Mountain or Molehill? A Simulation Study on the Impact of Response Styles

Hansjörg Plieninger
University of Mannheim

Even though there is an increasing interest in response styles, the field lacks a systematic investigation of the bias that response styles potentially cause. Therefore, a simulation was carried out to study this phenomenon with a focus on applied settings (reliability, validity, scale scores). The influence of acquiescence and extreme response style was investigated, and independent variables were, for example, the number of reverse-keyed items. Data were generated from a multidimensional IRT model. The results indicated that response styles may bias findings based on self-report data, and that this bias may be substantial if the attribute of interest is correlated with response style. However, in the absence of such correlations, bias was generally very small, especially for extreme response style and if acquiescence was controlled for by reverse-keyed items. An empirical example was used to illustrate and validate the simulation. In summary, it is concluded that the threat of response styles may be smaller than feared.

Keywords: response styles, simulation, item response theory

Introduction

There exists the widespread claim and fear that response styles—such as acquiescence response style (ARS) or extreme response style (ERS)—distort results based on self-report data. The goal of the present simulation study was to present data rather than claims and to scrutinize the effect of response styles. The study covered three scenarios of a prototypical psychological research process, namely, estimating the reliability of a scale, testing its validity via correlations, and assigning a score to every respondent. To closely mirror situations in the applied field, the simulated data were analyzed using basic procedures (e.g., Cronbach's alpha) without trying to control for response styles. The data generating model, however, was a rather complex item response model that allowed to flexibly cover a variety of conditions.

Hansjörg Plieninger, School of Social Sciences, Department of Psychology, University of Mannheim, Germany.

This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences funded by the German Research Foundation.

The author would like to thank Thorsten Meiser for comments that substantially helped to improve this manuscript.

Correspondence concerning this article should be addressed to Hansjörg Plieninger, Department of Psychology, University of Mannheim, 68131 Mannheim, Germany. E-mail: plieninger@uni-mannheim.de

Response Styles

Response styles are defined as the tendency to respond to questionnaire items irrespective of content (cf. Nunnally, 1978). This does not imply

that the subject matter is irrelevant to the respondent, but indicates that response styles act independently of content and that both sources influence the actual response. This theoretical notion is supported by empirical evidence showing that response styles are stable across content domains (e.g., Weijters, Geuens, & Schillewaert, 2010a; Wetzel, Carstensen, & Böhnke, 2013). Moreover, it is well documented that response styles are stable within a questionnaire as well as across periods of several years (e.g., Aichholzer, 2013; Weijters, Geuens, & Schillewaert, 2010b).

Response styles represent a source of interindividual variance—additional to the content-related variance—that is usually not taken into account in analyses of self-report data, at least in more applied settings. There seem to be three different viewpoints on the matter. First, probably the majority of practitioners and researchers ignore response styles, because they don't know enough about them or cannot implement (statistical) control for one reason or another. Second, some take the position that response styles are negligible because this source of variance is small, represents error variance, or is trifling compared to content (e.g., Rorer, 1965; Schimmack, Böckenholt, & Reisenzein, 2002). Third, many researchers believe that response styles are a serious threat to the quality of self-report data that potentially influence all kinds of measures scientists usually draw conclusions from. For example, Eid and Rauber (2000) stated that "differences in category use can distort the results [...]" (p. 21). Likewise, Weijters, Geuens, and Schillewaert (2010b) wrote that "response styles have been found to bias estimates of means, variances, and correlations [...], leading to potentially erroneous results and conclusions [...]" (p. 96).

Although individual findings support the impression that response styles form a severe threat, the literature lacks a systematic investigation of the amount of bias and the conditions under which bias occurs. Simulation studies are well suited to address this issue, because they allow a comprehensive analysis of a specific effect (e.g., of response styles) while having full control over all other influences. However, there are only very few studies published that attempt to look at response styles from the perspective of a simulation study. Heide and Grønhaug (1992) published a simulation in a marketing journal and found biasing effects of ARS and ERS, but their methodological approach was rather basic from today's perspective. The paper of Ferrando and Lorenzo-Seva (2010) also contains a simulation study on ARS with a limited range of conditions finding that ARS can bias results, but that this bias is minor for most practical purposes, at least with fully balanced scales (i.e., equal number of regular and reverse-keyed items). Savalei and Falk (2014) found that substantive factor loadings were only affected by ARS when its influence was strong. Wetzel, Böhnke, and Rose (2016) investigated trait recovery of different methods, which aim to control for ERS, and stated: "The results of our simulation study imply that ignoring ERS on average hardly affects trait estimates if ERS and the latent trait are uncorrelated or only weakly correlated [...]" (p. 17).

Statistical Models for Response Styles

A multitude of models to measure and/or control for response styles have been proposed, which vary greatly in terms of their objectives, requirements, and complexity (cf. Van Vaerenbergh & Thomas, 2013). For example, in confirmatory factor analysis, an additional acquiescence factor can be used to analyze scales comprised of both regular and reverse-keyed items (e.g., Billiet & McClendon, 2000). Different routes have been pursued in the family of item response theory (IRT). For example, mixture distribution Rasch models have been applied with the result that a 2-class solution could be interpreted as comprising non-extreme and extreme respondents (e.g., Eid & Rauber, 2000; Meiser & Machunsky, 2008; Wetzel et al., 2013). Böckenholt (2012) proposed a multidimensional IRT model in which the original response is separated into content- and response style-related

processes using dichotomous pseudoitems (cf. De Boeck & Partchev, 2012; Khorramdel & von Davier, 2014; Plieninger & Meiser, 2014). Another multidimensional IRT model, namely, a variant of Bock's *nominal response model*, was developed by Bolt and colleagues (Bolt & Newton, 2011; Johnson & Bolt, 2010) and further extended by Falk and Cai (2016). Furthermore, multidimensionality arising from random thresholds is accounted for in models suggested by Wang (e.g., Jin & Wang, 2014).

Most of the models proposed so far focus on only one response style and cannot be modified to accommodate another one. However, Wetzel and Carstensen (2015) recently proposed an approach in the framework of multidimensional Rasch models that allows to take into account both ARS and ERS.

Multidimensional Rasch Models

Multidimensional Rasch models date back to Georg Rasch (1961) himself and have, since then, been presented in multiple ways. Herein, the notation of Adams, Wilson, and Wang (1997), who call their approach *multidimensional random coefficients multinomial logit model*, is adapted. Therein, it is assumed that—possibly multiple—latent variables drive the item responses in an additive manner. The model has only one type of item parameter, namely, a difficulty parameter, which herein—for the sake of simplicity—was parametrized using a *rating scale model* approach (cf. Andrich, 1978), but other versions of the model for ordinal and binary items exist. In the current study, it is furthermore assumed that a symmetric, bipolar response format is used (e.g., ranging from *strongly disagree* to *strongly agree*).

Assume we have item i ($i = 1, \ldots, I$) with $K + 1$ response categories ($k = 0, 1, \ldots, K$) and person j ($j = 1, \ldots, J$). The model has d ($d = 1, \ldots, D$) latent dimensions and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)'$ is a column vector containing one person parameter per dimension. In the rating scale model, the item parameters comprise item location parameters β_i reflecting the overall difficulty of an item and threshold parameters τ_k, which are constant across items. This results in $I + K$ different item parameters overall contained in the vector $\boldsymbol{\xi} = (\beta_1, \ldots, \beta_I, \tau_1 \ldots, \tau_K)'$. The threshold parameters are constrained to sum to zero, $\sum_1^K \tau_k = 0$[1]. If the model parameters are to be estimated from empirical data, additional restrictions on the person or on the item parameters have to be made, because the model is otherwise not identified (cf. Adams et al., 1997).

Both the item parameters and the person parameters are mapped onto the category probabilities using a *design matrix* $\mathbf{A}$ and a *scoring matrix* $\mathbf{B}$, respectively. The linear combination of item parameters pertaining to category k of item i is defined by a row vector $\mathbf{a}_{ik}$ (of length $I + K$). The matrix $\mathbf{A}_i$ comprises $K+1$ of these row vectors stacked below each other and defines the design matrix for item i, and I of these matrices are then again stacked below each other defining the design matrix $\mathbf{A}$. An example for two items with three categories is depicted below:

$$\mathbf{A} * \boldsymbol{\xi} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 2 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 2 & 1 & 1 \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \beta_2 \\ \tau_1 \\ \tau_2 \end{bmatrix}.$$

The weight of category k of item i on each of the dimensions is defined by the row vector $\mathbf{b}_{ik}$ (of length D). The matrix $\mathbf{B}_i$ comprises $K + 1$ of these row vectors stacked below each other and defines the design matrix for item i, and I of these matrices are then again stacked below each other defining the design matrix $\mathbf{B}$. Three examples of scoring matrices for two items with three categories are depicted below. The first one is typically employed in polytomous, unidimensional models like the rating scale model. The second is an example

[1] All K τ parameters are explicitly displayed in the example below for consistency with the simulation setup even though the constraint makes one of the τ parameters redundant.

of *between-item multidimensionality*, where each item loads on only one dimension. The last one is an example of *within-item multidimensionality*, where the second item loads on both dimensions:

$$\mathbf{B}^{(1)} * \theta = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \\ 1 \\ 2 \end{bmatrix} * \begin{bmatrix} \theta_1 \end{bmatrix};$$

$$\mathbf{B}^{(2)} * \boldsymbol{\theta} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 2 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 2 \end{bmatrix} * \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix};$$

$$\mathbf{B}^{(3)} * \boldsymbol{\theta} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 2 & 0 \\ 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{bmatrix} * \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}.$$

Then, the probability of a response falling in category k of item i is modeled as

$$P(X_{ik} = 1; \mathbf{A}, \mathbf{B}, \boldsymbol{\xi} | \boldsymbol{\theta}) = \frac{\exp(\mathbf{b}_{ik}\boldsymbol{\theta} - \mathbf{a}_{ik}\boldsymbol{\xi})}{\sum_{k=0}^{K} \exp(\mathbf{b}_{ik}\boldsymbol{\theta} - \mathbf{a}_{ik}\boldsymbol{\xi})},$$

where $\mathbf{a}_{ik}$ and $\mathbf{b}_{ik}$, respectively, represent a row vector of $\mathbf{A}$ and $\mathbf{B}$, respectively, pertaining to the kth category of item i. The model reduces to Andrich's rating scale model for $D = 1$ and to the Rasch model for $K = 1$ and $D = 1$.

Multidimensional Rasch Models for Response Styles

Previously, multidimensionality within items has been investigated in situations where items measure more than one dimension at a time (cf. Adams et al., 1997). Wetzel and Carstensen (2015) extended the idea of within-item multidimensionality noting that not all of the latent dimensions need necessarily be related to the content of the items, but could also be related to, for example, response styles. This, in turn, requires different weights composing the matrix $\mathbf{B}$. Assuming that each response involves one attribute- and one response style-dimension, scoring matrices for an item with five categories involving ERS and ARS, respectively, may look as follows (cf. Wetzel & Carstensen, 2015):

$$\mathbf{B}^{(ERS)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & 0 \\ 3 & 0 \\ 4 & 1 \end{bmatrix}; \quad \mathbf{B}^{(ARS)} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 2 & 0 \\ 3 & 1 \\ 4 & 1 \end{bmatrix}.$$

For ERS, the direction (agreement vs. disagreement) of the response is still governed by the first, content-related dimension alone; however, the extremity of the response may be altered by ERS. Contrarily, ARS may alter the direction of the response and may lead to, for example, agreement with both regular and reverse-keyed items.

In summary, multidimensional Rasch models are an interesting alternative to existing response style models. First, the model is very flexible: Various forms of response styles can be implemented; the only restriction is to find sensible weights for the matrix $\mathbf{B}$. Second, this allows to simulate ERS and ARS from the same model facilitating the design of the study as well as the interpretation and comparison of the results. Third, multiple attribute-dimensions can be included. Fourth, the framework incorporates a unidimensional (content-only) model as a special case. Fifth, the model is parsimonious, because, for example, the number of item parameters is independent of the number of dimensions. These features make the model well-suited for the purposes of the present study, which aimed to investigate both ARS and ERS and which was intended to realistically cover situations of applied data analysis. Apart from that, even though it is a new model, the underlying notion of response styles is highly similar to that of established approaches (e.g., Johnson & Bolt, 2010; Weijters, Cabooter, & Schillewaert, 2010).

The Present Research

The present simulation study aimed to scrutinize the claim that response styles threaten the results of self-report data, especially so in applied settings. In more detail, the idea was to simulate data using the framework introduced above, to subsequently ignore response styles during data analysis (as is often done in the field), and finally to quantify the bias introduced by ARS or ERS. In order to cover a variety of settings, three different scenarios resembling prototypical steps of a research process were designed. First, Cronbach's alpha is arguably the most prominent measure of the reliability of a set of items, and it was investigated whether response styles would bias this measure (and how much). Second, the validity of a scale is often assessed using the correlation of two scale scores, and it was again investigated whether response styles would bias this measure (and how much). Third, the ultimate goal of assessment is to assign a score to every person. The accuracy of this was investigated (a) using correlations of true and observed scores and (b) by comparing the rank order of persons with and without response styles. In other words, it was examined how response styles may influence a decision (e.g., in health, education, work) that is based on self-report data. The analyses in all three scenarios employed raw score-based measures derived from classical test theory—for the reason that those measures are heavily used in applied research. This simulation study went beyond previous work in that the influence of both ERS and ARS was investigated under a wide range of conditions. Furthermore, the results of the simulation were verified and illustrated with an empirical example.

Method

Simulation Design and Set-Up

The simulation model had D dimensions comprising the attribute(s) of interest, θ_1 and possibly θ_2, (e.g., personality trait, attitude, symptom) and the response style θ_{RS}. In the case of two attributes, θ_1 and θ_2 each influenced a unique set of items (between-item multidimensionality), whilst θ_{RS} always influenced all items (within-item multidimensionality). In each replication, the person parameters were sampled from a multivariate normal distribution, $\boldsymbol{\theta} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \\ \mu_{\text{RS}} \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{1,2} & \rho_{1,\text{RS}} \\ \rho_{1,2} & \sigma_2^2 & \rho_{2,\text{RS}} \\ \rho_{1,\text{RS}} & \rho_{2,\text{RS}} & \sigma_{\text{RS}}^2 \end{bmatrix}.$$

If $\theta_{\text{RS}} \sim N(0, 0)$, the response style dimension effectively drops making the model a content-only model.

In order to manipulate the amount of response style variance relative to substantive variance, σ_1^2 ($= \sigma_2^2$) was fixed to a value of one. The amount of response style variance σ_{RS}^2 was varied between values of 0.00 and 1.00 (in steps of 0.10) indicating how diverse a sample is with respect to response styles, and higher values indicate more diversity. In each replication, the off-diagonal elements in $\boldsymbol{\Sigma}$ were drawn from a Wishart distribution with an identity matrix used as the *scale matrix* and $df = 10$; this results in the fact that the correlations have an expected value of zero and a variance of .10. The center of the response style distribution μ_{RS} was varied between values of -1.00 and 1.00 (in steps of 0.10). Positive (negative) values indicate that the sample overall tends to give more extreme responses (more non-extreme responses) for ERS and more (less) agree-responses for ARS, respectively.

Each replication entailed 200 persons and 10 items per attribute.[2] The number of categories was not varied and set to five (but see the Appendix). The number of reverse-keyed items was varied between zero and five per attribute. In each replication, the item location parameters β_i were drawn from a truncated normal distribution, $TN(0, 1, -1.5, 1.5)$. The item threshold parameters τ_k

[2] Pilot simulations revealed that N and I had virtually no effect on bias when varied between 100 and 1,000 and between five and 15, respectively.

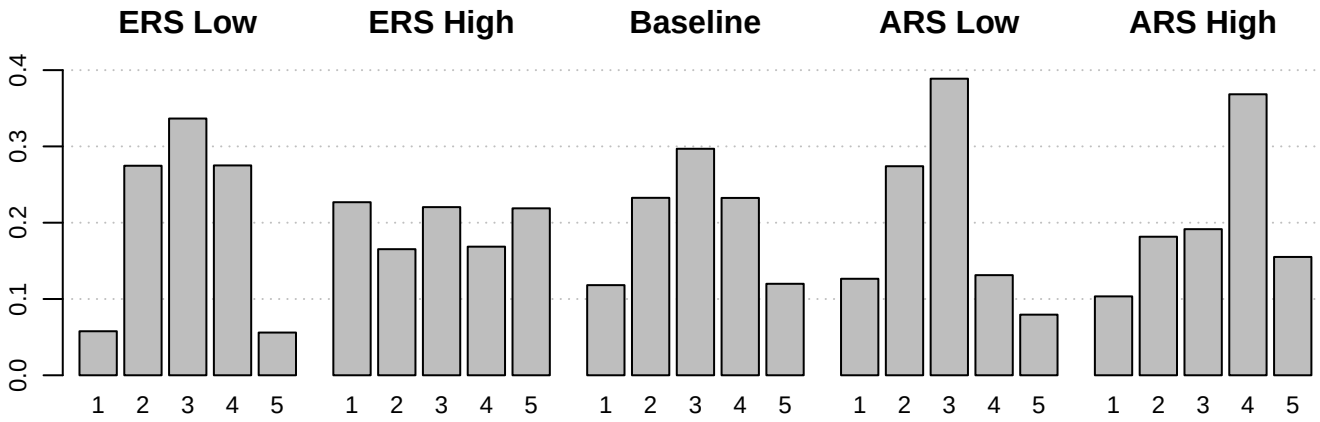

Figure 1. Illustration of the effect of response styles ($\mu_{\text{RS}} = 0$; $\sigma^2_{\text{RS}} = 1$; $\rho_{1,\text{RS}} = 0$). Displayed are responses to a 5-point item of the lower and the upper third of the—ERS or ARS—distribution as well as a baseline condition without response styles.

were each drawn from a uniform distribution, U(-2.5, 2.5), and they were sorted in ascending order to avoid category reversals.[3] Subsequently, the thresholds were centered because of the restriction $\sum_1^K \tau_k = 0$ (and it was made sure that none of the τ parameters exceeded the limits of ±2.5). To illustrate the effect response styles have in the present model with the given set-up, an example is shown in Figure 1; data were generated for 100,000 people and an item of intermediate difficulty with equally spaced threshold parameters between -1.5 and 1.5. The figure shows, for example, that the uppermost third of the ERS distribution used the extreme categories twice as much compared to the baseline condition without response styles.

The scoring matrix **B** of each simulation model was adopted from the following template according to the number of attributes, their assigned number of regular and reverse-keyed items, and the type of response style:

$$\mathbf{B}' = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} \text{(regular item)} \\ \text{(reverse-keyed item)} \\ \text{(ERS)} \\ \text{(ARS).} \end{matrix}$$

Dependent Variables

In the first scenario, Cronbach's alpha was used as an estimate of the reliability of a set of items. To compare this value to a response style-free measure, it was made use of the concept of *covariate-free reliability* recently introduced by Peter Bentler (2016). He proposed a measure of *covariate-free alpha*, which controls Cronbach's alpha for the influence of a covariate (i.e., response styles in the present case) via partialing.[4] The actual dependent variable that was used in the analyses was the amount of bias, that is, the difference between Cronbach's alpha and covariate-free alpha.

In the second scenario, a scale score $\bar{x}_d$ (i.e., the mean across items after recoding) was computed for both attributes. The correlation of these two

[3] In the case of two attributes, the threshold parameters were equal for both attributes.

[4] Regressing a true score T on a covariate Z yields a covariate-dependent part $T^{(Z)}$ and an orthogonal, covariate-free part $T^{\perp Z}$. Thus, it follows that $\sigma^2_T = \sigma^2_{T^{(Z)}} + \sigma^2_{T^{\perp Z}}$. Bentler (2016) showed that this holds also for the mean of the item-covariances (i.e., $\bar{\sigma}_{ij} = \bar{\sigma}^{(Z)}_{ij} + \bar{\sigma}^{\perp Z}_{ij}$), which is used in the equation of Cronbach's alpha, and proposed to decompose Cronbach's alpha into a covariate-dependent and a covariate-free part (i.e., $\alpha = \alpha^{(Z)} + \alpha^{\perp Z}$).

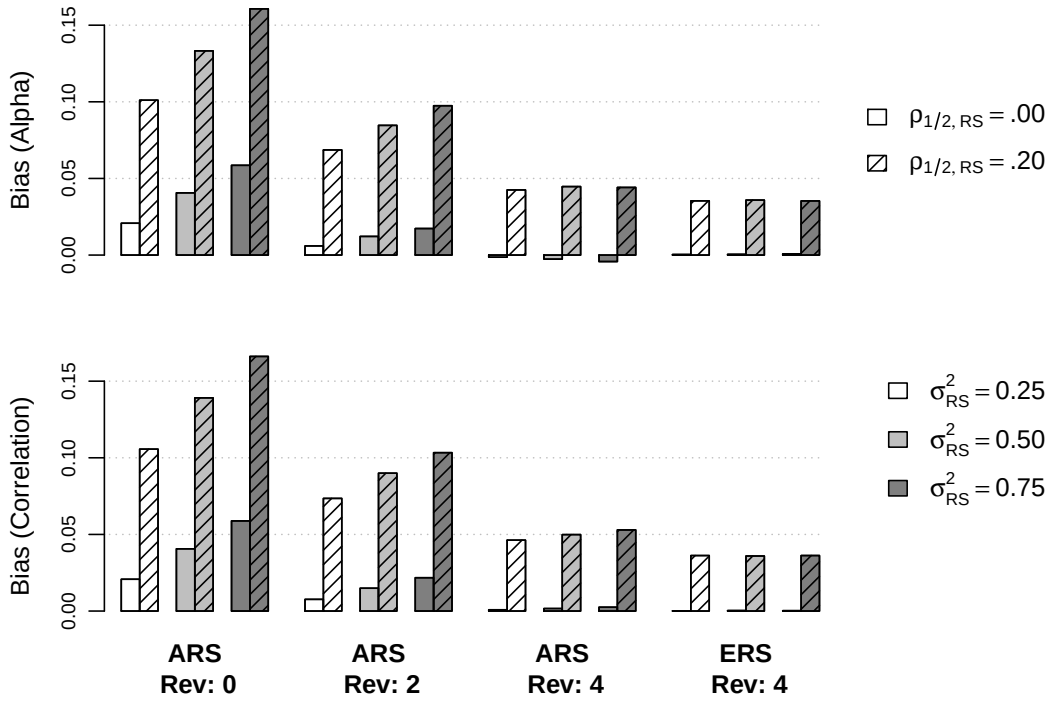

Figure 2. Overview of average bias with respect to Cronbach's alpha (upper panel) and the correlation of two scale scores (lower panel). Results are based on 1,000 replications in each of the selected conditions. $\rho_{1/2,\mathrm{RS}}$ stands for the correlation of the attribute (alpha) or each of both attributes (correlation) with the response style.

scale scores was compared to the partial correlation that controls the correlation of interest for response style ($r_{\bar{x}_1,\bar{x}_2\cdot\theta_{\mathrm{RS}}}$). Again, bias was used as the dependent variable, that is, the difference between the observed and the partial correlation.

In the last scenario, the true person parameters θ_1, which are independent of response styles, were compared with the observed scale scores, which are influenced by both the attribute and response styles. First, these two variables were correlated. Second, the rank order of persons was compared at different cutoffs. For example, at a cutoff value of c = .80, people at or above the 80th percentile of scale scores were classified as *positive*. Additionally, this classification was done using the true person parameters θ_1 and the same cutoff c resulting in four possible outcomes: *true positives* (TP; originally and observed positive), *false positives* (FP; originally negative but observed positive), *false negatives* (FN; originally positive but observed negative), and *true negatives* (TN; originally and observed negative). To illustrate the results, three different measures at 39 equally spaced cutoffs between c = .025 and c = .975 were calculated in every replication: the *true positive rate* (TPR = TP/[TP + FN]) or *sensitivity* indicating how many of the people originally above the cutoff were indeed selected, the *false positive rate* (FPR = FP/[FP + TN]) indicating how many of the people originally below the cutoff were falsely selected, and the overall *accuracy* (ACC = [TP + TN]/[TP + FP + FN + TN]) indicating the total rate of correct classifications.

Results

The results are based on 100,000 replications for each simulation. In each single replication, the values of all variables were randomly and independently drawn. The simulations and analyses were

Table 1
Effect of Response Styles on Cronbach's Alpha as a Function of Reverse-Keyed Items and Joint Distribution of Attribute and Response Style

	ARS				ERS			
	Model 1		Model 2		Model 1		Model 2	
	b	b^*	b	b^*	b	b^*	b	b^*
Intercept	0.082		0.010		0.079		0.001	
Reversed	−0.006	−0.09	−0.006	−0.10	0.000	0.00	0.000	0.00
μ_{RS}	0.001	0.00	0.000	0.00	−0.001	−0.01	−0.001	−0.01
σ^2_{RS}	0.047	0.13	0.010	0.03	0.042	0.12	0.000	0.00
$\rho_{1,\text{RS}}$	0.139	0.37	0.140	0.38	−0.001	0.00	0.001	0.00
$(\rho_{1,\text{RS}})^2$			0.792	0.65			0.858	0.74
Reversed × $\rho_{1,\text{RS}}$			−0.055	−0.25				
R^2		0.17		0.95		0.02		0.96

Note. All predictor variables were centered. All SEs for paramters $b \leq .001$.

conducted in R 3.2.1 (R Core Team, 2014).[5]

An overview of the average bias with respect to Cronbach's alpha (upper panel) and a correlation coefficient (lower panel) for selected conditions is given in Figure 2: ARS led to more bias compared to ERS, more reverse-coded items reduced bias, and more response style variance led to more bias. Furthermore, bias rarely exceeded levels of .05 if the attribute(s) and the response style were uncorrelated, but the opposite was true if the attribute(s) and the response style were moderately correlated. This figure gives already instructive insights, and more detailed results are reported in the following sections. In line with recommendations, for example, by Harwell (e.g., Harwell, Stone, Hsu, & Kirisci, 1996), it was chosen to refrain from presenting full-page tables with descriptive results. Rather, the results of each simulation were submitted to a regression model, which facilitates interpretation and makes it easier to detect effects of higher order. Unstandardized (b) and standardized (b^*) regression coefficients are reported.

Estimating the Reliability of a Scale in the Presence of Response Styles

Acquiescence. Two regression models were fit to the simulation results, one without and one with higher-order terms (see Table 1), and the following interpretation is based on the correctly specified, second model. Overall, the intercept indicated that—on average—the estimated alpha coefficient (which was .88) slightly overestimated the reliability by .01. Bias increased when fewer reverse-keyed items were used and when ARS variance was higher. Moreover, the substantive linear and quadratic effects of $\rho_{1,\text{ARS}}$ indicated that bias was most pronounced if ARS was related to the attribute of interest. Furthermore, an interaction effect indicated that reverse-keyed items buffer against the biasing effect of the attribute-ARS correlation. Both the interaction and the quadratic effect are illustrated in Figure 3 (left panel). There was no effect of μ_{RS} in this or any of the other simulations, because this parameter simply causes a shift of all responses without an effect on individual differences.

[5] It was made use of the packages **MASS** (Venables & Ripley, 2002), **truncnorm** (Trautmann, Steuer, Mersmann, & Bornkamp, 2014), and **magic** (Hankin, 2005).

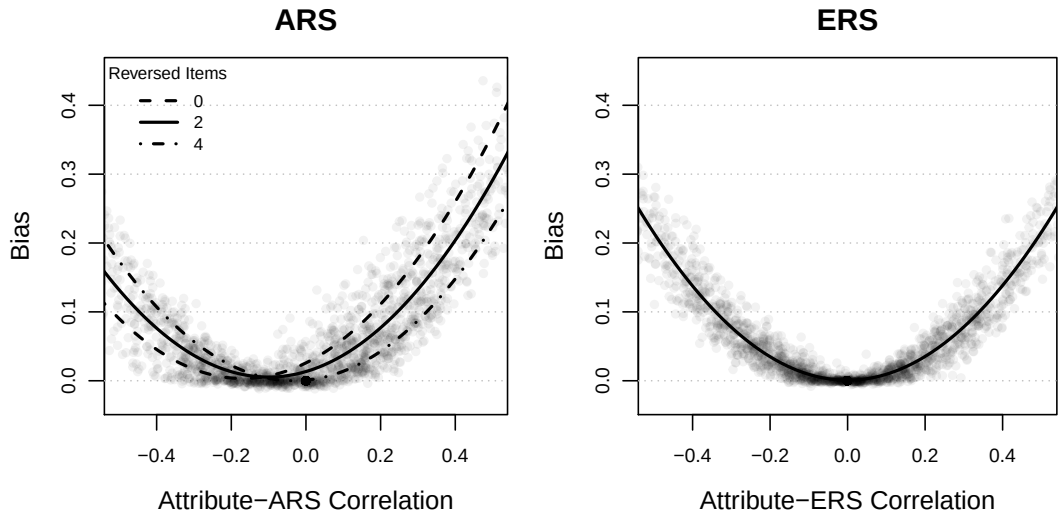

Figure 3. Effect of ARS and ERS, respectively, on Cronbach's alpha. Plotting region was restricted to $\left|\rho_{1,\mathrm{RS}}\right| < .5$ and a subset of 2,000 replications.

Extreme Response Style. This simulation focused on the effect of ERS on Cronbach's alpha. The intercept was virtually zero (Model 2) indicating that Cronbach's alpha was almost unbiased if $\rho_{1,\mathrm{ERS}} = 0$ (see Table 1). However, there was again a substantial quadratic effect of the attribute-ERS relationship, which is illustrated in Figure 3 (right panel). When the attribute and ERS were positively related, persons with a high (low) attribute level tend to give more (less) extreme answers. Thus, these responses undergo an upward-shift resulting in the fact that the items share additional variance that is due to ERS. When the relationship is negative, the responses are shifted downwards, which also increases the shared variance. This additional variance is wrongly attributed to the attribute if ERS is ignored leading to the observed bias.

Estimating the Correlation of Two Scales in the Presence of Response Styles

In addition to the previous scenario, the simulations now entailed further independent variables, namely, the correlation of the two attributes ($\rho_{1,2}$) as well as the correlation of the attributes with response style ($\rho_{1,\mathrm{RS}}$ and $\rho_{2,\mathrm{RS}}$, respectively).

Acquiescence. The results in Table 2 indicated that the actual correlation was, on average, slightly overestimated by a value of .01 when acquiescence was ignored as indicated by the intercept. Mirroring the results presented above, this bias became larger when fewer reverse-keyed items where employed and when ARS variance increased. Again, the center of the ARS distribution had no impact. Additionally, the negative slope of the true correlation between the two attributes indicated that bias became smaller the more positive the true relationship became. This is due to the fact that ARS makes correlations more positive, and the impact of this decreases the more positive the true correlation of the attributes already is. Note that this effect is only interpretable in the correctly specified, second model (see Table 2).

The second model revealed several interaction effects. The most important one was the interaction between the two attribute-ARS correlations ($\rho_{1,\mathrm{RS}} \times \rho_{2,\mathrm{RS}}$), which is also depicted in Figure 4. The correlation of interest was overestimated if the attribute-ARS relationships were either both positive or both negative, and the correlation of interest was underestimated if the attribute-ARS relationships were of opposite sign. How-

Table 2
Effect of Response Styles on Scale Score Correlation as a Function of Reverse-Keyed Items and Joint Distribution of Attributes and Response Style

	ARS				ERS			
	Model 1		Model 2		Model 1		Model 2	
	b	b^*	b	b^*	b	b^*	b	b^*
Intercept	0.012		0.012		0.000		0.000	
Reversed	−0.007	−0.12	−0.006	−0.11	0.000	0.00	0.000	0.00
σ^2_{RS}	0.024	0.08	0.024	0.08	0.000	0.00	0.000	0.00
μ_{RS}	0.002	0.01	0.002	0.01	0.000	0.00	0.000	0.00
$\rho_{1,2}$	0.005	0.02	−0.076	−0.25	0.015	0.05	−0.069	−0.24
$\rho_{1,\text{RS}}$	0.079	0.25	0.082	0.26	−0.001	0.00	0.000	0.00
$\rho_{2,\text{RS}}$	0.081	0.25	0.085	0.27	−0.001	0.00	0.001	0.00
$\rho_{1,\text{RS}} \times \rho_{2,\text{RS}}$			0.877	0.83			0.924	0.94
Reversed $\times$ σ^2_{RS}			−0.013	−0.07				
Reversed $\times$ $\rho_{1,2}$			0.006	0.03				
Reversed $\times$ $\rho_{1,\text{RS}}$			−0.031	−0.17				
Reversed $\times$ $\rho_{2,\text{RS}}$			−0.031	−0.16				
$\sigma^2_{\text{RS}} \times \rho_{1,2}$			−0.055	−0.06				
$\sigma^2_{\text{RS}} \times \rho_{1,\text{RS}}$			0.073	0.07				
$\sigma^2_{\text{RS}} \times \rho_{2,\text{RS}}$			0.075	0.07				
$\rho_{1,2} \times \rho_{1,\text{RS}}$			−0.075	−0.07				
$\rho_{1,2} \times \rho_{2,\text{RS}}$			−0.072	−0.07				
R^2		0.15		0.91		0.00		0.89

Note. All predictor variables were centered. All SEs for paramters $b \leq .001$.

ever, bias was small if at least one of the attributes was unrelated to ARS. Apart from that, reverse-keyed items buffered against the detrimental effect of an attribute-ARS relationship. However, this holds only for one attribute at a time and not for their interaction: The three way interaction (Rev $\times$ $\rho_{1,\text{ARS}} \times \rho_{2,\text{ARS}}$) did not explain additional variance. Furthermore, all effects became stronger the more ARS variance was in the data as revealed by the respective interactions.

Extreme Response Style. The previous simulation was repeated focusing now on ERS, and the results are displayed in Table 2. Both regressions indicated that bias was, on average, virtually zero. Moreover, none of the first-order predictors explained a substantial amount of variance. However, the interaction between the two attribute-ERS relationships was again large, which mirrors the quadratic effect (of $\rho_{1,\text{ERS}}$) observed in the scenario before. If all three intercorrelations were positive, people high (low) on both attributes gave more (less) extreme responses, which shifted these responses upwards. If both attribute-ERS relationships were negative, however, responses were shifted downwards. In both cases, the shared variance among items was inflated leading to on overestimation of the correlation between the two attributes. Contrarily, attribute-ERS correlations that were of opposite sign resulted in inverted patterns across the two attributes (upwards shift on one at-

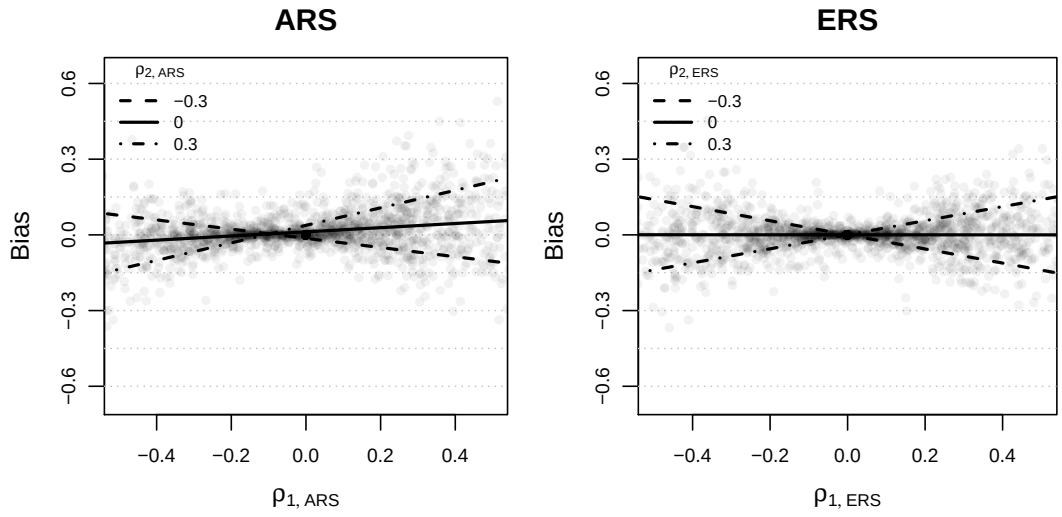

Figure 4. Effect of ARS and ERS, respectively, on the correlation of two scale scores. Plotting region was restricted to $|\rho_{1,\text{RS}}| < .5$ and a subset of 2,000 replications.

tribute and downwards shift on the other), which deflated the shared variance among items. Most important, however, was the effect that bias was virtually zero if one attribute-ERS relationship was close to zero. Furthermore, average bias did not exceed values of .08 for moderate attribute-ERS relationships ($|\rho| < .3$).

Estimating Respondents' Scale Scores in the Presence of Response Styles

In the final scenario, the effect of response styles on respondents' scale scores was investigated. The previous results already revealed that response styles can affect the relationship between observed and true scores (i.e., the reliability)—and this is of course due to distorted scores of respondents. The following simulations were intended to show a more fine-grained picture at the level of individual scores, because these are often the final goal in many situations. This made it necessary, to reduce the complexity of the simulation design in order to keep the presentation of the results concise. Therefore, an extreme condition with $\sigma^2_{\text{RS}} = 1$ was contrasted with a situation were response styles were absent (i.e., $\sigma^2_{\text{RS}} = 0$). Furthermore, it was decided to focus on reverse-keyed items, because this is the variable that can directly be controlled by the researcher. The effect of zero, two, and four reverse-keyed items was investigated for ARS (and arbitrarily set to four for ERS). Apart from that, 10 items, five categories, 200 persons, $\mu_{\text{RS}} = 0$, and $\rho_{1,\text{RS}} = 0$ was specified for each sample. Given the reduced number of conditions, only 10,000 replications were run in each simulation.

The results (i.e., correlations, TPR, FPR, and accuracy) are depicted in Figure 5. Even in the absence of response styles (depicted in gray), there was some natural discrepancy between the observed scales scores, $\bar{x}_1$, and the true person parameters, θ_1, due to the unreliable measurement with only 10 five-point items ($r = .93$). This was also reflected in the non-perfect accuracy, TPR, and FPR.

The effect of ERS and ARS, respectively, is mirrored in the difference between the response style condition (displayed in black) and the baseline condition (displayed in gray). In the uppermost panel of Figure 5, the influence of ERS is depicted, and the results indicated that ERS was problematic with respect to the TPR when selecting the highest performing individuals. For example, the TPR dropped from .84 to .81 at $c = .80$ (i.e., the up-

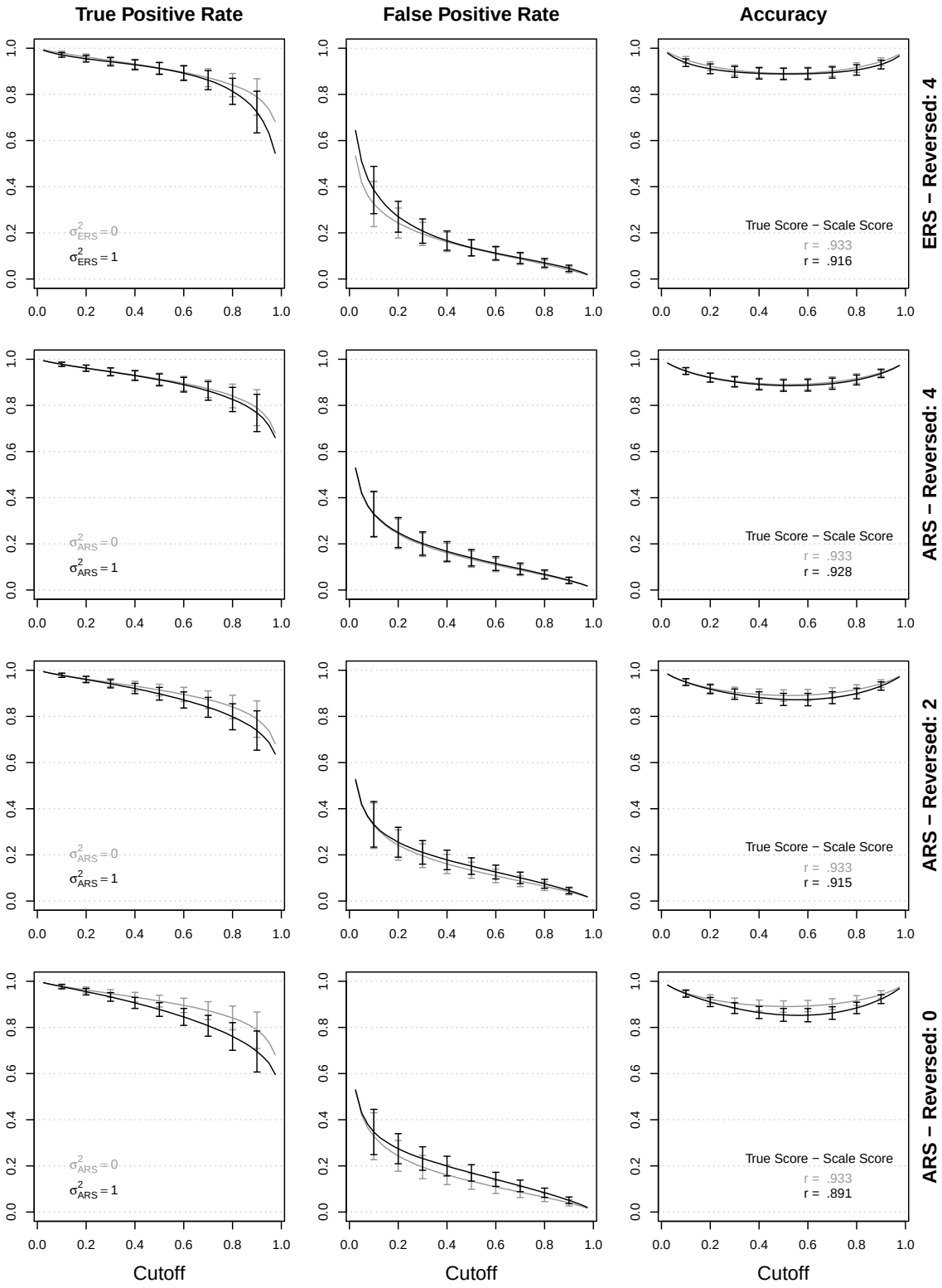

Figure 5. Influence of ERS and ARS on the classification of respondents, which is mirrored in the difference between the baseline condition without response styles (in gray) and the response style condition (in black). The lines represent the mean across all replications at a given cutoff value, error bars represent standard deviations.

permost 20% of a sample are selected). The accuracy indicated that ERS was most influential in the mildly extreme areas of the scale. In this range, ERS may make a difference between a 4- and a 5-response (or a 1 and a 2). In the outermost areas, the attribute level is so high or low making ERS less influential. Similarly, in the center of the scale, only very extreme ERS levels have the potential to alter responses in categories 2, 3, and 4.

The results for ARS with four, two, and zero reverse-keyed items are displayed in the three lower panels of Figure 5. All three measures were impaired in the presence of ARS, the more so the less reverse-keyed items were used. For example, with zero reverse-keyed items at $c = .80$, the TPR was only .76 (compared to .84 in the baseline condition), the FPR increased to .08 (compared to .06), and the accuracy was .88 (compared to .92). However, this effect was substantially reduced when using two or even four reverse-keyed items. In the latter case, ARS had virtually no effect at all. Note that the slight asymmetry in the impact of ARS (i.e., higher impact in the upper range of the scale) is simply due to an odd number of categories (ARS contrasts two agree-categories with three non agree-categories) and would disappear with an even number of categories.

Taken together, even though only an extreme condition—response style variance equal to content variance—was investigated herein, the effect of response styles was once again rather minor. This was especially true with respect to ERS and with respect to ARS controlled for by reverse-keyed items.

An Illustrative Example

An empirical data set from Jackson (2012) was analyzed in order to illustrate the effects of response styles in real data and to check whether the parameter values chosen in the simulations were reasonable. Respondents that were older than 80 ($n = 12$) or with unclear sex ($n = 56$) were excluded. Furthermore, 23 cases were removed because these persons showed no variability in the chosen response option across more than 25 subsequent items. The final data set included 8,745 persons who provided responses to 50 Big Five items. Openness, Conscientiousness, Extraversion, Agreeableness, and Emotional Stability were measured by ten 5-point items each (including 3, 4, 5, 4, and 8, respectively, reverse-coded items). Two models were fit to the data using a partial credit model parametrization[6]: Model 1 comprised one dimension for each of the five scales (between-item multidimensionality), and Model 2 included two additional dimensions for ARS and ERS, respectively, that where each measured by all 50 items. The model was fit using the R package **TAM** (Kiefer, Robitzsch, & Wu, 2015) employing Quasi-Monte Carlo integration with 5,000 nodes. For the sake of brevity, only model-based results (rather than raw score analyses) are reported in the following.

Model 2 had 13 parameters more than Model 1 (two variances, 11 covariances) and was clearly superior in terms of model fit (e.g., BIC-values of 1,130,022 for M1 and 1,093,083 for M2). The estimated item parameters of the two models were highly similar, $r > .99$, with a mean absolute difference (MAD) of .09. The correlations of the five pairs of corresponding person parameters (EAP) estimated by the two models were $r = .88$, $r = .95$, $r = .97$, $r = .90$, and $r = .96$; *MAD*s ranged between 0.14 and 0.26. In Model 2, the estimated variances of the Big Five dimensions were 0.59, 0.54, 1.20, 0.65, and 0.89, and those values were on average .04 smaller compared to Model 1. Furthermore, the estimated variance of ARS was 0.14 and that of ERS was 1.02. The latent intercorrelations of the Big Five dimensions in Model 2 ranged from $-.04$ to .46; the differences between these correlations and those from Model 1 ranged from $-.03$ to .04 with an average of .02 in abso-

[6] A rating scale model parametrization fit the data worse but did not affect the interpretation of the results. Few if any differences regarding the coefficients reported herein were observed in the second or third decimal place.

lute terms. The correlations between ARS and the Big Five dimensions ranged from −.09 to .11, and the correlations between ERS and the Big Five dimensions ranged from −.19 to .04. The estimated correlation between ARS and ERS was $r = .26$. Finally, in Model 2, the estimated EAP reliabilities for the Big Five ranged from .77 to .89 with an average of .83. Those values were on average .02 (between .01 and .04) smaller than those from Model 1 indicating that the reliability was slightly overestimated when response styles were ignored.

In summary, these results indicated that controlling for response styles increased model fit and led to different parameter values. However, these differences were rather small. Apart from that, the response style variances and covariances were in the range of the values chosen in the simulation above (with the only, small exception being σ^2_{ERS}).

Discussion

There is an increasing interest in response styles, and many models to measure and control for response styles have been developed. The justification for this research activity is—partly—the belief that not taking response styles into account would distort self-report data, which are used all over the place in (social) science. The goal of the present research was to scrutinize this belief with a focus on applied settings and, in turn, to take a more systematic look at the role of response styles.

Therefore, a simulation study was carried out for the two most prominent response styles, ARS and ERS, and for three different scenarios: one looking at Cronbach's alpha, one looking at correlations, and one looking at respondents' scores. These scenarios were selected to resemble typical situations of applied data analysis, where response styles are often ignored—either because response styles are believed to be negligible or because methodological control cannot be realized for one reason or another.

While the generated data were analyzed with everyday methods, they were simulated from a sophisticated, but straightforward IRT model. Wetzel and Carstensen (2015) extended the idea of within-item multidimensionality in the polytomous Rasch model to dimensions that are not related to content but to response styles. The only difference to traditional within-item multidimensionality is that the response style dimension receives weights that are different from the traditional, ordinal coding. This model was well suited for the present simulations, because it is highly flexible and different response styles and/or multiple attributes can be incorporated. Moreover, the underlying notion of response styles is similar to existing approaches.

The results were twofold. On the one hand, bias was large when the attribute of interest was correlated with response style, and bias got extreme for large correlations. Such attribute-response style relationships might perhaps explain empirical findings of a notable impact of response styles. However, correlations outside ±.20 were not observed in the empirical example presented herein and may in general be the exception rather than the rule. Moreover, if the correlation really is in the range of .40, .60, or even higher, the question arises what the items at hand actually measure and whether the problem may be socially desirable responding rather than ARS (cf. Paunonen & LeBel, 2012). In such situations, self-reports might not be a sensible way of data collection.

On the other hand, bias was small or even negligible in a large range of conditions. This holds especially if the attribute-response style relationship was small. Moreover, bias was lower when more reverse-keyed items were used (for ARS), when the attribute-attribute correlation was higher, and when response style variance was smaller. For example, in the conditions in Figure 2 where response styles were unrelated to the attribute(s), bias hardly exceeded levels of .05 or even .02 if at least two reverse-coded items were used. In summary, the findings are in line with previous work finding only small effects as long as the attribute-response style relationship is small (Ferrando & Lorenzo-Seva, 2010; Johnson & Bolt, 2010; Savalei & Falk, 2014; Wetzel et al., 2016).

The presented empirical example supported the interpretation that response styles can introduce bias, but that this bias is rather small and unlikely to alter results completely. Moreover, the example showed that the parameter values chosen for the simulation were reasonable and definitely not understated.

The issue of an attribute-response style relationship brings up further questions about causation and the nature of response styles themselves. Let's look at two examples with ARS. First, if a bivariate relationship between an attribute and ARS is observed, this may be simply due to a common cause or confounder (e.g., cultural background) whilst the bivariate relationship is in fact non-existing. Thus, a correct model would include the confounder but not necessarily ARS. Second, two independent attributes may both causally influence ARS—then called a collider. If ARS is wrongly included in the model, a spurious relationship between the two attributes may result. These examples highlight that a much deeper understanding of response styles and their causes and causal effects is needed in order to evaluate the impact of attribute-response style relationships.

If a rule of thumb should be derived from the present results, ⅓ of reverse-keyed items are probably a good way to control ARS. There was no evidence that a fully balanced scale would further improve the results markedly, at least if the attribute-response style correlation was reasonably small. However, it should be noted that the use of reverse-keyed items may have downsides, and there is ample literature on that topic (cf. Weijters, Baumgartner, & Schillewaert, 2013). Apart from that, the number of reverse-keyed items had no effect on the bias caused by ERS. This is not surprising given that the definition of ERS is independent of the direction of an item.

As with every simulation study, the generalizability of the results depends on the external validity (a) of the simulation model, (b) of the parameter values (fixed or varied), and (c) of the analysis model. First, the chosen model allowed for a very natural implementation of ERS—a tendency to select the endpoints—and ARS—a tendency to agree. Moreover, the model is highly similar to existing approaches and there is no reason to assume that a different simulation model (e.g., Johnson & Bolt, 2010) would lead to fundamentally different results. Furthermore, differences between the chosen rating scale approach and a partial credit approach would probably cancel each other out across items and replications. Second, the chosen parameter values seemed plausible given the empirical example. Moreover, the regression results make it straightforward to plug in values (e.g., $\sigma^2_{\text{RS}} = 2$) that were not covered herein. And, a wide range of conditions was realized by randomly sampling from the independent variables instead of restricting the study to, say, three levels of every factor. This, in turn, allowed to uncover quadratic and interaction effects. Third, the analyses focused on bias, which was based on partialing response style from the measure of interest. If the attribute and response style were correlated, this led to the fact that also attribute variance was—wrongly—partialed out inflating the amount of bias, the more so the stronger the correlation was. Thus, the extreme levels of bias (e.g., for $\rho = .5$) are probably a (too) pessimistic estimate. Apart from that, the analyses focused on only three scenarios, but the results translate to more complex situations, for example, when more than two attributes are investigated in a structural model.

Different outcomes, such as factor structure, model fit, threshold and loading parameters, or higher-order moments were not covered herein and remain a route for further research. Moreover, the relationship among different response styles and the effect of multiple response styles at a time may be of interest in future studies. Apart from that, this study focused on the effect of ignoring response styles in raw score-analyses; whether and how response styles can be controlled using appropriate (model-based) approaches is a different question (see, e.g., Wetzel et al., 2016).

In summary, the present results suggest that

the impact of response styles in applied settings is probably better described by a molehill than a mountain. The analyses demonstrated the importance of reverse-keyed items to control for the negative influence of ARS. The future will show whether the gap between the applied camp and the methods camp can be bridged such that practitioners take response styles into account where necessary and that psychometricians develop and refine the tools required to do so.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23. doi:10.1177/0146621697211001

Aichholzer, J. (2013). Intra-individual variation of extreme response style in mixed-mode panel studies. *Social Science Research*, *42*, 957–970. doi:10.1016/j.ssresearch.2013.01.002

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. doi:10.1007/BF02293814

Bentler, P. M. (2016). Covariate-free and covariate-dependent reliability. *Psychometrika*, *81*, 907–920. doi:10.1007/s11336-016-9524-y

Billiet, J. B. & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*, 608–628. doi:10.1207/S15328007SEM0704_5

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

Bolt, D. M. & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*, 814–833. doi:10.1177/0013164410388411

De Boeck, P. & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1–28. doi:10.18637/jss.v048.c01

Eid, M. & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*, 20–30. doi:10.1027//1015-5759.16.1.20

Falk, C. F. & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*, 328–347. doi:10.1037/met0000059

Ferrando, P. J. & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 427–448. doi:10.1348/000711009X470740

Finn, J. A., Ben-Porath, Y. S., & Tellegen, A. (2015). Dichotomous versus polytomous response options in psychopathology assessment: Method or meaningful variance? *Psychological Assessment*, *27*, 184–193. doi:10.1037/pas0000044

Hankin, R. K. S. (2005). Recreational mathematics with R: Introducing the "magic" package. *R News*, *5*(1), 48–51.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101–125. doi:10.1177/014662169602000201

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

Jackson, A. (2012). IPIP Big Five personality test answers [Data file]. doi:10.6084/m9.figshare.96542

Jin, K.-Y. & Wang, W.-C. (2014). Generalized IRT models for extreme response style. *Educa-*

tional and Psychological Measurement, *74*, 116–138. doi:10.1177/0013164413498876

Johnson, T. R. & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics*, *35*, 92–114. doi:10.3102/1076998609340529

Khorramdel, L. & von Davier, M. (2014). Measuring response styles across the Big Five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, *49*, 161–177. doi:10.1080/00273171.2013.866536

Kiefer, T., Robitzsch, A., & Wu, M. (2015). TAM: Test analysis modules (Version 1.11-0). Retrieved from https://CRAN.R-project.org/package=TAM

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

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Paunonen, S. V. & LeBel, E. P. (2012). Socially desirable responding and its elusive effects on the validity of personality assessments. *Journal of Personality and Social Psychology*, *103*, 158–175. doi:10.1037/a0028165

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

R Core Team. (2014). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Vol. 4. Contributions to Biology and Problems of Medicine* (pp. 321–333). Berkeley, CA: University of California. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200512895

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin*, *63*, 129–156. doi:10.1037/h0021888

Savalei, V. & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research*, *49*, 407–424. doi:10.1080/00273171.2014.931800

Schimmack, U., Böckenholt, U., & Reisenzein, R. (2002). Response styles in affect ratings: Making a mountain out of a molehill. *Journal of Personality Assessment*, *78*, 461–483. doi:10.1207/S15327752JPA7803_06

Trautmann, H., Steuer, D., Mersmann, O., & Bornkamp, B. (2014). truncnorm: Truncated normal distribution (Version 1.0-7). Retrieved from http://CRAN.R-project.org/package=truncnorm

Van Vaerenbergh, Y. & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*, 195–217. doi:10.1093/ijpor/eds021

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods*, *18*, 320–334. doi:10.1037/a0032121

Weijters, B., Cabooter, E. F. K., & 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. doi:10.1016/j.ijresmar.2010.02.004

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of ac-

quiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, *34*, 105–121. doi:10.1177/0146621609338593

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

Wetzel, E., Böhnke, J. R., & Rose, N. (2016). A simulation study on methods of correcting for the effects of extreme response style. *Educational and Psychological Measurement*, *76*, 304–324. doi:10 . 1177 / 0013164415591848

Wetzel, E. & Carstensen, C. H. (2015). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*. Advance online publication. doi:10.1027/1015-5759/a000291

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*, 178–189. doi:10.1016/j.jrp.2012.10.010

Appendix

Varying the Number of Categories and the Impact of the Weights

Researchers are regularly confronted with the question about the optimal number of categories for their instrument, and there is ample research trying to answer this question from different perspectives (e.g., Finn, Ben-Porath, & Tellegen, 2015). At first sight, it seems plausible to add to this literature with the present simulation study. Therefore, the number of categories was varied between 3 and 7 in a simulation focusing on the effect of ARS on Cronbach's alpha with a minimal setup ($\mu_{\text{ARS}} = 0; \sigma^2_{\text{ARS}} = 1; \rho_{1,\text{ARS}} = 0$). At first sight, more categories led to less bias ($b = -.01$). However, there is a confounding effect between the number of categories and the weights in the **B** matrix: With increasing categories, the content-related weights increase in size (e.g., from $[0, 1, 2]'$ to $[0, 1, 2, 3]'$), whereas the ARS weights are always fixed to zeros and ones (e.g., $[0, 0, 1]'$ and $[0, 0, 1, 1]'$). Thus, more categories did not lead to less bias for substantive reasons, but simply for the reason that the relative size of the ARS weights (i.e., the impact of ARS) decreased. For illustration, the ARS weights were set to $^{K}/_{2}$ in a follow-up simulation (e.g., $[0, 0, 1]'$ and $[0, 0, 1.5, 1.5]'$). Then, the effect of the number of categories on bias changed sign ($b = .01$). Thus, the confounding effect between the number of categories and the response style weights makes it impossible to draw conclusions about the relationship between the number of categories and response styles.

This confounding effect would play a role in all simulations reported herein. Moreover, it is present whenever weights for response styles are used and different numbers of categories are compared.

Apart from that, the described mechanism also applies to the comparison of different scoring schemes for a fixed number of categories. For example, weights of $(2, 1, 0, 1, 2)'$ for ERS instead of $(1, 0, 0, 0, 1)'$ may be seen just as valid. However, increasing the weights would also artificially increase the impact of ERS making the results incomparable. This is also mirrored in the fact that, in the empirical illustration, σ^2_{ERS} dropped from a value of 1.02 to 0.35 if the ERS weights were changed to $(2, 1, 0, 1, 2)'$.