Article

A Simulation Study on Methods of Correcting for the Effects of Extreme Response Style

Educational and Psychological Measurement
2016, Vol. 76(2) 304–324

DOI: 10.1177/0013164415591848
epm.sagepub.com
SAGE

Eunike Wetzel[1,2], Jan R. Böhnke[3], and Norman Rose[2]

Abstract

The impact of response styles such as extreme response style (ERS) on trait estimation has long been a matter of concern to researchers and practitioners. This simulation study investigated three methods that have been proposed for the correction of trait estimates for ERS effects: (a) mixed Rasch models, (b) multidimensional item response models, and (c) regression residuals. The methods were compared with respect to their ability of recovering the true latent trait levels. Data were generated according to a unidimensional model with only one trait, a mixed Rasch model with two populations of ERS and non-ERS, and a two-dimensional model incorporating a trait and an ERS dimension. The data were analyzed using the same models as well as linear regression where the trait estimate is regressed on an ERS score and the resulting residual is considered the corrected trait estimate. Over all conditions, the two-dimensional model achieved the best trait recovery, though the difference to the unidimensional model was rather small. Mixed Rasch models were in general inferior to the other correction methods. When the trait and ERS showed no to weak correlations, ERS appeared to have a minor impact on trait estimation.

Keywords

[1]University of Konstanz, Konstanz, Germany
[2]Eberhard Karls University Tübingen, Tübingen, Germany
[3]Mental Health and Addiction Research Group (MHARG), Hull York Medical School and Department of Health Sciences, University of York, York, UK

Corresponding Author:
Eunike Wetzel, University of Konstanz, Department of Psychology, Box 31, 78457 Konstanz, Germany.
Email: eunike.wetzel@uni-konstanz.de

response styles, extreme response style, mixed Rasch models, multidimensional item response models, regression residuals

One concern in psychological assessment based on self-report questionnaires is the potential influence of response styles on item responses. Response styles refer to systematic individual differences in response scale use that are characterized by a systematic preference or avoidance of certain response categories. If response styles exert an influence on item responses over and above the respondent's true latent trait level, trait estimates may be biased and in consequence inferences based on these trait estimates may be wrong (e.g., Austin, Deary, & Egan, 2006). This concern was already voiced rather dramatically by Cronbach (1950; p. 21) who said, ''The writer concludes that as a general principle, the tester should consider response sets an enemy to validity.'' Polytomous rating scales that are ubiquitous in the assessment of personality traits, attitudes, and many other constructs appear to be especially susceptible to response styles. Common response styles in self-report questionnaires applying polytomous rating scales include extreme response style (ERS), a preference for extreme response categories, and acquiescence response style, a preference for categories stating agreement (e.g., Austin et al., 2006; Baumgartner & Steenkamp, 2001; Bolt & Johnson, 2009). To quantify the influence of response styles on item responses, Wetzel and Carstensen (2015) computed differences in pseudo-R^2 values (Nagelkerke, 1991) between models ignoring response styles and two-dimensional item response models taking response styles into account. For ERS, they found an average incremental variance explanation of 25% across the 30 facets of the German Revised NEO Personality Inventory (NEO-PI-R; Ostendorf & Angleitner, 2004). For acquiescence response style, the incremental variance explanation was lower with an average of 4%. Furthermore, respondents appear to be largely consistent in their use of response styles across the scales in one instrument (Weijters, Geuens, & Schillewaert, 2010a; Wetzel, Carstensen, & Böhnke, 2013) as well as over time (Weijters, Geuens, & Schillewaert, 2010b; Wetzel, Lüdtke, Zettler, & Böhnke, 2015). The existence and persistence of response styles raises the question of how to account for response styles in the analysis of questionnaire data or, in other words, how to correct trait estimates for response style effects.

Several methods of correcting for response style effects have been suggested in the literature. The aim of this article is to evaluate three methods for the correction of ERS (mixed Rasch models, multidimensional item response models, and regression residuals) with respect to their ability of recovering the true latent trait levels using a simulation study. The article is structured as follows: First, the three methods will be explained in detail. Second, we present the design of the simulation study, including the choice of independent and dependent variables, the data generation procedures, and the statistical analyses. Third, the trait recovery achieved by the three methods will be reported. Last, the implications of these results for the correction of trait estimates for response style effects will be discussed. This study will

focus on one response style, namely ERS, because two of the three methods we evaluate have been suggested specifically for this response style (Bolt & Johnson, 2009; Bolt & Newton, 2011; Rost, Carstensen, & von Davier, 1997) and it also appears to be the most important response style in terms of its variance explanation incremental to the trait (Wetzel & Carstensen, 2015).

Methods of Correcting for Extreme Response Style Effects

In this simulation study, three methods that have been suggested to correct individual trait estimates for ERS will be evaluated: (a) mixed Rasch models, (b) multidimensional item response models, and (c) using regression residuals as trait estimates. In the following, the underlying rationale of these three methods and their application to response styles will be explained.

Mixed Rasch Models. Mixed Rasch models (Rost, 1990, 1991) are extensions of unidimensional Rasch models (e.g., Rasch, 1960) to multiple latent classes. In mixed Rasch models, item and person parameters are estimated separately for each latent class. This allows the investigation of qualitative differences between the latent classes as well as the investigation of quantitative differences within each latent class. In the case of rating scale data as we simulated in this study, the mixed partial credit model (Rost, 1991; von Davier & Rost, 1995) can be applied which adds class-specific parameters to the partial credit model (PCM) developed by Masters (1982).

In the mixed PCM the probability of a response in category x on item i with $m + 1$ response categories ($x = 0, \ldots, m$) depends on two latent variables: (a) the continuous latent trait θ and (b) the discrete latent class C, with $c = 1, \ldots, q$. The test takers' individual trait levels θ_v and their latent class membership are the person parameters in the mixed PCM. The effect of the latent class membership is taken into account by class-specific threshold parameters τ_{icj}. Hence, there are k class-specific model equations for the response category probabilities $P_{C=c}(X_i = x|\theta)$ of each item, which can be written as

$$P_{C=c}(X_i = x|\theta) = \frac{\exp \sum_{j=0}^{x} (\theta - \tau_{icj})}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} (\theta - \tau_{ick})}. \quad (1)$$

As the response categories are exhaustive and disjunctive within each latent class c it follows that $\sum_{k=0}^{m} P_{C=c}(X_i = x|\theta) = 1$, for all $c = 1, \ldots, q$ and $i = 1, \ldots, n$. For identification purposes, it is defined that $(\theta - \tau_{ic0}) \equiv 0$. The individual posterior probabilities $P(C = c \mid \boldsymbol{X}_v = \boldsymbol{x})$ can be used to estimate respondents' class membership. The unconditional latent class probabilities $P(C = c) = \pi_c$ are also estimable parameters in the

mixed PCM describing the distribution of C. Since the latent classes are exhaustive and disjunctive it follows that $\sum_{c=1}^{C} \pi_c = 1$.

Mixed PCMs have been used in several studies to differentiate latent classes of participants that differed systematically in their response scale use. These latent classes are interpreted as response style groups using the distribution of threshold parameters in each latent class. For example, Rost et al. (1997) showed that two response style groups (ERS and non–extreme response style [NERS]) existed in the Big Five as assessed by the German NEO Five-Factor Inventory (NEO-FFI; Borkenau & Ostendorf, 1993). In the ERS class, threshold parameters were close together, indicating that extreme categories had the highest probability of being chosen already at moderate trait levels. In contrast, in the NERS class, threshold parameters were widely spaced, indicating that extreme categories only had the highest probability of being chosen at very high or very low trait levels. The same pattern was found in the English NEO-FFI (Costa & McCrae, 1992) by Austin et al. (2006) and in several instruments assessing other constructs, such as a leadership performance scale (Eid & Rauber, 2000). According to Rost et al. (1997, p. 331), ''The trait parameter of the mixed Rasch model is automatically corrected for the effects of a response set on the sum scores'' since person parameters are estimated separately for each latent class and class-specific item parameters (signaling the response style) are therefore taken into account. Rost et al. (1997) argued that the resulting person parameters should then be on the same scale and consequently comparable between latent classes. Wetzel et al. (2013) illustrated this by contrasting trait estimates on the NEO-PI-R facets between a (one-class) PCM and a two-class mixed PCM where the latent classes were identified as ERS and NERS. Compared with the one-class PCM, trait estimates were contracted (i.e., less extreme) in the ERS class and extended (i.e., more extreme) in the NERS class for respondents with equal sum scores and thus equal trait estimates in the one-class PCM. This indicates a correction of trait estimates for response style effects, though it has not been investigated how accurately the trait estimates from mixed PCMs reflect the true latent trait levels.

Multidimensional Item Response Models. Multidimensional item response models are extensions of unidimensional item response models to two or more latent traits. The original multidimensional Rasch model for dichotomous items was presented by Rasch (1961). The unidimensional PCM (Masters, 1982) was extended to multidimensional data by Kelderman (1996). The multidimensional PCM defines the conditional response category probabilities $P(X_i = x \mid \boldsymbol{\theta})$ of a response in category x ($x = 0, \ldots, m$) of item i as a function of $q = 1, \ldots, s$ latent trait dimensions $\boldsymbol{\theta} = \theta_1, \ldots, \theta_s$. That is

$$P(X_i = x|\theta) = \frac{\exp\left[\sum_{j=1}^{x}\left(\sum_{q=1}^{s}\omega_{qij}\theta_q - \tau_{ij}\right)\right]}{\sum_{k=0}^{m}\exp\left[\sum_{j=0}^{k}\left(\sum_{q=1}^{s}\omega_{qij}\theta_q - \tau_{ij}\right)\right]}, \quad (2)$$

where $\sum_{j=1}^{0}(\bullet) \equiv 0$. As in the mixed PCM, τ_{ij} indicates the threshold parameter between two adjacent response categories $x = j-1$ and $x = j$. In Equation (2), ω_{qij} denotes an indicator variable which takes the value 1 if the response to category j of item i indicates dimension q and the value 0 if it does not. This assignment of items to dimensions has to be prespecified by the researcher. In the case of $s = 1$ and $\omega_{qij} = 1$ for all items and response categories in Equation (2), the unidimensional PCM results.

The application of multidimensional models to take into account response style effects was suggested by Bolt and Johnson (2009) who proposed modeling both the trait of interest and ERS in a multidimensional extension of Bock's (1972) nominal response model. Since both dimensions are modeled simultaneously, respondents' standing on the dimensions is estimated using information from both dimensions. Thus, Bolt and Johnson (2009) argued, the trait estimate derived from this model is corrected for response style effects. In the context of the multidimensional PCM described above, the substantive trait and ERS can be specified as follows. Let θ_1 be the latent trait and θ_2 the latent ERS. Then the indicator variables are $\omega_{1ij} = 1$ for all $j = 0, \ldots, m$. Only the responses of the two extreme response categories $X_i = 0$ and $X_i = m$ of each item i are indicative for θ_2. Therefore, $\omega_{2i0} = \omega_{2i(m)} = 1$ and $\omega_{2ij} = 0$ if $0 < j < (m)$.

Bolt and Newton (2011) extended this approach by adding a second trait dimension to the model and by modeling the ERS dimension on the combined items from both trait dimensions. According to Bolt and Newton, this method allows a better differentiation between the traits and ERS, which should lead to more accurate trait estimates. In a simulation, they analyzed the correlation between the generated trait level and the trait estimates from two-dimensional and three-dimensional models, which included either one trait and ERS, two traits, or two traits and ERS. Correlations between true trait levels and trait estimates for the first trait were high (at or above .90) for all models and did not differ notably between models that included one trait and ERS (average correlation .92) or two traits and ERS (average correlation .93).

However, the simulation study by Bolt and Newton (2011) did not investigate the possible influence of test length and sample size on the accuracy of trait recovery. In the present study, these factors will be taken into account. Thus, the trait estimates from two-dimensional PCMs, in which the trait and ERS dimensions are modeled using the same items, will be compared with the generated latent trait values for different test length and sample size conditions.

Regression Residuals. A third method that has been suggested for the correction of trait estimates for response style effects is based on linear regression (Baumgartner & Steenkamp, 2001; Webster, 1958; Weijters, Schillewaert, & Geuens, 2008). In this method, the procedure is to first compute a regression of the trait score on ERS, which in the simplest case could be a count of the extreme responses endorsed. Next, the regression residual is computed by subtracting the expected trait score from the observed trait score. This residual is then used as the new ''corrected'' trait estimate from which response style effects have been partialled out. For example, Webster (1958) illustrated how reliabilities decrease when scale scores have been adjusted with this method. Baumgartner and Steenkamp (2001) demonstrated that correlations between scales can both increase or decrease when residualized scores are used compared with observed scores, depending on the relationship between the scales and between different response styles.

This method based on regression residuals is the most convenient of the methods investigated here since it does not necessitate the estimation of item response models. However, several problems may exist with this approach, one of them being that only a linear relationship between the trait and ERS can be accounted for (see discussion). In our study, regression residuals will be compared with the true trait levels to investigate the trait recovery they achieve.

Aim of This Study and Hypotheses

The aim of this study is to compare the methods of correcting for ERS described above regarding their ability to recover the true latent trait levels. It is important to note that the three models differ substantially. In the mixed Rasch model, it is assumed that persons with a tendency toward extreme responses are a qualitatively distinct subpopulation from persons who do not have this tendency. In the multidimensional PCM, each person's ERS is formalized as a continuous latent variable. Hence, each test taker has a value on this latent trait representing his or her tendency to endorse the highest or lowest response category. The use of the regression residual assumes that regressing test scores on response style indicators will remove trait-unrelated variability caused by response styles from the test scores. Taking the residual as the adjusted test score assumes that the underlying trait of interest and the responses style are uncorrelated. If this assumption does not hold, then reliable variance of the latent trait estimate will also be partialled out. This is an important difference compared with the other two models since the mixed Rasch model as well as the multidimensional Rasch model allow for stochastic dependencies between the latent trait variable and ERS.

As the approaches differ in their assumptions they are expected to vary in their suitableness in real applications depending on the true data generating mechanism. Unfortunately, the latter is unknown in most cases. To study the performance of the different methods more generally, we simulated data using different population models and applied the three methods to all of them. For purposes of comparison, the

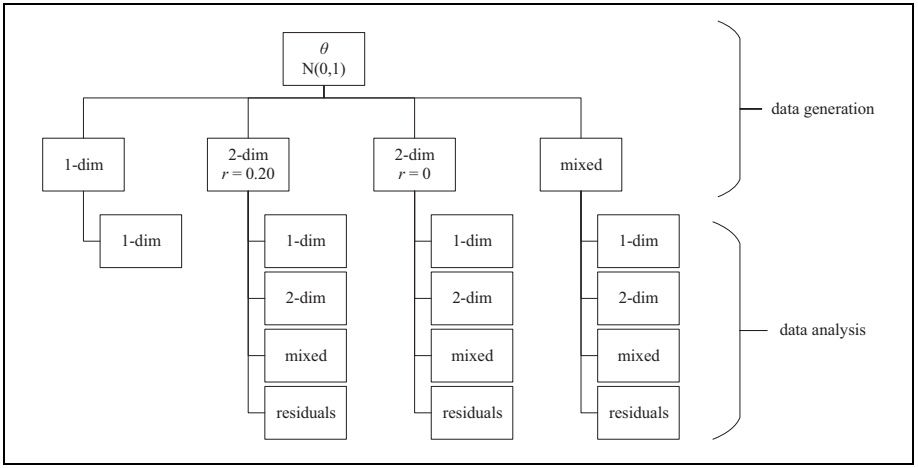

Figure 1. Simulation design with data generation and data analysis factors. The factors test length and sample size are not shown here.
Note. 1-dim = 1-dimensional; 2-dim = 2-dimensional; mixed = mixed partial credit model.

same items will be used for modeling the trait and ERS, though modeling ERS based on other items than the trait items is of course also possible (see, e.g., Weijters et al., 2010b). We hypothesize that both methods based on trait estimates from item response models will achieve a better trait recovery than regression residuals. Since there are no previous studies evaluating the trait recovery in mixed PCMs and the simulation study on multidimensional models (Bolt & Newton, 2011) only considered a limited number of conditions, the comparison of these two methods is exploratory.

Method

The basic structure of the simulation was to first generate true trait levels θ and item responses for these θ under different true models. The second step was to analyze the data with the three correction methods in order to obtain estimated trait levels $\hat{\theta}$ for different conditions of test length and sample size. This structure is illustrated in Figure 1 and will be explained in detail in the following. Then, we will describe the analyses conducted to investigate the degree of trait recovery by each of the correction methods.

Simulation Design

The design of the simulation study included three data generation methods and four data analysis methods. Three of the data analysis methods corresponded to the correction methods described above, namely, mixed PCM, multidimensional PCM, and regression residuals. In addition, the data were analyzed using a unidimensional (one-class) PCM to investigate the degree of trait recovery when the presence of ERS in

the data is ignored. The other two factors in the simulation design were test length and sample size. The factor levels were chosen to cover many realistic applications of psychometric tests. The test length was indicated by the number of items in the test (5, 10, 25, and 50 items). Three different sample sizes were considered (200, 500, and 2,000).

Data Generation Methods. The data were generated to reflect responses on a polytomous rating scale with four response categories (e.g., *strongly disagree, disagree, agree, strongly agree*) since this response format is widely used in psychological assessment with self-report questionnaires. Data were simulated for three true models: (a) a unidimensional PCM without ERS (in the following *1-dimensional*), (b) a mixed PCM (in the following denoted by *mixed*) in which data were simulated from two populations (persons with ERS and persons without ERS, i.e., NERS), and (c) a multidimensional model in which both the trait and ERS influenced item responses (in the following *2-dimensional*). The mixed PCM and the 2-dimensional model correspond to the two correction methods that are based on analyses using these models. Data were additionally generated based on a 1-dimensional model as a baseline model to obtain benchmark values for the trait recovery measures (see Figure 1).

For each of the sample size conditions the corresponding number of θ were drawn from a standard normal distribution. Then, item responses were generated under the different true models. Data generation under the 1-dimensional and 2-dimensional models was conducted for the whole sample. Previous research indicates that some traits may be weakly to moderately associated with ERS (e.g., $r = .22$ with extraversion and conscientiousness; Austin et al., 2006). Thus, for the 2-dimensional data, two variations were simulated: one with a correlation of 0 between the trait and ERS and one with a correlation of .20 between the trait and ERS.

For data generation under the mixed PCM, the sample was halved and item responses were then generated separately for the two halves. We therefore assumed that the population consisted of 50% ERS and 50% NERS. One set of item location parameters was drawn from a standard normal distribution for the whole sample. Threshold parameters for the two populations were computed as the sum of the respective item location parameter and a deviation factor sampled randomly from distributions with differing means (-0.5, 0, 0.5 for ERS and -2.75, 0, 2.75 for NERS). This ensured that the resulting response data were characteristic for ERS and NERS classes.

Data were generated in R (R Core Team, 2013). In all conditions the number of response categories was four. Thus, ERS would be characterized by a high proportion of responses in the categories 0 (e.g., *strongly disagree*) and/or 3 (e.g., *strongly agree*). A total of 100 replications were conducted per sample size $\times$ test length condition. For each replication, new θ and new item (threshold) parameters were drawn, though response data were generated according to the three true models based on the same set of θwithin each condition.

Data Analysis Methods. Response data simulated under the 2-dimensional model and mixed PCM were analyzed using all three correction methods. In addition, the data were analyzed with the 1-dimensional PCM to investigate the bias that occurs when data containing ERS are analyzed without taking ERS into account. Response data generated under the 1-dimensional model were only analyzed with the 1-dimensional PCM to obtain a baseline for the correlation between θ and $\hat{\theta}$ as well as for the bias measures (see Figure 1).

Parameter estimation for the 1-dimensional and 2-dimensional PCMs was conducted in ConQuest (Wu, Adams, Wilson, & Haldane, 2007). Mixed PCMs were estimated in the software for multidimensional discrete latent trait models (mdltm; von Davier, 2005, 2008). Linear regressions for obtaining the regression residuals were computed in R (R Core Team, 2013). For all methods, expected a posteriori (EAP) estimates (Bock & Aitkin, 1981) were used as trait (and if applicable ERS) estimates. For the correction method mixed PCM, the EAPs estimated for the latent class with the highest probability of class membership were used. For the correction method regression residuals, the observed trait estimates used in the regression were based on the EAPs derived from a 1-dimensional analysis of 2-dimensional or mixed data. The predictor in the regression was a centered ERS index based on the frequency of extreme responses. The resulting residuals were standardized to make them comparable with θ and the EAPs resulting from the other correction methods.

Comparison of Methods Regarding Trait Recovery

Three measures were computed to investigate the degree of trait recovery by each of the three correction methods: (a) the correlation between the true trait level θ and the estimated trait level $\hat{\theta}$, (b) the mean absolute bias, and (c) the mean squared error.[1] The mean absolute bias was defined as $MAB(\hat{\theta}) = E\left[\left|\theta - \hat{\theta}\right|\right]$ and the mean squared error was defined as $MSE(\hat{\theta}) = E\left[\left(\theta - \hat{\theta}\right)^2\right]$ to quantify the efficiency of the person parameter estimation method.

The correlation $r(\theta, \hat{\theta})$ from the 1-dimensional data analyzed with the 1-dimensional model provides a benchmark of the expected correlation when there is no response style in the data. The correlations found for the three correction methods can then be compared against this benchmark as well as against each other. The same is true for the bias measures.

Results

All the estimated models converged. In the following, the results on the recovery of the true trait levels will first be reported regarding the correlation $r(\theta, \hat{\theta})$ and second regarding the bias measures.

Table 1. Simulation Results for $r(\theta, \hat{\theta})$.

(a) Data generation method: 1-Dimensional

			Data analysis method	
			1-Dimensional	
		N	*M* (*SD*)	Min; Max
Nr. items	5	200	0.796 (0.025)	0.742; 0.850
		500	0.797 (0.018)	0.757; 0.837
		2,000	0.797 (0.012)	0.764; 0.828
	10	200	0.883 (0.015)	0.844; 0.919
		500	0.882 (0.011)	0.854; 0.910
		2,000	0.882 (0.006)	0.869; 0.894
	25	200	0.948 (0.007)	0.930; 0.963
		500	0.947 (0.005)	0.934; 0.958
		2,000	0.948 (0.003)	0.940; 0.953
	50	200	0.972 (0.004)	0.963; 0.981
		500	0.972 (0.002)	0.968; 0.976
		2,000	0.973 (0.001)	0.970; 0.975

(b) Data generation method: 2-Dimensional, $r(\theta, ERS) = 0.20$

			Data analysis method							
			1-Dimensional		2-Dimensional		Mixed		Residuals	
		N	*M* (*SD*)	Min; Max	*M* (*SD*)	Min; Max	*M* (*SD*)	Min; Max	*M* (*SD*)	Min; Max
Nr. items	5	200	0.734 (0.031)	0.666; 0.820	0.751 (0.032)	0.642; 0.814	0.559 (0.083)	0.312; 0.719	0.723 (0.032)	0.632; 0.793
		500	0.735 (0.022)	0.680; 0.794	0.753 (0.019)	0.699; 0.806	0.577 (0.045)	0.470; 0.689	0.725 (0.022)	0.672; 0.781
		2,000	0.735 (0.014)	0.692; 0.769	0.754 (0.012)	0.718; 0.784	0.573 (0.031)	0.463; 0.647	0.726 (0.014)	0.680; 0.768
	10	200	0.830 (0.021)	0.757; 0.868	0.850 (0.018)	0.796; 0.884	0.729 (0.073)	0.495; 0.868	0.819 (0.022)	0.756; 0.862
		500	0.826 (0.014)	0.788; 0.864	0.845 (0.013)	0.816; 0.874	0.751 (0.057)	0.553; 0.836	0.816 (0.015)	0.780; 0.855
		2,000	0.826 (0.009)	0.800; 0.845	0.846 (0.006)	0.829; 0.858	0.757 (0.043)	0.604; 0.818	0.816 (0.010)	0.790; 0.837
	25	200	0.905 (0.013)	0.876; 0.935	0.913 (0.011)	0.884; 0.938	0.806 (0.062)	0.592; 0.895	0.893 (0.016)	0.830; 0.930
		500	0.903 (0.009)	0.878; 0.925	0.913 (0.007)	0.897; 0.929	0.813 (0.052)	0.532; 0.878	0.891 (0.010)	0.862; 0.914
		2,000	0.903 (0.005)	0.891; 0.914	0.913 (0.003)	0.905; 0.920	0.793 (0.054)	0.612; 0.867	0.893 (0.005)	0.881; 0.909
	50	200	0.934 (0.007)	0.919; 0.948	0.938 (0.007)	0.919; 0.954	0.848 (0.050)	0.620; 0.907	0.920 (0.013)	0.874; 0.947
		500	0.933 (0.004)	0.921; 0.942	0.937 (0.005)	0.925; 0.948	0.826 (0.046)	0.680; 0.893	0.922 (0.008)	0.897; 0.938
		2,000	0.933 (0.003)	0.926; 0.941	0.938 (0.002)	0.932; 0.943	0.810 (0.055)	0.602; 0.883	0.922 (0.005)	0.908; 0.932

(*continued*)

Table 1. (continued)

(c) Data generation method: 2-Dimensional, $r(\theta, ERS) = 0$

			Data analysis method							
			1-Dimensional		2-Dimensional		Mixed		Residuals	
		N	*M* (*SD*)	Min; Max	*M* (*SD*)	Min; Max	*M* (*SD*)	Min; Max	*M* (*SD*)	Min; Max
Nr. items	5	200	0.735 (0.033)	0.651; 0.796	0.750 (0.034)	0.618; 0.807	0.572 (0.075)	0.311; 0.724	0.733 (0.033)	0.642; 0.798
		500	0.733 (0.022)	0.684; 0.786	0.750 (0.021)	0.706; 0.797	0.582 (0.047)	0.404; 0.669	0.732 (0.022)	0.682; 0.785
		2,000	0.733 (0.015)	0.696; 0.772	0.752 (0.013)	0.717; 0.784	0.579 (0.032)	0.494; 0.651	0.733 (0.014)	0.696; 0.772
	10	200	0.830 (0.020)	0.762; 0.874	0.848 (0.018)	0.803; 0.880	0.732 (0.071)	0.503; 0.851	0.828 (0.020)	0.762; 0.876
		500	0.825 (0.014)	0.787; 0.856	0.843 (0.012)	0.817; 0.870	0.753 (0.046)	0.575; 0.824	0.825 (0.014)	0.781; 0.854
		2,000	0.826 (0.008)	0.803; 0.844	0.845 (0.006)	0.829; 0.859	0.754 (0.042)	0.629; 0.818	0.826 (0.008)	0.807; 0.843
	25	200	0.904 (0.012)	0.874; 0.936	0.913 (0.011)	0.889; 0.941	0.807 (0.063)	0.543; 0.902	0.902 (0.013)	0.873; 0.933
		500	0.902 (0.008)	0.880; 0.922	0.913 (0.007)	0.893; 0.929	0.815 (0.043)	0.701; 0.893	0.902 (0.008)	0.880; 0.922
		2,000	0.903 (0.005)	0.890; 0.915	0.913 (0.004)	0.904; 0.922	0.802 (0.041)	0.664; 0.863	0.903 (0.005)	0.892; 0.915
	50	200	0.933 (0.008)	0.913; 0.953	0.938 (0.007)	0.916; 0.954	0.852 (0.039)	0.715; 0.910	0.931 (0.008)	0.909; 0.950
		500	0.932 (0.005)	0.919; 0.944	0.938 (0.005)	0.924; 0.950	0.829 (0.037)	0.715; 0.896	0.932 (0.005)	0.919; 0.944
		2,000	0.933 (0.003)	0.925; 0.940	0.938 (0.002)	0.932; 0.944	0.803 (0.043)	0.680; 0.874	0.933 (0.003)	0.924; 0.940

(d) Data generation method: Mixed Rasch model

			Data analysis method							
			1-Dimensional		2-Dimensional		Mixed		Residuals	
		N	*M* (*SD*)	Min; Max	*M* (*SD*)	Min; Max	*M* (*SD*)	Min; Max	*M* (*SD*)	Min; Max
Nr. items	5	200	0.800 (0.029)	0.725; 0.863	0.805 (0.028)	0.738; 0.868	0.624 (0.092)	0.318; 0.782	0.783 (0.037)	0.654; 0.865
		500	0.801 (0.020)	0.740; 0.843	0.810 (0.018)	0.757; 0.846	0.632 (0.069)	0.468; 0.782	0.784 (0.038)	0.629; 0.829
		2,000	0.803 (0.017)	0.686; 0.835	0.810 (0.014)	0.716; 0.842	0.646 (0.072)	0.475; 0.773	0.790 (0.028)	0.662; 0.832
	10	200	0.867 (0.020)	0.797; 0.904	0.885 (0.018)	0.819; 0.917	0.838 (0.040)	0.688; 0.908	0.860 (0.024)	0.769; 0.908
		500	0.869 (0.017)	0.808; 0.897	0.886 (0.012)	0.842; 0.907	0.857 (0.024)	0.762; 0.895	0.862 (0.027)	0.692; 0.895
		2,000	0.870 (0.011)	0.831; 0.885	0.887 (0.008)	0.861; 0.900	0.868 (0.021)	0.771; 0.891	0.862 (0.019)	0.774; 0.885
	25	200	0.914 (0.011)	0.878; 0.942	0.944 (0.007)	0.925; 0.957	0.906 (0.029)	0.803; 0.944	0.910 (0.015)	0.840; 0.938
		500	0.916 (0.009)	0.895; 0.945	0.946 (0.005)	0.931; 0.956	0.915 (0.019)	0.845; 0.946	0.912 (0.012)	0.880; 0.943
		2,000	0.918 (0.008)	0.889; 0.934	0.946 (0.003)	0.936; 0.952	0.917 (0.021)	0.841; 0.942	0.915 (0.010)	0.857; 0.933
	50	200	0.934 (0.008)	0.911; 0.953	0.970 (0.004)	0.957; 0.980	0.928 (0.025)	0.802; 0.960	0.930 (0.012)	0.873; 0.949
		500	0.935 (0.006)	0.921; 0.952	0.970 (0.002)	0.963; 0.975	0.925 (0.025)	0.824; 0.952	0.932 (0.007)	0.902; 0.946
		2,000	0.935 (0.006)	0.920; 0.948	0.970 (0.002)	0.965; 0.975	0.929 (0.025)	0.805; 0.954	0.933 (0.008)	0.904; 0.948

Note. 1-dimensional = 1-dimensional partial credit model; 2-dimensional = 2-dimensional partial credit model; mixed = mixed partial credit model; residuals = regression residuals; min = minimum; max = maximum.

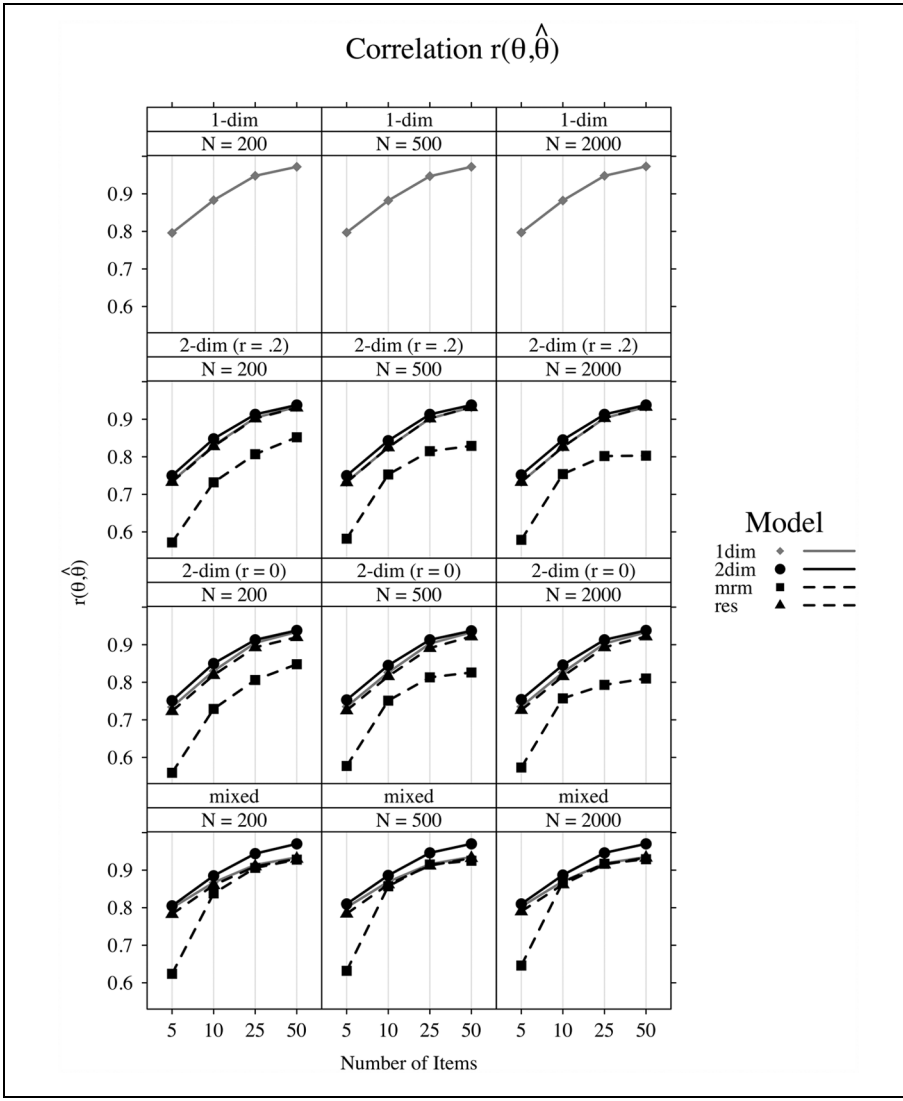

Figure 2. Correlation $r(\theta, \hat{\theta})$.
Note. 1-dim = 1-dimensional; 2-dim = 2-dimensional; mixed = mixed partial credit model.

Recovery of True Trait Levels: $r(\theta, \hat{\theta})$

Correlations between the original true trait levels used for generating data and the estimated trait levels from the analysis of the data indicate how well the ordering of respondents regarding their true trait levels could be recovered in each of the simulation design conditions. Table 1 shows average values across the 100 replications for

the mean, standard deviation, minimum and maximum of the correlation $r(\theta, \hat{\theta})$ separately for each of the four data generation methods. For interpretational convenience the mean correlations are also depicted in Figure 2.

Benchmark. The 1-dimensional analysis of the data that was generated according to the 1-dimensional model provides a benchmark of the correlation when there is no ERS in the data (Part (a) in Table 1). For shorter test lengths (5, 10 items) the mean correlation was .80 for five items and .88 for 10 items across all sample size conditions. For moderate test lengths (25 items) it was .95 and for long test lengths (50 items) it was .97, again with no notable variation across sample size conditions. The variation of the correlations across the 100 replications was small ($SD < .03$ for all conditions). The range of $r(\theta, \hat{\theta})$ across all sample size and test length conditions was .74 to .98. Thus, with 10 or more items, the degree of trait recovery for the 1-dimensional model was very high and close to perfect for the condition with 50 items.

Two-Dimensional Data With r(θ, ERS) = .20. For data that was generated to contain ERS that correlated at .20 with the true trait levels, recovery of the true trait levels was best when the data were analyzed with a 2-dimensional PCM (see Table 1, Part (b), and Figure 2). The total range of $r(\theta, \hat{\theta})$ across all sample size and test length conditions was .64 to .95 with mean values of .75, .85, .91, and .94 for the test lengths 5, 10, 25, and 50 items, respectively. The correlations were only slightly lower for the 1-dimensional analysis and regression residuals. For example, for a test length of 10 items the mean correlation was .83 for a 1-dimensional analysis, .85 for a 2-dimensional analysis, and .82 for a regression residuals analysis of the data for all sample size conditions. In contrast, the mixed PCM yielded a far worse trait recovery with mean values of .58, .75, .81, and .83 for $r(\theta, \hat{\theta})$ for 5, 10, 25, and 50 items and a sample size of 500 in all cases. Furthermore, the results also fluctuated more strongly across the 100 replications with the average $SD = .05$ for mixed PCM and .01 for 2-dimensional.

While sample size did not appear to have a notable effect on trait recovery, test length clearly did (see the increase in correlations with increasing test lengths in Figure 2). The difference in $r(\theta, \hat{\theta})$ between test length = 5 and test length = 50 was .20 for the 1-dimensional analysis, .19 for both the 2-dimensional and regression residuals analysis, and .26 for the mixed PCM analysis. In comparison with the benchmark, the mean correlations were between .03 (10 and 50 items) and .05 (5 items) lower for the 2-dimensional analysis and accordingly differences were slightly larger for the 1-dimensional and regression residuals analysis. The mixed PCM showed the largest differences to the benchmark: between .12 (50 items, sample size 200) and .24 (5 items, sample size 200).

In sum, taking into account ERS by applying a 2-dimensional model or regression residuals led to trait recovery that was only slightly worse than the trait recovery found for data that did not contain ERS. However, ignoring the presence of ERS in the data and instead analyzing it using a 1-dimensional model yielded a similar and

only slightly worse trait recovery. The correction method mixed PCM was not able to adequately recover the true trait levels.

Two-Dimensional Data With r(θ, ERS) = 0. The pattern of results for the 2-dimensional data that was generated with a correlation of 0 between the trait and ERS was practically identical to the one described for the 2-dimensional data with $r(\theta, \text{ERS}) = .20$. Here the 2-dimensional analysis also yielded the best recovery of the true trait levels, followed by the 1-dimensional analysis and the regression residuals. The mixed PCM again did not yield a satisfactory recovery of the true trait levels (see Table 1 and Figure 2, Part (c)).

Mixed Data. The mixed data was generated based on two populations, one ERS and one NERS population. The recovery of the two populations in the analysis of response data with the mixed PCM was overall very good. The mean probability of class membership to the most likely class across 100 replications ranged from .87 for test length 5 to 1.0 for test length 50. Analyzing the mixed data with the mixed PCM yielded a mean correlation of .63, .86, .92, and .93 between true trait level and estimated trait level for the test lengths 5, 10, 25, and 50 items and a sample size of 500 (see Part (d) of Table 1).[2] The mean correlation was similar for a 1-dimensional or regression residuals analysis of the data, except for a test length of 5 items where these two methods achieved better trait recovery than the mixed model ($r = .80$ for 1-dimensional and .78 for regression residuals). Interestingly, the 2-dimensional analysis yielded the best trait recovery with mean correlations of .81, .89, .95, and .97 for the different test lengths and across sample size conditions. These values are practically identical with average benchmark correlations (see above), indicating that a 2-dimensional analysis of data from two populations that differ regarding their use of ERS was able to recover the true trait levels to the same degree as a 1-dimensional analysis of data that did not contain ERS. Note also that the effect of test length on trait recovery was not as strong as in the other data generation methods since the difference between the correlation at a test length of five items and at a test length of 50 items was between .14 (1-dimensional analysis) and .16 (2-dimensional analysis) for a sample size of 500 with the exception of the mixed PCM analysis where the difference was still very large at .30. In sum, the mixed PCM analysis of the mixed PCM data achieved a better trait recovery than the one found for the mixed PCM analysis of other data, but it was still inferior to the trait recovery achieved by the 2-dimensional analysis which was close to benchmark values.

Recovery of True Trait Levels: Mean Absolute Bias

The results were extremely similar for both bias measures (mean absolute bias and mean squared error). Thus, this section will only report the results for the mean absolute difference between θ and $\hat{\theta}$.[3] Plots of the mean absolute bias over the 100 replications in each of the simulation design conditions are depicted in Figure 3.

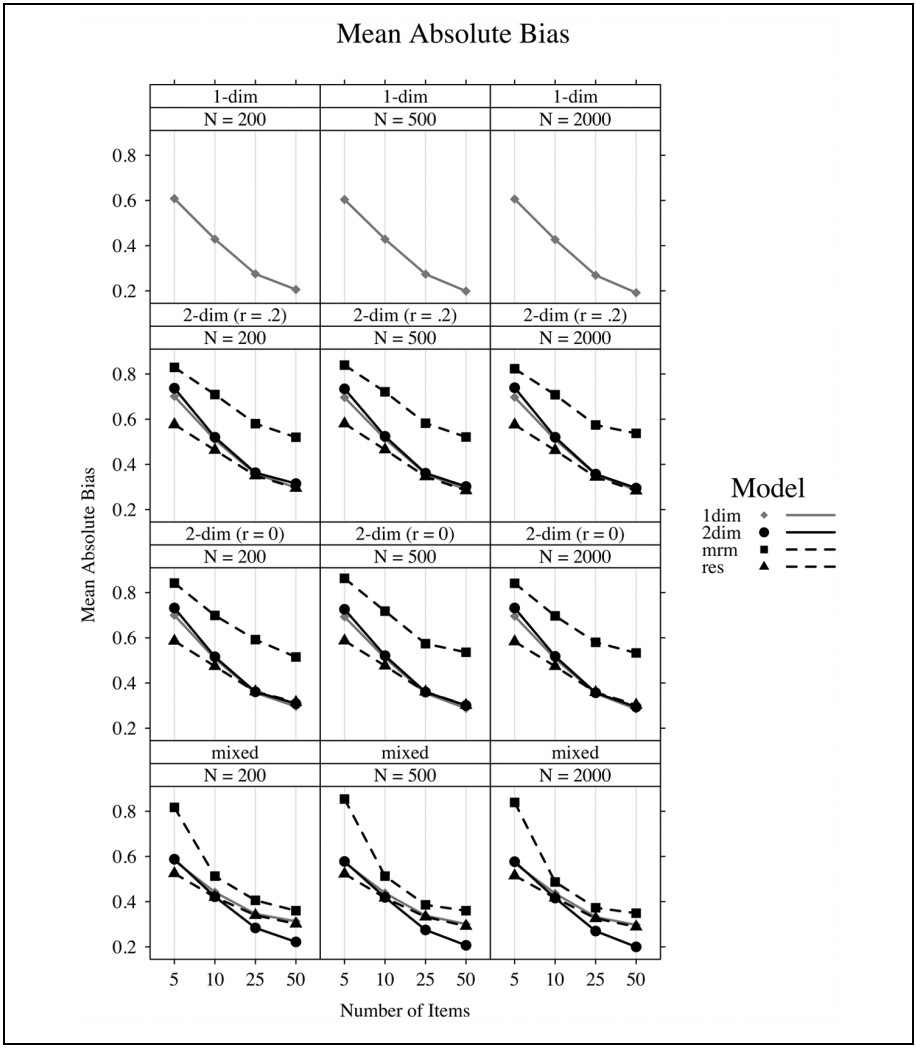

Figure 3. Mean absolute bias.
Note. 1-dim = 1-dimensional; 2-dim = 2-dimensional; mixed = mixed partial credit model.

Benchmark. For the benchmark analysis the mean absolute bias for a sample size of 500 was .60 for a test length of five items, .43 for 10 items, .27 for 25 items, and .20 for 50 items. The mean absolute bias differed only slightly across sample size conditions (e.g., for 50 items it was .21 for $N = 200$, .20 for $N = 500$, and .19 for $N = 2{,}000$). The variation in absolute bias decreased with increasing test lengths and sample sizes (from .05 for 5 items and $N = 200$ to .01 for 50 items and $N = 500$).

Two-Dimensional Data With r(θ, ERS) = .20. For the 2-dimensional data with a correlation of .20 between true trait levels and ERS, the mean absolute bias showed the same general pattern for all analysis methods, namely decreasing values with increasing test lengths and sample sizes (see Figure 3). For test lengths of 25 and 50 items, the 1-dimensional, 2-dimensional, and regression residuals analysis of the data yielded similarly high trait recovery with a mean absolute bias of about .36 for 25 items and .30 for 50 items and a sample size of 500. These bias values are approximately .10 higher than the ones found for the benchmark. For shorter test lengths, regression residuals yielded the best trait recovery with a mean absolute bias of .59 for 5 items and a mean absolute bias of .48 for 10 items (sample size = 500). The mixed PCM analysis of the 2-dimensional data resulted in the worst trait recovery of all analysis methods with a mean absolute bias of .86, .72, .57, and .54 for 5, 10, 25, and 50 items and a sample size of 500. Furthermore, mean absolute bias values varied more strongly across replications with a mean standard deviation of .10 across all test length and sample size conditions.

In sum, the analysis methods 1-dimensional, 2-dimensional, and regression residuals were able to recover the true trait levels to a great extent, though differences to the benchmark still existed. The mixed PCM analysis did not recover the true latent trait levels accurately. In contrast to the correlations described above, the 2-dimensional analysis method was not superior to the 1-dimensional or regression residuals analysis methods with respect to the mean absolute bias.

Two-Dimensional Data With r(θ, ERS) = 0. As depicted above for the correlations, the results for the mean absolute bias for analyses of the 2-dimensional data in which the true trait levels and ERS did not correlate were practically identical to the results found for 2-dimensional data with $r(\theta, \text{ERS}) = .20$ (see Figure 3).

Mixed Data. Analyzing the mixed data with a mixed PCM led to lower mean absolute bias values for test lengths of 10 and above compared with when data generated under a different true model was analyzed with a mixed PCM. For example, for a sample size of 500, the mean absolute bias was .51 for 10 items, .39 for 25 items, and .36 for 50 items. For a test length of 5 items the mean absolute bias was very similar to the one found when 2-dimensional data were analyzed: .85 for the mixed data, .86 for the 2-dimensional data with $r(\theta, ERS) = .20$, and .84 for the 2-dimensional data with $r(\theta, ERS) = 0$. However, the mixed PCM analysis still showed a poorer trait recovery than the other three analysis methods, in particular than the 2-dimensional analysis for which the mean absolute bias for a sample size of 500 was .58, .42, .28, and .21 for 5, 10, 25, and 50 items, respectively. Note that these mean absolute bias values are similar to the ones found in the benchmark analysis. For the 1-dimensional and regression residuals analysis, the mean absolute bias was very similar to the one found when these methods were applied to analyze 2-dimensional data.

In sum, trait recovery for the mixed PCM analysis of the mixed data was improved compared with the mixed PCM analysis of other data, though it was still worse than for the other three analysis methods. As for the correlations, the 2-dimensional

analysis of the mixed data yielded a degree of trait recovery that was similar to the one in the benchmark analysis.

Discussion

Response styles in self-report questionnaires have been a source of concern for researchers and practitioners for a long time, leading to the development of methods to correct for their effects. However, research systematically investigating the ability of these correction methods to recover the true latent trait levels and research on the actual impact of response styles on trait estimates has been scarce. This simulation study showed that when the trait and ERS are uncorrelated or only weakly correlated, ignoring the presence of ERS in the data does not have the detrimental effect on trait estimation that is often assumed. Across all conditions, the analysis of data with ERS using a 2-dimensional model yielded the best trait recovery, though trait recovery for a 1-dimensional model was not substantially worse. The correction method using mixed Rasch models was consistently inferior to the 2-dimensional, regression residuals, and 1-dimensional analysis of the data, even when data were specifically generated according to the mixed Rasch model. In the following, we will discuss the results of our simulation study before noting some limitations of this study and providing some directions for future research.

Discussion of Simulation Results

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 as typically found in empirical applications (Austin et al., 2006; Wetzel & Carstensen, 2015). Stronger relationships between ERS and the trait would presumably have a larger impact on trait estimation. As multidimensional item response models with a latent ERS variable did not yield biased or inaccurate trait estimates, they emerge as the preferred method. In comparison with Bolt and Newton (2011), who found a correlation of .92 between the true trait level and the estimated trait level in a 2-dimensional model with ERS based on a five-item scale, trait recovery was substantially worse in our study with an average correlation of .75. A similar trait recovery as in Bolt and Newton (2011) was achieved in the conditions with 25 items with an average correlation of .91.

The results concerning the application of mixed Rasch models revealed that model-based approaches may even do more harm than good since—for shorter test lengths—this method was even inferior to a 1-dimensional analysis when the data was generated according to the mixed Rasch model. There are several factors that may affect the trait recovery by mixed Rasch models. First, using the most likely trait estimate from mixed Rasch models makes the assumption that all respondents are allocated to the latent class that accurately reflects their response style. If persons are manifestly allocated to the wrong response style class, they receive a trait estimate that is different (and adjusted in the opposite direction) from the one they would receive if

they had been allocated to the right response style class. However, in our study classification accuracy was very good. Furthermore, results were extremely similar between the most likely trait estimate and the average trait estimate (see Note 2). Second, item parameter recovery may affect trait recovery. In our study, item parameter recovery for the mixed PCM was good for test length 5 and very good for test lengths 10 and above (e.g., correlations of .95, .98, and .95 between true and estimated first, second, and third thresholds for 10 items and a sample size of 500). Thus, classification accuracy and item parameter recovery cannot explain the poor trait recovery by the mixed PCM found in our study. One issue that might serve as an explanation is that the proposition that person parameter estimates are automatically corrected for response style effects as put forward by Rost et al. (1997) is really an assumption and not a property of the model. Differences in item parameters between latent classes are assumed to adjust the latent trait in the model for ERS. However, in real applications, it is impossible to distinguish between an appropriate model-based adjustment for ERS and potential measurement noninvariance across unknown latent classes. Equivalent result patterns can occur in both cases when the mixed Rasch model is used. However, if the latent heterogeneity reflects measurement noninvariance, the trait levels are on different metrics and are therefore incomparable across latent classes. There is no quantity in the model that could ensure the invariance of θ scales across latent classes.

Interestingly, the correction method based on regression residuals showed satisfactory trait recovery despite several disadvantages inherent to the method. One disadvantage is that the person parameters are obtained in a two-step procedure. Hence, item response theory standard errors or marginal reliability estimates are not available. Nevertheless, using regression residuals as point estimates is quite easy and, therefore, may be attractive to applied researchers. The major disadvantage of regression residuals is the implicit assumption that the latent trait and ERS are uncorrelated. The fairly good performance of the regression method in our simulation study with a maximum $r(\theta, ERS) = .20$ indicates robustness of this approach for low correlations. However, we cannot extrapolate the robustness of the regression approach to increasing correlations between the latent trait and ERS. As the multidimensional item response model does not require a zero correlation it is theoretically superior. Furthermore, the assumptions of linear regression are violated. It is implausible to expect a linear relationship between the trait and ERS since—when the same items are used for the trait score and ERS—high ERS scores are more likely to occur at very low or very high trait score levels. Or, in other words, respondents with very low or very high trait scores automatically receive a high ERS score since they must have endorsed a certain amount of extreme categories to obtain a very low or very high trait score. This implies a nonlinear relationship between the latent trait and the number of extreme responses, calling into question the suitability of correcting trait levels by using residuals of standard linear regression models.

A major drawback of all methods is their inability to differentiate between persons with extreme trait levels and persons with moderate to high levels of ERS. That is, the more extreme an individual's trait level is, the higher is also the probability of an

extreme response pattern even if the person does not have a general tendency to prefer extreme response categories.

Limitations and Future Directions

For purposes of comparison the same items were used to model the trait and ERS in this study, though for multidimensional PCMs and regression residuals, it is also possible to use separate item sets. An interesting question would be how results on trait recovery change when a heterogeneous item set that has no overlap with the trait items were used to measure ERS as, for example, in Weijters et al. (2010b) and Wetzel et al. (2015). This study only investigated ERS, but of course a number of other response styles exist such as acquiescence or midpoint response style. Thus, before making any general claims about the impact of response styles on trait estimation in data from questionnaires with polytomous rating scales, it would be important to investigate the trait recovery achieved by different correction methods and the 1-dimensional model for other response styles. In addition, to address the practical relevance of correcting for response styles in applied contexts, it would be important to test whether the correction methods achieve an increase in criterion-related validity compared with a 1-dimensional model.

It is also conceivable that the true data generating process underlying rating scale data is different from those considered here. However, we generally do not know the underlying true model in reality. Our results indicate that for both mixed and multidimensional data, which make widely differing assumptions regarding the nature of response styles, trait recovery is best when a 2-dimensional model is applied. Other methods that have been proposed for the modeling of response styles but that were not considered in this study include Böckenholt (2012), who developed a model that divides the response process into subprocesses related to the trait and subprocesses related to response styles, and Rossi, Gilula, and Allenby (2001), who developed a Bayesian hierarchical model for modeling response styles. Thus, future research could investigate trait recovery in these models and multidimensional models based on alternatives to the PCM such as the graded response model.

In sum, this study showed that when ERS and the trait are not or only weakly correlated, the impact of ERS on trait measurement is minor. Of the methods investigated here the 2-dimensional item response model achieved the best trait recovery.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The difference between θ and $\hat{\theta}$ was not used as a bias measure because it was on average 0 due to model identification constraints.
2. As noted above, these results are based on the EAP for the latent class with the highest class membership probability (most likely EAP). Alternatively, one could also use an average EAP based on the posterior probabilities of membership to the two latent classes. If this average EAP is applied for the analyses, results are overall the same. Minor differences occur for the test lengths 5 and 10 items. For example, for 5 items, $r(\theta, \hat{\theta})$ for the most likely EAP was 0.05 lower compared to $r(\theta, \hat{\theta})$ for the average EAP (0.01 lower for 10 items). As classification accuracy improves with increasing test lengths, the difference between the most likely EAP and the average EAP decreases and therefore results on the recovery of θ become practically identical.
3. Results on the mean squared error can be obtained from the first author on request.

References

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. doi:10.1016/j.paid.2005.10.018

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *28*, 143-156.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in 2 or more nominal categories. *Psychometrika*, *37*, 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum-likelihood estimation of item parameters—Application of an EM algorithm. *Psychometrika*, *46*, 443-459. doi:10.1007/Bf02293801

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

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. doi:10.1177/0146621608329891

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

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar [NEO Five Factor Inventory]*. Göttingen, Germany: Hogrefe.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*, 3-31. doi:10.1177/001316445001000101

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*(1), 20-30.

Kelderman, H. (1996). Multidimensional Rasch models for partial-credit scoring. *Applied Psychological Measurement*, *20*, 155-168. doi:10.1177/014662169602000205

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174. doi:10.1007/Bf02296272

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691-692. doi:10.1093/biomet/78.3.691

Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae* [NEO-PI-R: NEO Personality Inventory after Costa and McCrae]. Göttingen, Germany: Hogrefe.

R Core Team (2013). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.

Rasch, G. (1961). *On general laws and the meaning of measurement in psychology*. Berkeley, CA: University of California Press.

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association*, *96*(453), 20-31.

Rost, J. (1990). Rasch models in latent classes—An integration of 2 approaches to item analysis. *Applied Psychological Measurement*, *14*, 271-282. doi:10.1177/014662169001400305

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*, 75-92.

Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324-332). Kiel, Germany: Leibniz Institute for Science and Mathematics Education. Retrieved from http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/ltlc.htm

von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report* (Vol. No. RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2008). The mixture general diagnostic model. In G. R. Hancock & K. M. Samuelson (Eds.), *Latent variable mixture models* (pp. 255-274). Charlotte, NC: Information Age.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 371-379). New York, NY: Springer.

Webster, H. (1958). Correcting personality scales for response sets or suppression effects. *Psychological Bulletin*, *55*(1), 62-64. doi:10.1037/H0048031

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

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, *36*, 409-422. doi:10.1007/s11747-007-0077-6

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