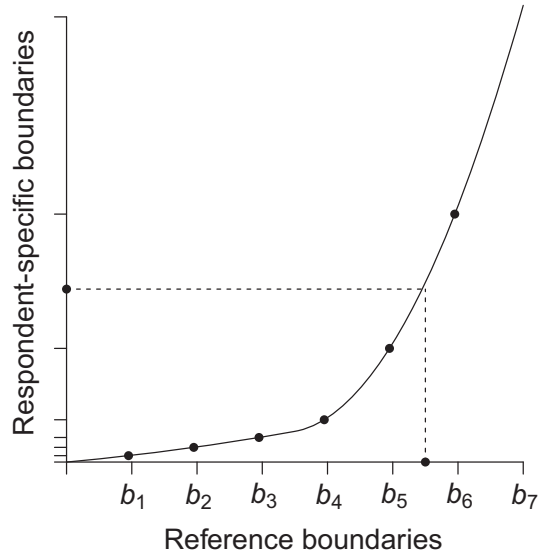

Figure 2. Example depicting how the observed value, $x_{ij} = 6$, corresponds to the corrected value. The solid line indicates the response function, ϕ_i. The horizontal axis represents the reference boundary (scale), while the vertical axis represents the respondent-specific boundary.

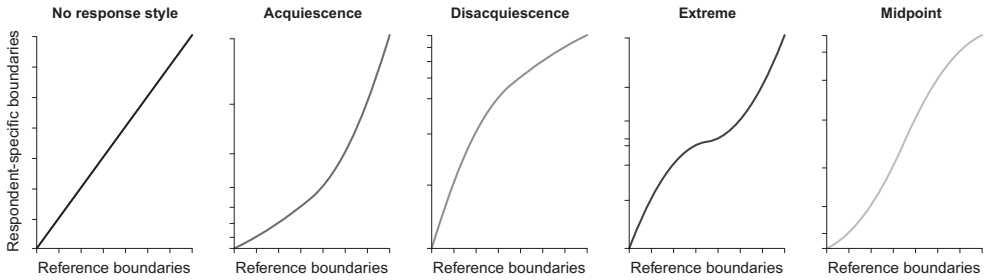

Figure 3. Response functions. The horizontal axis represents the reference boundary (scale), while the vertical axis represents the respondent-specific boundary. [Colour figure can be viewed at wileyonlinelibrary.com]

continuous, the output value of the response function corresponding to the representative value of the category (i.e., the point on the vertical axis in Figure 2), can be read (i.e., interpolated) off the vertical axis. The resulting value, y_{ij} in this case, is the corrected value.

Response functions can be used to interpret the respondents' response styles. Figure 3 shows examples of typical response functions corresponding to respondents who have no, acquiescence, disacquiescence (a tendency to disagree), midpoint, or extreme response styles.

3. Correcting and clustering preference data in the presence of response-style bias

Based on the ideas and definitions introduced in Section 2, we consider estimation of respondent-specific response functions. Moreover, we show that the estimated response functions can be used to correct for response-style bias and, at the same time, to find clusters of respondents based on their corrected item preferences.

3.1. Modelling response functions

To estimate a response function, data that represent respondent-specific boundaries are required. Here, similar to dual scaling and CDS, we code the preference data as 'rank-ordered boundary data'. This means that the indicated item preferences and the reference boundaries are converted to rank-orders for each respondent. The boundary rankings obtained reflect respondents' tendencies to select certain rating categories.

Suppose that q-point scale preference data $\mathbf{X} = (x_{ij})(i = 1, \ldots, n;\ j = 1, \ldots, m)$ are given with the reference boundaries $b_1, b_2, \ldots, b_{q-1}$. Then the rank-ordered boundary data $f_{i\ell}$ $(\ell = 1, \ldots, q-1)$ can be obtained as

$$f_{i\ell} = \sum_{t=1}^{m+q-1} \left(I\{c_{it} < b_\ell\} + \frac{1}{2} I\{c_{it} = b_\ell\} \right) - \frac{1}{2}$$
$$\text{where } c_{it} = \begin{cases} \frac{b_\ell + b_{\ell-1}}{2} & (t = 1, \ldots, m, x_{it} = \ell) \\ b_{t-q+1} & (t = m+1, \ldots, m+q-1). \end{cases} \tag{1}$$

For $t = 1, \ldots, m$, c_{it} indicates the representative values of a category, in our case, $(b_\ell + b_{\ell-1})/2$. On the other hand, for $t = m+1, \ldots, m+q-1$, c_{it} indicates reference boundaries.

To illustrate how this works in practice, consider seven-point scale preference data, $\mathbf{x}_i = (5, 6, 7)$, as given. Using equation (1), we obtain $\mathbf{c}_i = (4.5, 5.5, 6.5, 1, 2, 3, 4, 5, 6)$, where $\mathbf{c}_i = (c_{it})$ $(t = 1, \ldots, m+q-1)$. Then, sorting and converting these to rank-orders (starting from 0) yields

$$\begin{array}{lccccccccc} \mathbf{c}_i^{\text{sorted}} = (1 & 2 & 3 & 4 & 4.5 & 5 & 5.5 & 6 & 6.5) \\ \text{rank} : & \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} & 4 & \mathbf{5} & \mathbf{6} & \mathbf{7} & 8 \end{array}$$

Since $c_{i4} = 1, c_{i5} = 2, c_{i6} = 3, \ldots, c_{i9} = 6$ correspond to ranks 0, 1, 2, 3, 5, 7, respectively, we get $\mathbf{f}_i = (0,1, 2, 3, 5, 7)$. Figure 4 (left) plots these reference boundaries against the $\mathbf{f}_i = (f_{i\ell})$ $(\ell = 1, \ldots, 6)$. Using this converted $\mathbf{f}_i$, we see that respondent i demonstrates an acquiescence response style. For example, for $f_{i1}, \ldots, f_{i4}$, the values increase one by one, which indicates that respondent i does not frequently select categories between the first and fourth reference boundaries (i.e., the respondent does not often assign a rating smaller than 4). On the other hand, there is a large gap between f_{i4} and f_{i6}, which indicates that categories between the fourth and sixth reference boundaries are often selected.

Using $f_{i\ell}$, we consider a model for response functions corresponding to Definition 2.1, using I-spline basis functions. Let $\bar{f}_{i\ell} = f_{i\ell}/p$, where $p = m+q-1$, so that $\bar{f}_{i\ell} \in [0, 1]$. Also, from here on, we use $b_\ell = \ell/q$ $(\ell = 1, \ldots, q-1)$. In CCRS, $\bar{f}_{i\ell}$ is approximated as

$$\begin{aligned} \bar{f}_{i\ell} &\approx \phi_i^{\text{CCRS}}(\ell/q) \quad (i = 1, \ldots, n;\ \ell = 1, \ldots, q-1), \\ &\text{where } \phi_i^{\text{CCRS}}(x) = \sum_{r=1}^{3} \beta_{ir} M_r(x) \\ &\text{s.t.} \sum_{r=1}^{3} \beta_{ir} = 1, \quad \beta_{ir} \geq 0 \quad (r = 1, 2, 3), \end{aligned} \tag{2}$$

$$M_1(x) = \begin{cases} \frac{2t(x-L)-(x^2-L^2)}{(t-L)^2} & (L \leq x < t) \\ 1 & (t \leq x \leq U), \end{cases}$$

$$M_2(x) = \begin{cases} \frac{(x-L)^2}{(t-L)(U-L)} & (L \leq x < t) \\ \frac{t-L}{U-L} + \frac{2U(x-L)-(x^2-t^2)}{(U-t)(U-L)} & (t \leq x \leq U), \end{cases}$$

$$M_3(x) = \begin{cases} 0 & (L \leq x < t) \\ \frac{(x-t)^2}{(U-t)^2} & (t \leq x \leq U) \end{cases}$$

and $x \in [L, U]$, $t = L + 0.5(U-L)$. Here M_r $(r = 1, 2, 3)$ are I-spline basis functions, and β_{i1}, β_{i2} and β_{i3} are the coefficients of M_1, M_2 and M_3, respectively (see, for example, Ramsay, 1988). In CCRS, we use $L = 0$, $U = 1$. Non-negative conditions, $\beta_{ir} \geq 0$ $(r = 1,2,3)$, are required for ϕ_i to be a monotone increasing function. See Appendix S1 in the Supporting information for a more detailed justification of the rationale underlying

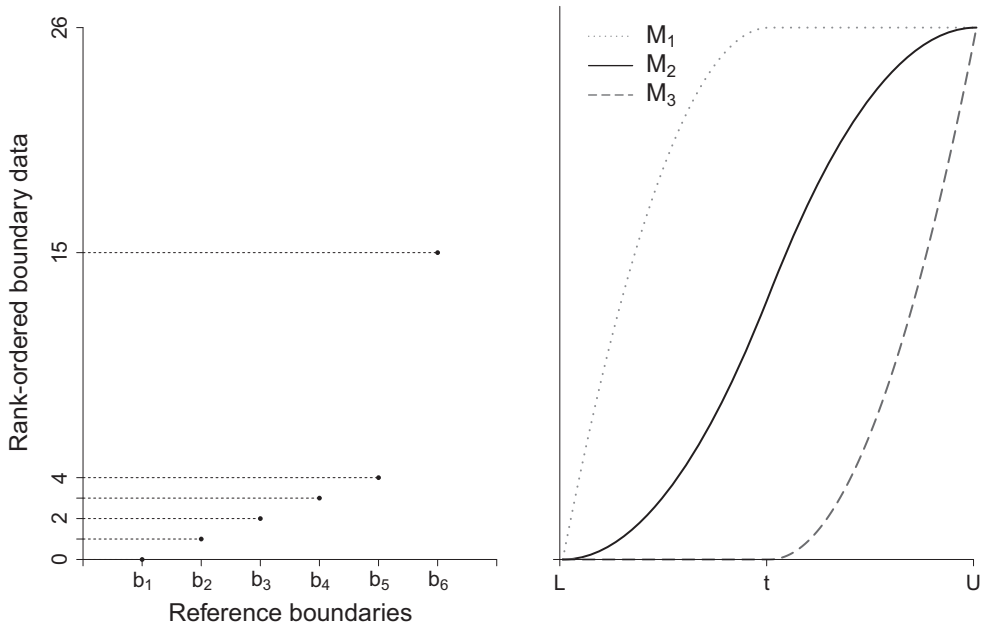

Figure 4. (Left) An example of rank-ordered boundary data. The horizontal axis corresponds to reference boundaries, while the vertical axis shows f_{il} values corresponding to each boundary. Each dot represents $f_{i1}, \ldots, f_{i6}$. (Right) Three I-spline basis functions. $M1$, $M2$, $M3$ are shown as solid, dotted and dashed lines, respectively. [Colour figure can be viewed at wileyonlinelibrary.com]

the scaling of $[L, U]$, $f_{i\ell}$ and b_ℓ to [0, 1] as well as the advantages of adding the constraint $\sum_{r=1}^{3} \beta_{ir} = 1$.

By using three I-spline basis functions (as shown in Figure 4 (right)), we can handle the five types of response style shown in Figure 3. Further, in this model, only β_{i1}, β_{i2} and β_{i3} need to be considered to interpret the response styles. For example, a greater β_{i3} value indicates a stronger tendency to select high categories because it results in more weight being placed on M_3, which alters the shape of function to be more similar to the shape of the response function corresponding to the acquiescence response style (shown in Figure 3).

Now we can define a new correction method. Using the model defined in equation (2), the response function can be estimated by 'smoothing' via the constrained least squares method. In other words, given a $q \times 1$ vector $\bar{\mathbf{f}}_i = (\bar{f}_{i\ell})$ and a $(q-1) \times 3$ matrix $\mathbf{M} = (M_r(\ell/q))$ $(\ell = 1, \ldots, q-1;\ r = 1, 2, 3)$, $\boldsymbol{\beta}_i$ is obtained by minimizing

$$\sum_{i=1}^{n} ||\bar{\mathbf{f}}_i - \mathbf{M}\boldsymbol{\beta}_i||^2, \quad \text{s.t.} \sum_{r=1}^{3} \beta_{ir} = 1, \quad \beta_{ir} \geq 0,$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \beta_{i3})$. Using the estimated value of $\hat{\boldsymbol{\beta}}_i$, we can construct the 'estimated' response function (see Definition 2.3), $\hat{\phi}(x) = \sum_{r=1}^{3} \hat{\beta}_{ir} M_r(x)$. By transforming all responses in the preference data $\mathbf{X}$ using $\hat{\phi}(x)$, we obtain an $(n \times m)$ 'corrected data' matrix, where response-style bias is removed. Note that our new correction method can be considered as a special case of the framework introduced in Section 2.

In order to cluster respondents based on content in corrected data matrix, content-based clustering, such as *k*-means clustering, can be applied to the corrected data. We shall refer to this type of analysis as CCRS tandem.

Sequentially applying two methods (smoothing and clustering) may not yield optimal results for the correction and content-based clustering as the criteria of correction and clustering are optimized separately (e.g., Arabie & Hubert, 1994). Therefore, we propose a method to conduct these two procedures simultaneously.

3.2. CCRS: Correcting and clustering response-style-biased data

Simultaneous smoothing and clustering can be achieved by simply adding the two minimization criteria (e.g., Hwang, Dillon, & Takane, 2006). Let K be the number of content-based clusters. Then we define the objective function of CCRS as follows:

$$L(\mathbf{B},\mathbf{G},\mathbf{U}|\bar{\mathbf{F}},\mathbf{Z},M_1,M_2,M_3) = \lambda\sum_{i=1}^{n}||\bar{\mathbf{f}}_i - \mathbf{M}\boldsymbol{\beta}_i||^2 + (1-\lambda)\sum_{i=1}^{n}\sum_{k=1}^{K}u_{ik}||\mathbf{Z}_i\widetilde{\mathbf{M}}\boldsymbol{\beta}_i - \mathbf{g}_k||^2$$
$$\text{s.t.}\sum_{r=1}^{3}\beta_{ir} = 1, \quad \beta_{ir}\geq 0 \quad (r=1,2,3;\ i=1,\ldots,n), \tag{3}$$

where $\mathbf{B}=(\boldsymbol{\beta}_i)$, $\mathbf{G}=(\mathbf{g}_k)$, $\mathbf{U}=(u_{ik})$, $\mathbf{F}=(\mathbf{f}_i)$ $(i=1,\ldots,n;\ k=1,\ldots,K)$, and $\mathbf{Z}=(\mathbf{Z}_i)$, $\mathbf{Z}_i=(z_{ij\ell})$ $(j=1,\ldots,m;\ \ell=1,\ldots,q)$. The first term in equation (3) is the smoothing term, and the second term is the content-based clustering term. Note that $\lambda\in[0,1]$ weights these two terms and needs to be determined prior to the analysis.

In the content-based clustering term, *k*-means clustering is performed on the corrected data, namely, $\mathbf{Z}_i\widetilde{\mathbf{M}}\boldsymbol{\beta}_i=(\hat{y}_{ij})$ $(i=1,\ldots,n;j=1,\ldots,m)$. Specifically, the $q\times 1$ vector $\mathbf{z}_{ij}=(z_{ij\ell})$ $(\ell=1,\ldots,q)$ is a dummy vector that takes $z_{ij\ell}=1$ if respondent i selects category ℓ for the jth item; otherwise, $z_{ij\ell}=0$. The $q\times 3$ matrix $\widetilde{\mathbf{M}}=(M_r(\tau(\ell)))$ $(\ell=1,\ldots,q;\ r=1,2,3)$ is a basis function matrix; however, unlike $\mathbf{M}$, it takes the midpoints of the boundaries as arguments to construct the corrected data in Definition 2.3. The $K\times 1$ vector $\mathbf{u}_i=(u_{ik})$ $(k=1,\ldots,K)$ is an indicator vector for the content-based cluster, where $u_{ik}=1$ if respondent i belongs to the kth content-based cluster; otherwise, $u_{ik}=0$. $\mathbf{G}$ is the $K\times m$ content-based cluster centroid matrix.

Choosing an appropriate value for λ is a complicated task as there is no clear criterion that can be used. In Section 5 we show how different values of λ affect the clustering results, and in Section 6 we propose a pragmatic approach to determine λ and K at the same time.

3.3. CCRS parameter estimation

To obtain parameters $\mathbf{B},\mathbf{G},\mathbf{U}$, two operations – estimation of the response functions (estimation of $\mathbf{B}$) and content-based clustering (estimation of $\mathbf{G}$ and $\mathbf{U}$) – are performed sequentially. For fixed $\mathbf{B}$, minimizing equation (3) reduces to *k*-means clustering of the (response-style-corrected) data $\mathbf{Z}_i\widetilde{\mathbf{M}}\boldsymbol{\beta}_i$ $(i=1,\ldots,n)$. On the other hand, when $\mathbf{G}$ and $\mathbf{U}$ are fixed, solving for $\mathbf{B}$ is less trivial as this appears in both terms in equation (3). However, minimizing equation (3) with respect to $\mathbf{B}$ can be reduced to a simple constrained least squares problem as follows:

$$L(\mathbf{B},\mathbf{G},\mathbf{U}|\bar{\mathbf{F}},\mathbf{Z},M_1,M_2,M_3)=\sum_{i=1}^{n}\left\|\begin{pmatrix}\sqrt{\lambda}\bar{\mathbf{f}}_i\\(\sqrt{1-\lambda})\mathbf{G}^T\mathbf{u}_i\end{pmatrix}-\begin{pmatrix}\sqrt{\lambda}\mathbf{M}\\(\sqrt{1-\lambda})\mathbf{Z}_i\widetilde{\mathbf{M}}\end{pmatrix}\boldsymbol{\beta}_i\right\|^2. \quad (4)$$

For a proof, see Appendix S2. The CCRS parameters can be estimated using the following algorithm.

Step 1: *Initialization*. Set λ and a convergence criterion ε, randomly choose an initial value for $\mathbf{B},\mathbf{G},\mathbf{U}$, and set the number of iterations w to $w = 1$.
Step 2: *Response function estimation*. For fixed $\mathbf{G},\mathbf{U}$, update $\mathbf{B}$ in such a way that equation (4) is minimized with the constraint in equation (1) (Haskell & Hanson, 1981).
Step 3: *Content-based clustering*. For fixed $\mathbf{B}$, update $\mathbf{G},\mathbf{U}$ using the formula

$$\mathbf{g}_k=\frac{\sum_{i=1}^{n}u_{ik}(\mathbf{Z}\widetilde{\mathbf{M}}\boldsymbol{\beta}_i)}{\sum_{i=1}^{n}u_{ik}},$$

$$u_{ik}=\begin{cases}1 & (k=\underset{s\in\{1,\ldots,K\}}{\operatorname{argmin}}\ ||\sqrt{1-\lambda}(\mathbf{g}_s-\mathbf{Z}_i\widetilde{\mathbf{M}}\boldsymbol{\beta}_i)||^2)\\0 & (\text{otherwise})\end{cases}\qquad (i=1,\ldots,n;k=1,\ldots,K).$$

Step 4: Convergence test. Compute $L^{(w)}$, the value of the objective function (3) using updated parameters and, for $w > 1$, if $L^{(w)} - L^{(w\text{-}1)} < \varepsilon$, terminate; otherwise, let $w = w + 1$ and return to step 2.

Convergence of the algorithm is guaranteed because the objective function (3) is monotonically decreasing in subsequent steps. Note that in step 1 of the algorithm, initial values for $\mathbf{B}$, $\mathbf{G}$, $\mathbf{U}$ need to be selected. This can be done randomly, for example, by randomly generating values from a uniform distribution. Alternatively, one could consider initial values for $\mathbf{B}$, $\mathbf{G}$, $\mathbf{U}$ by solving $\boldsymbol{\beta}_i\ (i=1,\ldots,n)$ for the first term of equation (3), that is, the optimal fitting of the response functions to the boundary data, and applying k-means to corrected data $\mathbf{Z}_i\widetilde{\mathbf{M}}\boldsymbol{\beta}_i\ (i=1,\ldots,n)$ to obtain initial values for $\mathbf{G}$, $\mathbf{U}$. We shall refer to this type of initialization as CCRS tandem initialization.

4. Correcting preference data in the presence of response-style bias by CDS

Schoonees *et al.* (2015) used constrained dual scaling to estimate a response function defined similarly as in Section 2. In dual scaling, which is equivalent to correspondence analysis when analysing contingency tables (van de Velden, 2000), category quantifications are obtained such that the quantifications best capture variance in the data in low-dimensional space. For the analysis of preference data, dual scaling aims to quantify respondents, items and boundaries. In particular, in CDS, one-dimensional quantifications for respondents and boundaries are obtained to model monotonically increasing response functions for clusters of respondents. Response-style bias can then be corrected for in a manner similar to that described in Section 2. A sequential analysis where we first correct for response-style effects using CDS, after which k-means is applied to the corrected data, can be seen as an alternative to the CCRS approach. We refer to such an approach as CDS tandem analysis.

As CDS is based on dual scaling, there are several restrictions. To explain this in detail, let v_i and $w_{b\ell}$ denote values quantified by CDS for respondent i, and the ℓth boundary for the bth response-style-based cluster ($b = 1, \ldots, H$), respectively. In addition, suppose that a respondent i belongs to the bth response-style-based cluster. In CDS, $w_{b\ell} = \phi_b^{\mathrm{CDS}}(\ell)$, where ϕ_b^{CDS} is the CDS response function for the bth response-style-based cluster. Then ϕ_b^{CDS} approximates the rank-ordered boundary data $f_{i\ell}$ as

$$\begin{aligned} &\tilde{f}_{i\ell} \approx v_i \phi_b^{\mathrm{CDS}}(\ell) \quad (i = 1, \ldots, n;\ \ell = 1, \ldots, q-1) \\ &\text{where } \phi_b^{\mathrm{CDS}}(x) = \mu_b + \sum_{r=1}^{3} \alpha_{br} M_r(x) \\ &\text{s.t. } \alpha_{br} \geq 0 \quad (r = 1, 2, 3) \end{aligned} \tag{5}$$

and $\tilde{f}_{i\ell} = f_{i\ell} - p/2$. For the spline basis function M_r in CDS, L and U are respectively set to 0 and q (rather than 0 and 1 as is the case in CCRS). For more details, see Schoonees *et al.* (2015).

Comparing equation (5) with equation (2), it is clear that CDS only estimates response functions for response-style-based clusters $b = 1, \ldots, H$. Hence, due to the one-dimensional approximation only one parameter v_i $(i = 1, \ldots, n)$ in equation (5) is respondent-specific. Therefore, estimating response functions in CDS could incur a significant loss of respondent-specific information.

Note that, by setting $H = n$ and fixing the cluster indicator, CDS may be used to estimate respondent-specific $\boldsymbol{\alpha}_b$ ($b = 1, \ldots, n$) values. However, in practice, this process only yields degenerate solutions in which the parameters are zero or close to zero due to the one-dimensional reduction.

5. Simulation study

We conducted a simulation study to evaluate the performance of CCRS. In addition to the proposed CCRS method, we applied k-means and tandem CDS to the simulated data. In tandem CDS, preference data are first corrected using CDS. Then k-means is applied to the corrected data.

To assess the performance of the methods, we consider two scenarios. In scenario I we assume that there are two kinds of underlying clustering structures: content-based and response-style-based clusters. In scenario II only a content-based clustering structure is assumed. By considering these two scenarios, data are generated corresponding to situations that are assumed to underlie, either implicitly or explicitly, both the tandem CDS and the CCRS methods.

5.1. Data generation

The data-generation process can be divided into two steps: generation of true preferences $x_{ij}^* \in \mathbb{R}$ and mapping of the true preferences to q-point scale data $x_{ij} \in \{1, \ldots, q\}$. Content-based clusters and, for scenario I only, response-style-based clusters, are induced in the first and second steps, respectively. We generate data in such a way that each scenario represents a 'realistic' situation. A detailed description of the data-generation process can be found in Appendix S3.

5.2. Simulation study design

We use a full factorial design with $n = 300, 600$, $m = 20, 30$, $q = 5, 7$, and $K = 2, \ldots, 5$, to assess the performance of the methods in different settings. In addition, for scenario I, in which response-style-based clusters exist, we generated data with $H = 3$ and 5, where H denotes the number of response-style-based clusters. For $H = 3$, the response styles considered are acquiescence, midpoint and no response style. For $H = 5$, disacquiescence and extreme response styles are added. In this simulation we assume that the true number of content-based and response-style-based clusters K and H are known.

For each combination of parameters in our simulation, we randomly generated 100 different data sets. For each data set we apply all methods and assess both content-based and response-style-based clustering.

5.2.1. Evaluation

To evaluate the content-based clustering, the accuracies of CCRS, tandem CDS, and *k*-means clustering are compared. On the other hand, for the response-style-based clustering, the accuracy of CDS is compared to that of CCRS with *k*-means clustering applied to the estimated $\boldsymbol{\beta}_i$ values ($i = 1, \ldots, n$). The adjusted Rand index (ARI) is used to evaluate the retrieval of the underlying structure (Hubert & Arabie, 1985). The ARI assesses the similarity between two cluster allocations (a true and estimated cluster allocation, in this case). It takes a value of 1 for a perfect recovery, and this value decreases as performance worsens.

5.2.2. Selecting the number of response-style-based clusters for CDS

Constrained dual scaling requires a choice for the number of response-style-based clusters H. In scenario II, no response style clusters exist and we therefore need to find an estimate for this. Schoonees *et al.* (2015) use a scree plot of the optimized objective function over different H. However, this approach cannot be used when we want to compare the results from different methods. Therefore, both in our simulation and empirical study, we use the Krzanowski–Lai cluster (KL) index (Krzanowski & Lai, 1988) to determine the number of clusters. The KL index is based on an idea similar to the scree plot, but also takes into account the number of variables.

In the simulation study, we selected different H depending on different n and K values by first running a small simulation. For example, for the $n = 300$ and $K = 2$ case, 10 data sets were simulated for each combination of $m = 20$, 30 and $q = 5$, 7. The KL index was calculated for the results obtained for each data set generated. Among the resulting $2 \times 2 \times 10 = 40$ KL indices, the most frequently selected H value was used as the number of response-style-based clusters for all settings with $n = 300$ and $K = 2$. This process was performed for all different n and K values. The result of this procedure was that $H = 3$ was selected for $n = 300$ and $K = 2, \ldots, 5$, and $H = 4$ was selected for $n = 600$ and $K = 2, \ldots, 5$.

5.2.3. Other setting

Concerning the choice of λ in CCRS, we considered values of .2, .5, .8 and compared the results in each case. In addition, all methods require some type of initialization. For CDS we use the defaults from the *cds* package (Schoonees, 2016) in R (R Core Team, 2017); for *k*-means, we use 100 random starts; for the CCRS method, we consider the CCRS tandem initialization as well as 49 random starts.

5.3. Scenario I: Clustered response styles

Here we only show the results for $n = 300$. The results for $n = 600$ are similar and can be found in Appendix S4.

5.3.1. Content-based clusters retrieval

The ARI results for the content-based clusters (content ARI) are shown in Figure 5. As can be seen, the *k*-means results are poor, possibly due to the presence of response-style bias. However, CDS tandem, which corrects for response-style bias, also demonstrates poor results. Apparently, the joint but uncorrelated presence of content and response clusters makes it difficult for CDS to detect the true content clustering structure.

CCRS tandem and $\lambda = .8$ appear to work well compared to all other methods. A general tendency of the content ARI obtained by CCRS is that greater q and n, and smaller m, K and H values yield better results. Larger q values may yield good results as the estimation of the response function improves when there are more rating categories and hence more boundaries. Note also that the performance of CCRS does not appear to be strongly affected by an increase (from $H = 3$ to $H = 5$) in the number of response styles.

5.3.2. Response-style-based clusters retrieval

The ARI results for the response-style-based clusters (response style ARI) are shown in Figure 6. As can be seen, the mean ARIs for CDS are always below those of CCRS. Furthermore, CCRS with $\lambda = .8$ outperforms the other methods in nearly all cases. Note that the response style ARI results for $\lambda = .8$ are generally better than those for CCRS tandem. An explanation for this could be that CCRS tandem only uses the boundary data $\bar{\mathbf{f}}_t$ to estimate the response functions, while simultaneous CCRS also exploits the underlying content related cluster structure in its estimation.

5.4. Scenario II: Respondent-specific response styles

5.4.1. Content-based cluster retrieval

The ARI results for the content-based cluster structure are shown in Figure 7. As can be seen, there are no big differences from scenario I, with the exception of the *k*-means results. The *k*-means results improved considerably compared to the results in scenario I. An explanation for this could be that the underlying content-based cluster structure in this scenario is no longer obscured by an additional (uncorrelated) response-style-based cluster structure. However, despite this improvement, CCRS still consistently outperforms *k*-means and appears to be useful for obtaining content-based clusters even when no response-style-based clusters are present.

5.5. Conclusions from the simulation study

The results of the simulation study demonstrate that the proposed CCRS method outperforms CDS tandem, and *k*-means in all cases. Moreover, CCRS performed approximately equally well in both scenarios I and II. On the other hand, *k*-means clearly performed worse in scenario I. These results indicate that the proposed CCRS method appears to be robust to having both content-based and response-style-based cluster structures. Overall, CCRS performs better for greater q and smaller K. In addition, the

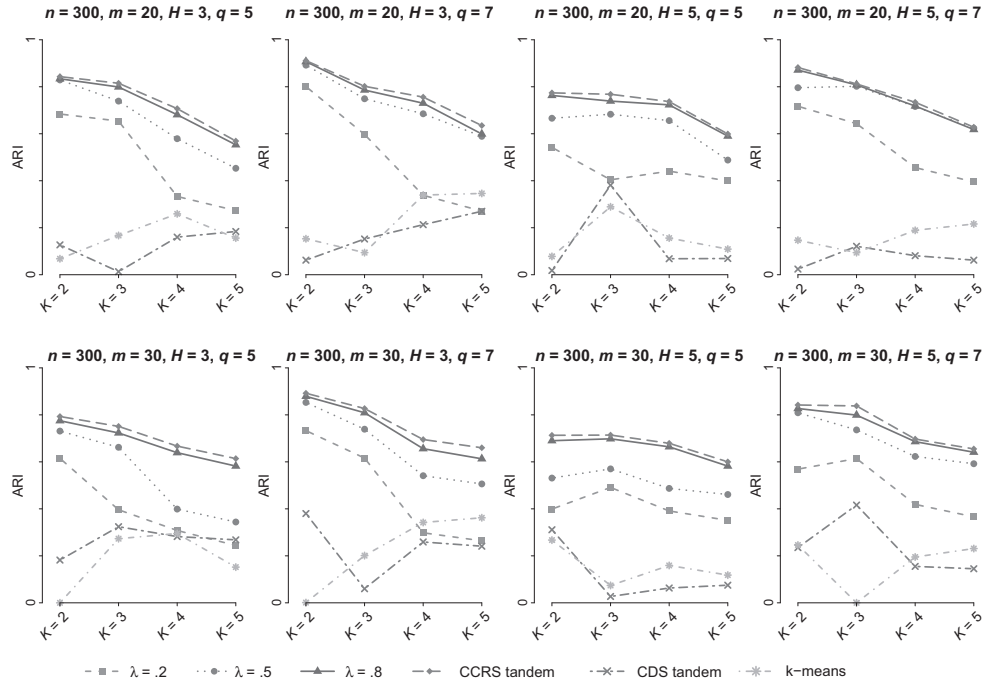

Figure 5. Parallel plot showing mean content ARIs for scenario I (presence of content and response-style-based clusters), $n = 300$ and different parameter settings and methods. [Colour figure can be viewed at wileyonlinelibrary.com]

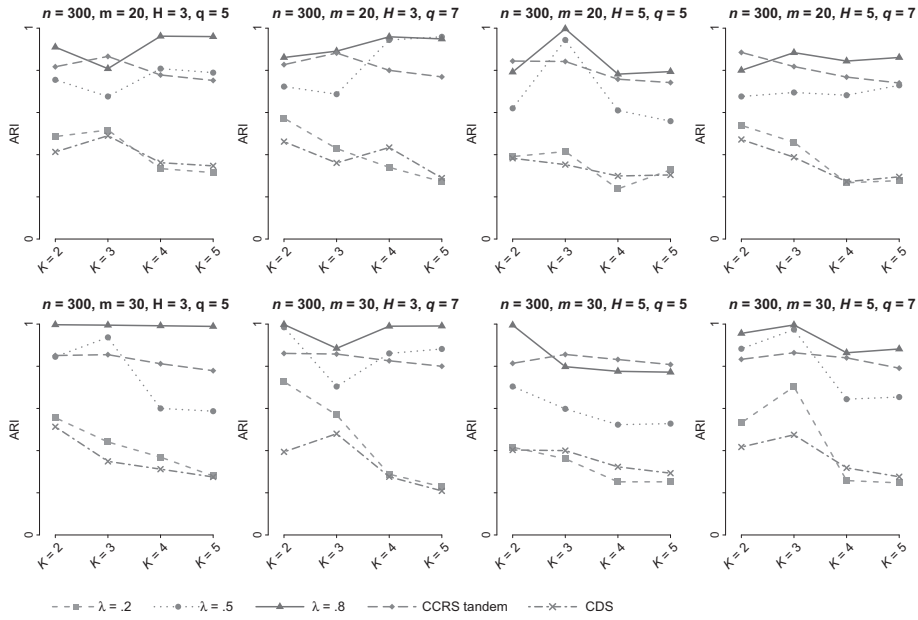

Figure 6. Parallel plot showing mean response style ARIs for scenario I (presence of content and response-style-based clusters), n = 300 and different parameter settings and methods. [Colour figure can be viewed at wileyonlinelibrary.com]

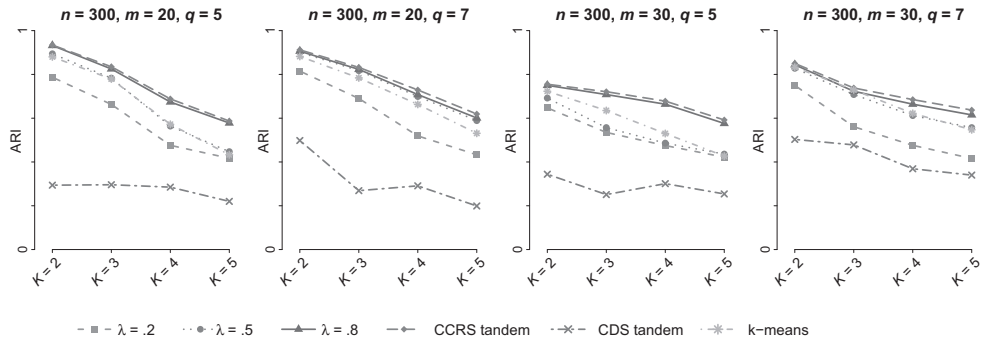

Figure 7. Parallel plot showing mean content ARIs for scenario II (no response style clusters) and different parameter settings and methods. [Colour figure can be viewed at wileyonlinelibrary.com]

performance of CCRS does not appear to be strongly affected by an increase in the number of response styles H, indicating that CCRS can account for more response styles.

It should be noted that, although the results are not reported here, we also ran simulations using $n = 120$. Even though for such small-sample cases the ARI decreased slightly, CCRS still works quite well when compared to the other methods.

The simulation study results showed that the content-based clustering results of CCRS improved when λ increased, although differences between the cluster retrieval results for $\lambda = .8$ and CCRS tandem were very small. However, if a response-style-based clustering structure was present, this structure was better retrieved by selecting $\lambda = .8$. Therefore, we suggest using $\lambda \geq .8$ in order to obtain optimal results for both response-style-based and content-based clustering.

6. Application

6.1. Data

We illustrate the use of CCRS with an empirical application based on survey data collected in 2008 by the East Asian Social Survey (EASS). The survey data include 8,745 respondents in four countries and 107 questions (except demographic variables). The data were downloaded from the ICPSR website (http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/34607). More information about the data is available in Chang, Iwai, Li, and Kim (2014). For our application, we selected 10 items from the survey in which respondents were asked to evaluate five values: patriarchy/gender role, harmony, in-group orientation, hierarchy/authority, and uncertainty avoidance/risk taking. For each of these values respondents were asked to assess two statements using a seven-point Likert scale ranging from 1 (not important at all) to 7 (very important). The 10 statements and corresponding values can be found in Table 1.

In Figure 8 we see that there appears to be considerable difference in response tendencies among the four countries. For example, the Chinese and Taiwanese respondents selected the second highest category much more often than the Korean and Japanese respondents. Moreover, the Japanese respondents tended to select the midpoint more often than respondents from the other countries.

Below, we show the content and response-style-based clustering results obtained by the CCRS and CDS tandem methods, as well as the content clustering results obtained by *k*-means.

Table 1. Value research selected items. Each statement is rated from 1 (strongly disagree), to 7 (strongly agree)

j	Statement	Value
1	It is more important for a wife to help her husband's career than to pursue her own career[a]	Patriarchy/gender role
2	The authority of father in a family should be respected under any circumstances	Patriarchy/gender role
3	It is not desirable to oppose an idea which the majority of people accept, even if it is different from one's own	Harmony
4	One should not express one's complaints about others in order to have good relationship with them	Harmony
5	When hiring someone at a private company, even if an unacquainted person is more qualified it would still be better to give the opportunity to relatives or friends	In-group orientation
6	I would be honored when people who come from the same town play an important role in society	In-group orientation
7	A subordinate should obey the superiors' instructions, even if the person cannot agree with them	Hierarchy/authority
8	If we have capable leaders, it is better to let them decide everything	Hierarchy/authority
9	A life full of risks and chances is more desirable than an ordinary and stable life	Uncertainty avoidance/risk taking
10	With extra money, I would invest in items for high returns even if they are risky	Uncertainty avoidance/risk taking

[a]For Japanese respondents, this statement was phrased differently, even though the same value was measured. Specifically, in the Japanese version the statement was: "A husband's job is to earn money; a wife's job is to look after the home and family."

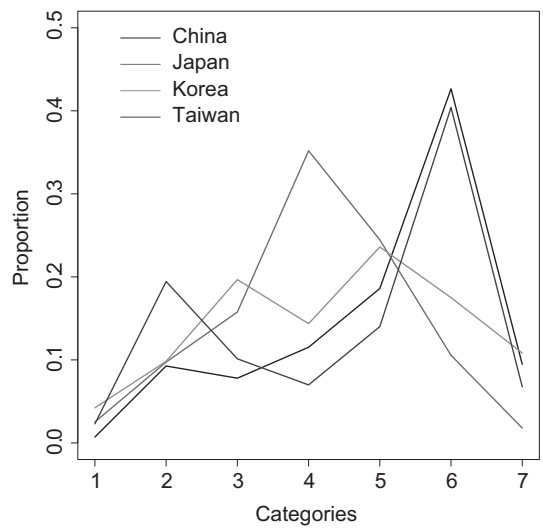

Figure 8. Proportion of rating categories 1–7 selected in different countries for all items. [Colour figure can be viewed at wileyonlinelibrary.com]

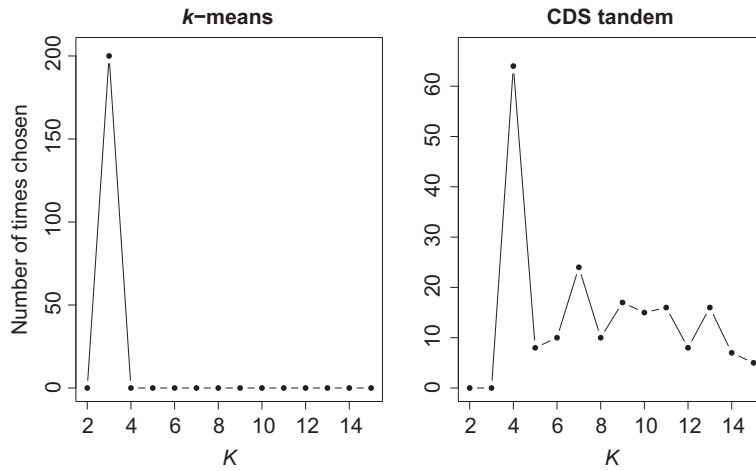

Figure 9. : The number of times that different values of the parameter *K* were selected for content-based clustering methods using the KL index.

6.2. Setting

As these are empirical data, no known true clustering structure exists and all parameters need to be determined based on the data. Similar to the situation in cluster analysis, where selection of the number of clusters is a complex task, selection of such parameters in CDS and CCRS is difficult. In our application, we employed a pragmatic approach and based our selections on the KL index also used in our simulation study. Furthermore, to ensure stability of the selected number of clusters, we based our choice on the results for 200 bootstrap samples. That is, from the complete sample we drew 200 bootstrap samples and, for each bootstrap sample, we selected the *K* value that maximized the KL index. Next, the *K* value that was most often selected in these 200 samples was used in the final estimation. For the CCRS method, which requires two parameters (i.e., *K* and λ), the combination of parameters that was selected most frequently was used.

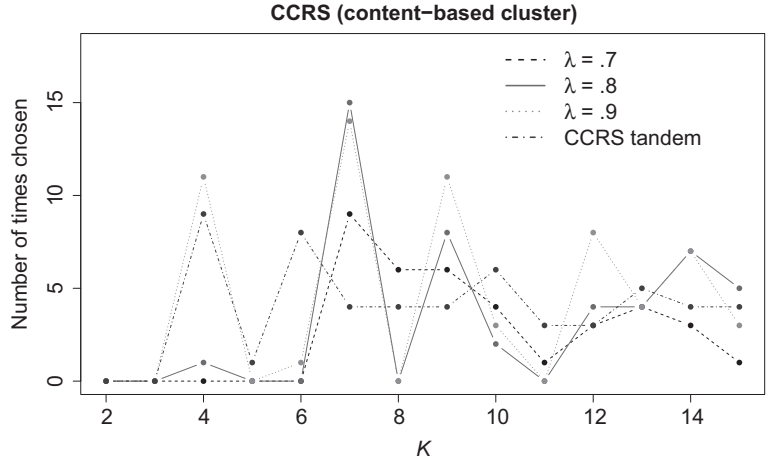

Figure 10. The number of times that different values of the parameter *K* were selected for content-based clustering methods using the KL index for different values of λ. [Colour figure can be viewed at wileyonlinelibrary.com]

To reduce computation times, we used 20 different initial values for each CCRS run on bootstrap sample. In addition, for CDS tandem, K and H were selected sequentially. That is, first H was selected in a similar fashion as described above. Then, using the optimal H, K was determined in the same way.

Using the candidate values $K, H = 2, \ldots, 15$ and $\lambda = .7, .8, .9$ and CCRS tandem, we obtained $K = 7, \lambda = .8$ and $H = 6$ for CCRS, $K = 4$ and $H = 3$ for CDS tandem, and $K = 3$ for *k*-means.

The total number of times each parameter was selected is shown in Figures 9 and 10 for content-based clusters and in Figure 11 for response-style-based clusters. Note that, for *k*-means, $K = 3$ was always selected. For the other methods, the value of K (or H) selected by the KL index varied among the bootstrap samples. However, a clear peak can be identified for most cases.

Once the parameters were set, the methods were applied using 500 different initial values in the same manner as used in the simulation study.

6.3. Clustering results

The content-based clustering results obtained by CCRS are shown in Figure 12. From the boxplots, we see how the clusters differ with respect to the assessment of the items. In some cases these differences are limited to only one item (e.g., clusters 1 and 5), but most differences concern at least two items corresponding to the same value (e.g., clusters 5 and 7).

In most cases, items corresponding to the same values are similarly evaluated within a cluster. However, for the third value (in-group orientation), this does not appear to be the case. Apparently, the evaluation of in-group orientation differs depending on the group considered (i.e., relatives or people from the same town). However, the difference may also be due to the phrasing of the two items. In particular, in question 5 respondents assess whether they would favour relatives, whereas question 6 merely asks respondents whether they appreciate the success of others (from the same town in this case).

In CCRS, respondent-specific response functions are used. Clustering the resulting functions leads to a six-cluster solution. The corresponding response functions are

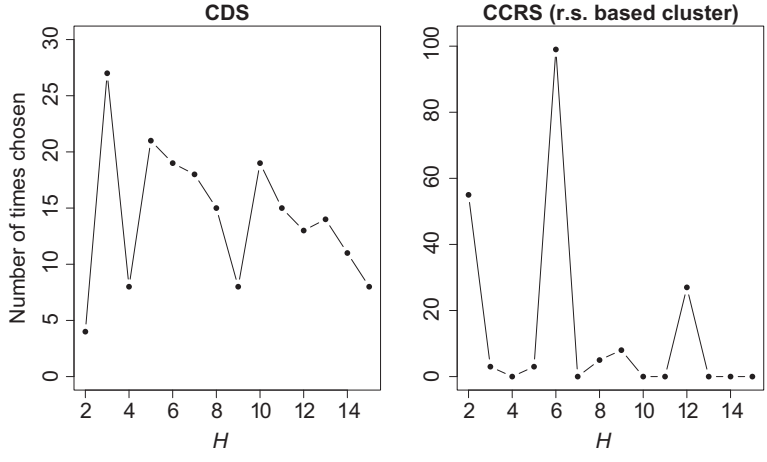

Figure 11. The number of times that different values of the parameter H were selected, using the KL index, for response-style-based clustering.

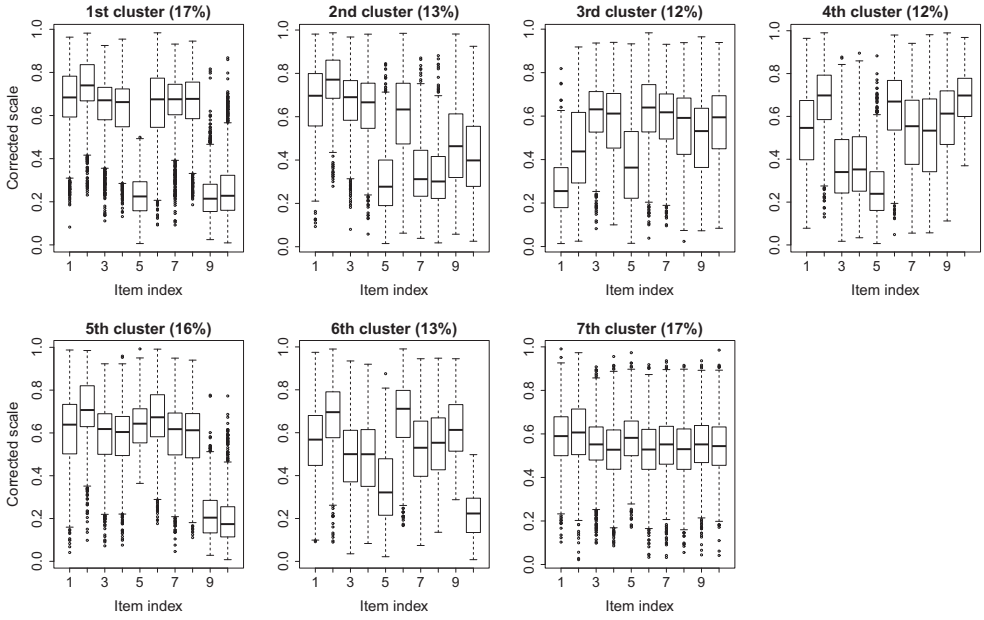

Figure 12. Boxplots for the 10 items (horizontal axis) for the content-based (cont.) clusters obtained with CCRS. The vertical axis indicates the scale of the corrected data using estimated response functions.

depicted in Figure 13 (left). Moreover, recall that the coefficients in the response functions estimated by CCRS can be used to visually capture the characteristics of response styles (cf. Section 5.2). The results are shown in Figure 13 (right). The second and fourth response-style-based clusters correspond to a high β_1 and a low β_3 value, indicating an 'acquiescence' response style. Similarly, the third response-style-based cluster demonstrates low values for both β_1 and β_3 corresponding to a 'midpoint' response style.

To see whether the response-style-based clusters are related to nationalities, we consider the distributions over the countries in Table 2. In the second and fourth clusters (acquiescence), most respondents are Chinese. On the other hand, the third response-style-based cluster (midpoint) comprises over 50% Japanese.

To see how the response-style-based clusters and content-based clusters correspond, we consider a mosaic plot that visualizes the cross-tabulation of the two cluster solutions. Figure 14 shows that there does not appear to be significant overlap of respondents between the content-based and response-style-based clusters. In each content-based cluster respondents from all response-style-based clusters are present. That is, the content-based clusters and the response-style-based clusters do not coincide.

To assess whether results obtained by CCRS are 'better' than results obtained by the CDS tandem and *k*-means methods is cumbersome as we do not know whether there are true underlying cluster structures. Nevertheless, a comparison of results may be insightful and help in their interpretation.

The CDS tandem content-based clustering results and cluster-wise response functions are shown in Figures 15 and 16, respectively. Looking at the association between the content-based and response-style-based clusters as shown in Figure 17, we see that there is significant overlap between respondents in the content-based and response-style-based

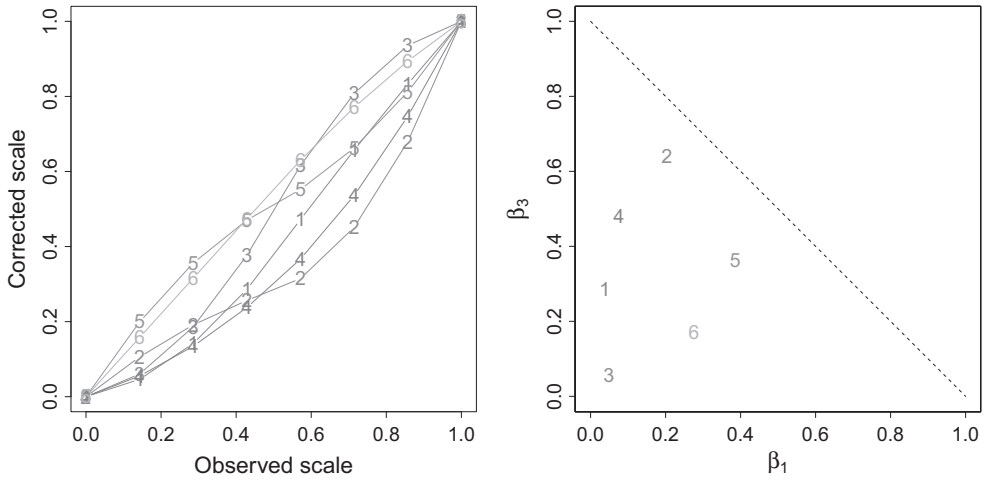

Figure 13. (Left) Estimated response functions of the response-style-based clusters obtained by CCRS. (Right) Low-dimensional plot for β_b of the response-style-based clusters. Numbers indicate response-style-based clusters, and correspond to those used in the left plot. [Colour figure can be viewed at wileyonlinelibrary.com]

Table 2. Distribution of respondents' nationalities (%) over the response-style-based clusters ($b = 1, \ldots, 6$) obtained by CCRS

b	China	Japan	Korea	Taiwan
1	43.8	14.0	20.5	21.6
2	52.3	1.7	20.4	25.6
3	17.7	53.2	23.8	5.3
4	59.7	2.3	8.4	29.6
5	27.7	6.8	18.0	47.4
6	16.6	24.7	28.6	30.2

clusters. This indicates that the cluster-wise correction of response-style results in content-based clusters that are similar to those response-style-based clusters. Consequently, when interpreting content-based clusters one may merely be considering response-style-based differences.

The *k*-means clustering results are shown in Figure 18. Here, the clusters also appear to correspond to some response tendencies. In the first and third content-based clusters, for example, we see that respondents predominantly select high and midpoint ratings, respectively. An interpretation relative to the item content appears difficult for this solution.

In summary, it appears that the *k*-means results may only reflect response tendencies. Moreover, when using CDS to correct for response-style effect, the corrected data strongly reflect certain response-style-based clustering results. Consequently, the content-based cluster results obtained from the corrected data do not yield additional content-related insights. On the other hand, with the proposed CCRS method, we obtain content-based clusters that are dissimilar to the response-style-based clusters.

Finally, results of the empirical data application cannot be easily validated, and the fact that we find 'dissimilar' clusters does not provide evidence that CCRS should be preferred

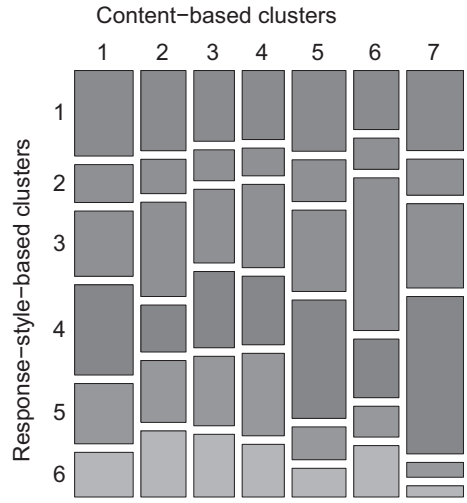

Figure 14. Mosaic plot of the CCRS content-based and response-style-based clusters. The index of response-style-based clusters corresponds to the number used in Figure 13. [Colour figure can be viewed at wileyonlinelibrary.com]

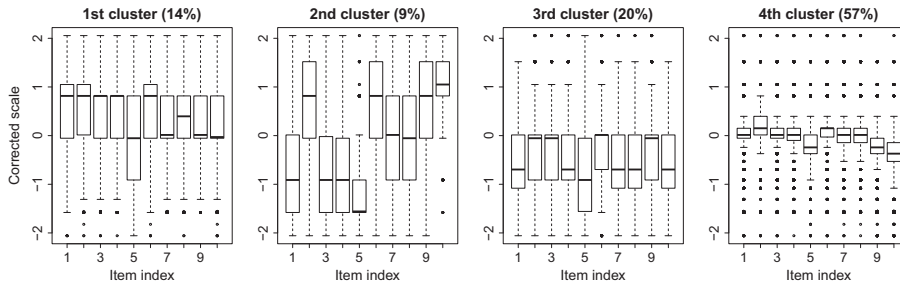

Figure 15. Boxplot of 10 items (horizontal axis) of the content-based clusters obtained by CDS tandem. The vertical axis indicates the scale of the corrected data using the estimated response functions.

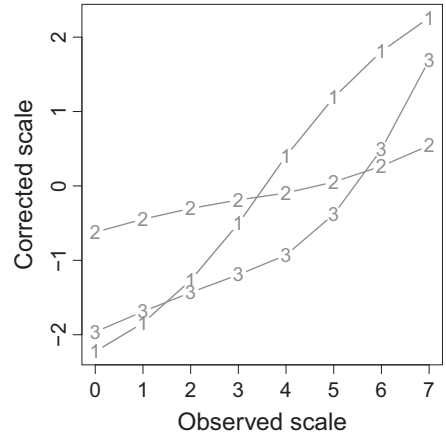

Figure 16. Estimated response functions of response-style-based cluster obtained by CDS. [Colour figure can be viewed at wileyonlinelibrary.com]

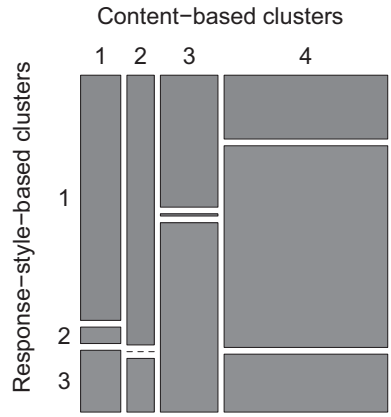

Figure 17. Mosaic plot of the content-based and response-style-based clusters by CDS tandem. The index of response-style-based clusters corresponds to the number in Figure 16. [Colour figure can be viewed at wileyonlinelibrary.com]

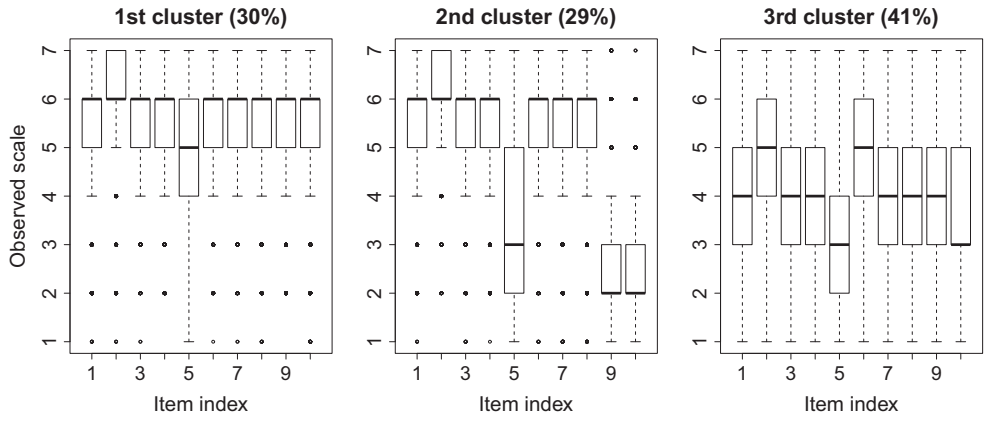

Figure 18. Boxplot of 10 items (horizontal axis) of the content-based clusters obtained by *k*-means. The vertical axis indicates the scale of the original preference data.

over CDS. However, the results of this application, in combination with the results of the simulation study, do suggest that CCRS is better able to retrieve content-based cluster structures.

7. Conclusion and discussion

In this study we introduced a new method, called CCRS, for simultaneously correcting for response-style bias and performing content-based clustering. By generalizing the concept of a response function as introduced by van de Velden (2008) and Schoonees *et al.* (2015), respondent-specific response functions were estimated without first applying a dimension reduction technique. In CCRS, we obtain clusters which are not affected by response-style bias. Note that our new correction method, explained in Section 3.1, which is a part of CCRS, can also be used to correct for response-style bias in combination with other methods and applications.

We demonstrated in a simulation study that our proposed CCRS method outperforms existing methods such as CDS tandem (CDS and *k*-means) as well as *k*-means in most cases. In particular, when both content-based and response-style-based clustering structures exist, CCRS performs better at retrieving the content-based clustering structure. Overall, having fewer clusters and more rating categories (i.e., a larger rating scale) yields better CCRS results for both content-based and response-style-based clusters. In addition, we showed that the performance of CCRS is not strongly affected by an increase in the number of response styles H and a decrease in the sample size n.

Using an empirical data set, we illustrated that CCRS yields different content-based and response-style-based clusters, whereas both CDS tandem and *k*-means lead to content-based clusters that are hard to distinguish from response-style-based clusters. Obviously, the results of the empirical data are difficult to validate, as is often the case in cluster analysis studies. Nevertheless, these results do illustrate that the potential challenge associated with existing methods (i.e., identifying clusters that are merely related to response tendencies) can be mitigated with the proposed approach.

We implemented our method in the R package ccrs for the statistical computing environment R (R Core Team, 2017), which can be downloaded from the comprehensive R archive network (Takagishi, 2019, https://cran.r-project.org/web/packages/ccrs).

There are many opportunities for future work based on the proposed approach. While we focus on evaluating the clustering retrieval in this study, it is also important to validate the accuracy of the correction itself. In addition, only content-based clusters that differ from response-style-based clusters were extracted in this paper; however, if there exists a response-style-like content-based cluster (such as a content-based cluster of mostly midpoint values), additional tools, such as anchoring vignettes (King, Murray, Salomon, & Tandon, 2004), may be required to distinguish them.

Finally, for the new framework described in Section 2, only the relationship with CDS was considered in this paper. However, it would be interesting to investigate its relationship with IRT methods as well as the method proposed by van Rosmalen *et al.* (2010). Such an evaluation could possibly result in a very general framework for correction that includes various existing correction methods and would facilitate a comparison of the correction accuracies of different correction methods.

Acknowledgements

We thank the East Asian Social Survey (EASS) who conducted the survey which is used in our empirical data analysis. In addition, this work was supported by a JSPS KAKENHI Grant Number 17J06200.

References

Arabie, P., & Hubert, L. (1994). Cluster analysis in marketing research. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 160–189). Oxford, UK: Blackwell.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences*, *40*, 1235–1245. https://doi.org/10.1016/j.paid.2005.10.018

B¨ockenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*, 159–181. https://doi.org/10.1111/bmsp.12086

Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*, 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*, 665–678. https://doi.org/10.1037/a0028111

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, *22*, 69–83. https://doi.org/10.1037/met0000106

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*, 335–352. https://doi.org/10.1177/0146621608329891

Chang, Y. H., Iwai, N., Li, L., & Kim, S. W. (2014). *East Asian Social Survey (EASS), cross-national survey data sets: Culture and Globalization in East Asia, 2008*. Ann Arbor, MI: EASSDA, Inter-university Consortium for Political and Social Research (ICPSR).

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, *31*, 187–212. https://doi.org/10.1177/0022022100031002003

Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods*, *11*(1), 14.

Haskell, K. H., & Hanson, R. J. (1981). An algorithm for linear least squares problems with equality and nonnegativity constraints. *Mathematical Programming*, *21*, 98–118. https://doi.org/10.1007/BF01584232

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218. https://doi.org/10.1007/BF01908075

Hwang, H., Dillon, W. R., & Takane, Y. (2006). An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. *Psychometrika*, *71*(1), 161–171. https://doi.org/10.1007/s11336-004-1173-x

Javaras, K. N., & Ripley, B. D. (2007). An "unfolding" latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association*, *102*, 454–463. https://doi.org/10.1198/016214506000000960

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*, *68*, 563–583. https://doi.org/10.1007/BF02295612

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*(1), 191–207. https://doi.org/10.1017/S000305540400108X

Krzanowski, W. J., & Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, *44*, 23–34. https://doi.org/10.2307/2531893

MacQueen. (1967). Some methods for classification and analysis of multivariate observations. In *Proceed-ings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. *1*, pp. 281–297). Berkeley, CA: University of California Press.

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, *44*, 1539–1550. https://doi.org/10.1016/j.paid.2008.01.010

Meiser, T., & Machunsky, M. (2008). The personal structure of personal need for structure. *European Journal of Psychological Assessment*, *24*(1), 27–34. https://doi.org/10.1027/1015-5759.24.1.27

Moors, G. (2012). The effect of response style bias on the measurement of transformational, transactional, and laissez-faire leadership. *European Journal of Work and Organizational Psychology*, *21*, 271–298. https://doi.org/10.1080/1359432X.2010.550680

Morren, M., Gelissen, J., & Vermunt, J. (2012). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. *Methodology: European Journal of*

Research Methods for the Behavioral and Social Sciences, *8*, 159–170. https://doi.org/10.1027/1614-2241/a000048

Nishisato, S. (1980). Dual scaling of successive categories data. *Japanese Psychological Research*, *22*, 134–143. https://doi.org/10.4992/psycholres1954.22.134

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*, 875–899. https://doi.org/10.1177/0013164413514998

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, *3*, 425–441. https://doi.org/10.1214/ss/1177012761

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*, 75–92. https://doi.org/10.1111/j.2044-8317.1991.tb00951.x

Schoonees, P. C. (2016). *cds: Constrained dual scaling for detecting response styles*. (R package version 1.0.3)

Schoonees, P. C., Velden, M. van. de., & Groenen, P. J. (2015). Constrained dual scaling for detecting response styles in categorical data. *Psychometrika*, *80*, 968–994. https://doi.org/10.1007/s11336-015-9458-9

Stukovsý, R., Palat, M., & Sedlakova, A. (1982). Scoring position styles in the elderly. *Studia Psychologica*, *24*, 145–154.

Takagishi, M.(2019). *ccrs: Correct and cluster response style biased data. (R package version 0.1.0)*. Vienna, Austria: R Foundation for Statistical Computing.

van Rosmalen, J., Van Herk, H., & Groenen, P. J. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research*, *47*, 157–172. https://doi.org/10.1509/jmkr.47.1.157

van de Velden, M. (2000). Dual scaling and correspondence analysis of rank order data. In R. D. H. Heijmans, D. S. G. Pollock & A. Satorra (Eds.), *Innovations in multivariate statistical analysis* (pp. 87–99). Dordrecht, Netherland: Kluwer Academic Publisher. https://doi.org/10.1007/978-1-4615-4603-0

van de Velden, M. (2008). Detecting response styles by using dual scaling of successive categories. In K. Shigemasu, A. Okada, T. Imaizumi & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 517–524). Tokyo, Japan: Universal Academy Press.

von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 99–115). New York, NY: Springer. https://doi.org/10.1007/978-0-387-49839-3

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods*, *15*, 96–110. https://doi.org/10.1037/a0018721

Received 4 April 2018; revised version received 22 February 2019

Supporting Information

The following supporting information may be found in the online edition of the article:

Appendix S1. Properties and interpretation of CCRS.
Appendix S2. Proof for constrained least squares expression.
Appendix S3. Data generation in simulation study.
Appendix S4. Simulation results for $n = 600$ case.