Psychological Test and Assessment Modeling, Volume 65, 2023 (2), 220-258

Detecting and Differentiating Extreme and Midpoint Response Styles in Rating Scales using Tree-Based Item Response Models: Simulation Study and Empirical Evidence

Artur Pokropek[1], Lale Khorramdel[2] & Matthias von Davier[2]

[1] Institute of Philosophy and Sociology of the Polish Academy of Sciences, Poland
[2] TIMSS & PIRLS International Study Center at Boston College, USA

Abstract

An extended the Tree-Based Item Response Models **(**IRTree) approach to detect response styles in rating scale data and to differentiate between extreme response style (extreme response style) and midpoint response style is introduced and validated with a simulation study and using empirical data. The Tree-Based Item Response Models extension is based on the decomposition of rating data into binary pseudo items, which are examined using the multidimensional Item Response Theory modelling framework. Different scenarios, levels, and consistencies of extreme and midpoint response styles are simulated. The approach is further applied to selected scales of the PISA questionnaire. Results show that the approach is a useful and valid tool to detect and correct for response styles in rating scales.

Keywords: large scale assessment, response styles, simulation study

Author Note

This research is financed by the Polish National Science Centre (NCN; Narodowe Centrum Nauki) research grant (2019/33/B/HS6/00937) *Understanding response styles in self-report data: consequences, remedies and sources*.

Modelling latent constructs in sociological, psychological, educational, and workforce assessments and surveys is growing in importance but is not without problems related to measurement error and bias. The most popular way of measuring latent constructs, like attitudes, values or sentiments, is through questionnaires that use a rating scale for respondents to rate themselves. There are several potential problems with rating scales (cf. Khorramdel and von Davier, 2014) – for example, the assumption of interval-scale level, which is seldom true (Rost, 2004; von Davier, 2010a); cultural and group-related differences in how the single response options are interpreted; and fakability in high-stakes measurements (cf. Birkeland, Manson, Kisamore, Brannick, and Smith, 2006; Robie, Brown, and Beaty, 2007). Another is that so-called response styles might be present.

Response styles are defined as construct-irrelevant responses (Paulhus, 1991; Rost, 2004). They are invalid responses that show a specific pattern or style but have nothing to do with the construct of interest to be measured. Hence, response styles can harm the validity (Baumgartner and Steenkamp, 2001; De Jong, Steenkamp, Fox, and Baumgartner, 2008; Dolnicar and Grun, 2009; Weijters, Schillewaert, and Geuens, 2008) and the dimensionality of the measurement (Rost, 2004) by contributing to systematic errors (van Vaerenbergh and Thomas, 2013). response styles are assumed to be broadly stable within single questionnaire administrations (Nunnally, 1967; Javaras and Ripley, 2007) and across longitudinal survey data (Weijters, Geuens, and Schillewaert, 2010). Moreover, gender differences (De Jong, Steenkamp, Fox, and Baumgartner, 2008; Weijters, Geuens, and Schillewaert, 2010) and cultural differences (Bachman and O'Malley, 1984; Buckley, 2009; Bolt and Newton, 2011; Chen, Lee, and Stevenson, 1995; Dolnicar and Grun, 2009; Hamamura, Heine, and Paulhus, 2008; Hui and Triandis, 1989; Van Herk, Poortinga, and Verhallen, 2004) can contribute to response styles. Consequently, response styles cause bias to the survey data and can lead to false inferences about group differences in investigated constructs.

Depending on the number of response categories in the rating scale, different types of response styles might occur (e.g., the tendency toward the midpoint of the scale, the tendency toward the extremes of the scale, acquiescence or the tendency to agree even to contrary statements, the tendency to avoid specific categories, etc.). There are multiple reasons for response styles. They can result from a problem understanding the item content (e.g., low reading ability or ambiguous, inconsistent, or complex statements), a consequence of a lack of test-taking motivation or acceptance for the assessment, or simply fatigue effects toward the end of the assessment. Low motivation is especially problematic in low-stakes assessments, where the test results have no consequences for single respondents.

To ensure higher validity that allow for fair group comparisons, data should first be tested for response styles and corrected for them (if present). A promising line of approaches to test and correct response styles are Tree-Based Item Response Models (IRTree models). Böckenholt (2012) proposed an IRTree model for single questionnaire scales in which responses to a rating scale are decomposed into multiple response subprocesses represented by binary pseudo items. Each examined response styles is represented by a pseudo item and can be modelled using simple structure Item Response Theory models. This approach and its extension to multiple scales (Khorramdel and von Davier, 2014; Plieninger and Meiser, 2014; von Davier and Khorramdel, 2013) have already been applied to empirical data using measures of the Big Five personality scales (Khorramdel and von Davier, 2014; von Davier and Khorramdel, 2013). Moreover, Khorramdel, von Davier and Pokropek (2019) recently proposed a comprehensive approach which combines a multidimensional IRTree model with a mixture distribution Item Response Theory model to examine different latent classes of respondents with different response styles and behaviours. They applied their mixture IRTree modelling approach to the Programme for the International Assessment of Adult Competencies (PIAAC) data and illustrated that different response styles and behaviours could be related to external variables such as response time and other process data (omitted responses and a number of clicks/actions in computer-based items).

While several methodologies, such as Henninger's sum-to-zero constraint for varying thresholds within the Partial Credit Model (Henninger, 2018) and Tutz's approach of introducing finite mixtures (Tutz et al., 2018) have significantly advanced our ability to account for extreme and mid-response styles, the focus of this paper is on the extension of IRTree-Based Item Response Models. It's crucial to note that direct comparisons among these models can be challenging due to differing theoretical assumptions about the response processes. Each approach operates under unique assumptions about how respondents engage with rating scales. However, our work aims not to compare these different approaches directly. Instead, we seek to explore the potential of enriching IRTree models to handle varied response styles, thereby contributing to the array of tools available for analyzing and interpreting response patterns in survey-based research.

The different studies utilising IRTree approaches provided promising results and indicated that these approaches could be used to test and correct for response styles in rating data. However, validity studies are still sparse. To our knowledge, there are only two published studies using extraneous criteria to validate IRTree approaches: Plieninger and Meiser (2014) use academic grades and the relationship between self-concept and reading performance to prove the usefulness of their proposed approach, and Khorramdel, von Davier and Pokropek (2019) utilise response times and process data and relate them to different response styles behavior. However, these studies provide only indirect pieces of evidence for the validity of IRTree approaches, and only the first study (Plieninger and Meiser, 2014) focusses on Böckenholt’s initial IRTree model.

Therefore, one goal of the current study is to further examine the validity of the Böckenholt approach in two ways: First, a simulation study is conducted testing the power and Type I error of the approach; second, the approach is applied to empirical data coming from the background questionnaire of the PISA 2012 main survey. But this paper is not limited to testing Böckenholt’s approach. This paper also introduces a model extension which allows the differentiation between extreme response style and midpoint response style. In the following, a more detailed description of the applied IRT approach is provided, and the simulation and empirical studies are illustrated.

Methods

The method to detect response styles in the simulated and empirical data is based on Böckenholt's (2012) approach. Rating data from 5-point rating scales are decomposed into multiple response subprocesses using the same BPI coding as described by von Davier and Khorramdel (2013) as well as Khorramdel and von Davier (2014). The binary pseudo items are coded to reflect an extreme response style, a midpoint response style, and construct-related responses, and modeled through unidimensional and multidimensional item response theory models. Those models are equivalent to confirmatory factor multidimensional models for categorical data with small differences in parametrisation (Asparouhov & Muthén, 2020). For the simulation study, data are simulated to reflect responses to a 5-point rating scale. The rating scale is one of the most common scales used in large-scale surveys and is frequently used by sociologists, psychologists, and educational scientistsIn the empirical study, data from PISA 2012 student questionnaire scales administered with a 5-point rating scale were selected.Decomposing scales with higher, and in some situations lower, numbers of categories is possible but requires some additional consideration and research on the processes behind the decomposition. For those interested in using a 7-point rating scale, we recommend the works by Spratto, E. M., Leventhal, B. C., & Bandalos, D. L. (2021) and Plieninger, H., & Meiser, T. (2014), who provide detailed guidance on decomposing such scales and fitting IRTree models. Plieninger and Meiser (2014) also offer guidance for 4-point scales. Additionally, Böckenholt (2017) provides an extensive discussion on the decomposition of a 6-point rating scale. All IRT models applied in this study were estimated by applying the mixture general diagnostic modelling framework (MGDM; von Davier, 2008, 2010b), which allows the specification of a discrete mixture model with a hierarchical component (von Davier, 2010b) using the software *mdltm* (von Davier, 2005) for multidimensional discrete latent traits models. The software provides marginal maximum likelihood estimates obtained using customary expectation-maximisation methods[1].

[1] The code used for data simulation and analyses in our study can be provided upon request. Please contact the first author directly for this information.

Decomposition of Rating Data into binary pseudo items

The simulated and empirical rating data were decomposed into binary pseudo items following the procedure illustrated in Khorramdel and von Davier (2014) as well as von Davier and Khorramdel (2013). The 5-category responses to all items were decomposed assuming three latent variables. Thus, every questionnaire item was recoded into three different kinds of binary pseudo items (see Table 1):

- one accounting for extreme positive and negative responses (*e*-items; responses to extreme categories coded as 1, otherwise as 0 or missing value)
- one accounting for the middle category (*m*-items; responses to the middle category coded as 1, otherwise as 0)
- one for only positive (including extreme and nonextreme) responses (*d*-items; negative responses coded as 0, responses to the midpoint category coded with a missing value, moderate and extreme positive responses both coded as 1).[2]

The score based on *e*-items represents a possible measure of extreme response style, and the score based on *m*-items represents a possible measure of response styles. The scalewise scores based on *d*-items, on the other hand, aim to model the trait-relevant responses that are not biased by extreme and midpoint response styles. If the middle category of the rating scale was chosen, *e*-items and *d*-items received a missing value code because no dependencies were implied between binary pseudo items *e*, *d*, and *m*. This is required by standard IRT modelling, which assumes conditional independence of items used in the estimation (cf. Khorramdel and von Davier, 2014; von Davier and Khorramdel, 2013).

Table 1

Example for Coding BPI

Original Scoring (5-point rating scale)	BPI *e* (Extreme Responses)	BPI *m* (Midpoint Responses)	BPI *d* (Trait Responses)
1	1	0	0
2	0	0	0
3	–	1	–
4	0	0	1
5	1	0	1

[2] It should be noted that before the rating data were decomposed into binary pseudo items, missing responses were coded as missing values. Usually, negatively worded items in the empirical dataset would also be recoded so that endorsement on the recoded negative items and the positively phrased items all indicated higher levels of the construct. However, there were no negatively worded items in the questionnaire scale used in this study.

Note: BIP - binar pseudo item

IRT Modelling of binary pseudo items

After decomposing rating data into binary pseudo items (BPIs), unidimensional and multidimensional IRT models were applied to test whether the binary pseudo items are measures of response styles or construct-related responses. In the current paper, all estimated IRT models are based on the 2-parameter logistic model (2-PL model; Birnbaum, 1968). The 2-PL model generalises the Rasch or 1-parameter logistic model (1-PL model; Rasch, 1960). The 1-PL or Rasch model postulates that the probability for response *x* to item *i* for respondent *v* (or for answering toward a trait) depends on only two parameters: the item parameter β_i (difficulty of endorsement) and the person parameter θ_{v} . The 2-PL model postulates an additional item parameter, the discrimination parameter α_i . For unidimensional scales, the model equation of the 2-PL model is defined as:

$$P(x = 1 \mid \theta_v, \beta_i, \alpha_i) = \frac{exp\left(\alpha_i(\theta_v - \beta_i)\right)}{1 + exp\ \left(\alpha_i(\theta_v - \beta_i)\right)} \qquad (1)$$

The discrimination parameter α_i describes how well an item discriminates between examinees with different trait levels, independent of the difficulty of an item.

The 2-PL model can be specified for multiple scales in multidimensional item response theory models. It is assumed that the 2-PL model holds, with the qualifying condition, that it holds with a different person parameter for each set of different subsets (scales) of items (von Davier, Rost, and Carstensen, 2007). For the case of a multidimensional 2-PL model with between-item multidimensionality (each item loads on only one scale), the probability of response *x=1* to item *i* in scale *k* by respondent *v* can be defined as:

$$P(x = 1 \mid \boldsymbol{\theta}_v, \beta_i, \alpha_i) = \frac{exp\left(\sum_{k=1}^{K} \alpha_{ik}(\theta_{vk} - \beta_i)\right)}{1 + exp\ \left(\sum_{i=1}^{K} \alpha_{ik}(\theta_{vk} - \beta_i)\right)} \qquad (2)$$

where **θv** is a vector of scales and α*ik* is the item loading for item *i* on scale *k* with the restriction that each item loads on only one scale.

For every scale of interest, the binary pseudo items are modeled with both unidimensional and multidimensional 2-PL models and the overall model fit is compared. The Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978) are used for this comparison. In Table 2, an example illustrates how IRT models are applied to the three binary pseudo items described above in line with the approach described by Böckenholt (2012). A 1-dimensional 2-PL model (1D) is compared to a 3-dimensional 2-PL model (3D) to test whether the BPI data can best be described by a unidimensional factor (all three types of binary pseudo items loading on the same factor) or by three different factors (each type of BPI loading on a different factor: the factor for *d*-items representing the questionnaire scale, the factor for *e*-items representing extreme response style, and the factor for *m*-items representing response styles).

Table 2

Loading Matrix in Case of a Single Questionnaire Scale (Böckenholt Approach)

	1D			3D		
BPIs	Construct	ERS	MRS	Construct	ERS	MRS
d	1	0	0	1	0	0
e	1	0	0	0	1	0
m	1	0	0	0	0	1

Note: BIP - Binar Pseudo Item; ERS - extreme response style; MRS - midpoint response style

In this approach, it is assumed that both extreme and midpoint response styles are present in the data; it is tested for both types of response styles simultaneously. However, what if there were an extreme response style but no midpoint response styles or the other way around? An extension has to be added to the Böckenholt approach to account for this hypothesis. To test for extreme response style and midpoint response styles separately, two additional 2-dimensional models are needed. One model accounts for extreme response style where *d*-items and *m*-items are assigned to one factor measuring the questionnaire construct and *e*-items are assigned to a second factor measuring extreme response style (2D(ERS)). The other accounts for midpoint response style where *d*-items and *e*-items are assigned to a construct factor and *m*-items to a midpoint response style factor (2D(MRS)).

These models are designed to gauge different patterns of response bias: extreme response style (ERS), indicated by BPIe, and midpoint response style (MRS), indicated by BPIm. When response styles are absent, the unidimensional model serves as our null model, providing a benchmark for the scenarios where ERS or MRS do not contribute substantively to the trait measurement. For the two-dimensional models, we separately introduce ERS and MRS to examine their effects. Without the respective response style in the data, we expect low discrimination for the BPIe and BPIm items as they would not contribute much to trait measurement. The additional dimensions dedicated to capturing these response styles would not yield substantial loadings, and thus, the unidimensional model should provide a better fit, accounting for penalties for model complexity in fit statistics like AIC or BIC. The three-dimensional model brings ERS and MRS into play simultaneously. In the absence of these response styles, the 3D model should not significantly improve fit over the 1D model due to the lack of substantial loadings on the additional dimensions. Thus, the 1D model would be expected to have a better fit.

Conversely, when response styles are present in the data, the multidimensional models should provide a better fit than the unidimensional model. This is because the BPIe and BPIm items can better capture the trait variation influenced by the respective response styles, allowing the multidimensional models to separate the influences of trait and response styles into distinct dimensions.

In summary, the comparison between the unidimensional model and the multidimensional models allows us to discern and measure specific response styles within the data. Better fit of a multidimensional model signals a systematic response style bias in the data. Thus, this method provides a nuanced way to identify and account for response styles, enhancing the robustness and validity of our research outcomes.

An alternative would be to use new coding schemes for pseudo items dedicated for different response styles. This is feasible, and some alternatives for such additional schemes exist and could be investigated in the future. Table 3 illustrates the loading matrix of the 2D models compared to the 1D and 3D models.

Table 3

Extended Loading Matrix in Case of a Single Questionnaire Scale

	1D			3D			2D(ERS)		2D(MRS)	
BPIs	Construct	ERS	MRS	Construct	ERS	MRS	Construct	ERS	Construct	MRS
d	1	0	0	1	0	0	1	0	1	0
e	1	0	0	0	1	0	0	1	1	0
m	1	0	0	0	0	1	1	0	0	1

Note: BIP - Binar Pseudo Item; ERS - extreme response style; MRS - midpoint response style

Hypotheses

Our hypothesis in the simulation study is that extreme response style and midpoint response style, given that they exist in the data because respondents give invalid responses, can be measured with *e*-items and *m*-items, respectively, and that, in this case, scale scores based on *d*-items are a more valid measure of the construct of interest. In this case, a 3D model would fit the data relatively better than a 1D model. If a 1D model shows a better model fit, we cannot assume that extreme and midpoint response styles are present. We further assume that the models 2D(ERS) and 2D(MRS) can be used to differentiate between extreme and midpoint response styles in cases where only one of these is present in the data.

For differentiating types of response styles, we use two strategies. One compares the four models (1D, both 2D, and 3D), assuming that the best-fitted model would indicate processes underlying the data. In the second approach, the 1D model is tested against 2D and 3D models.

For the empirical study, we assume that respondents with either low test-taking motivation or low reading ability are more likely to show response styles in questionnaire data administered with a rating scale. For this study, different measures of motivation are defined and cognitive test scores are used to examine whether the presented IRT approach can detect response styles in critical subgroups of the sample showing low motivation or low reading ability in contrast to subgroups of students with higher motivation and reading ability. More details on the measures and the procedure are given in section 4.

Simulation Study

The simulation study presented in this paper aims to simulate data for one scale with and without response styles, and to validate the IRT approach (illustrated above) to measure and correct for response styles. Different levels of extreme response style and midpoint response style in different scenarios, as well as data without response styles, are simulated in a five-step procedure using the Stata$^{®}$13 statistical package. First, three random variables were drawn reflecting the latent trait, the tendency for midpoint response style, and the tendency for extreme response style. In the second step, responses to simulated items measuring the latent trait were generated. In the third step, responses to the simulated items were exposed to response styles (midpoint response style or extreme response style) and observed responses affected by response styles were generated. In the fourth step, responses exposed to response styles were recorded into pseudo items as described in section 2.2. In the last step, unidimensional and multidimensional IRT models were estimated, and their fit was assessed. Figure 1 presents the overview of the simulation design and is followed by a detailed description of each section.

In previous research different approaches to simulate response styles data were used. Böckenholt (2012) made direct use of IRTree models, Flack and Cai (2016) used the multidimensional nominal response model, Plieninger (2017) extended the multidimensional Rasch (1966) model to modelling response styles. Wetzel and collogues (Wetzel, Böhnke, and Rose, 2016), among others, used mixture models. In this study, we decided not to relay on any given parametrisation of the model. Instead, a stepwise procedure is used, allowing us to generate data flexibly for producing different scenarios, such as different consistency levels of response styles (a unique feature of the presented study). This study is not focused on recovering model parameters but on testing the ability to detect response styles. Therefore, a strict alignment between the generating model (which cannot be known in real settings) and the estimated model (a simplification of the real processes) is unnecessary.

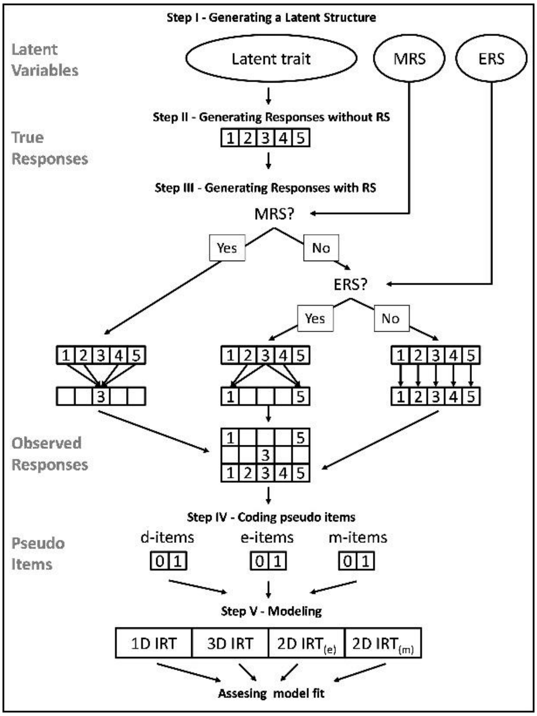

Figure 1.

Simulation design. Note: ERS - extreme response style; MRS - midpoint response style

Step I - Generating the Latent Structure

For each simulated dataset, 1,000 subjects were generated. This sample size was chosen as a typical effective sample size for most large-scale surveys. For each subject, three continuous latent variables were drawn from the standard normal distribution:

1. latent trait to be measured
2. a tendency for midpoint response style
3. a tendency for extreme response style

We generated latent variables according to two scenarios. In the first scenario (matrix A), correlations between latent variables were set to zero. A more complex scenario was provided by employing correlation matrix B depicted in Table 4, where we specified small to moderate correlations between all variables and large negative correlations between two response styles. By doing this, we assume that, in most cases, respondents tend to use only one response style. The indicators of latent variables were sampled from a multivariate normal distribution according to specified correlation matrices. The values for the correlation matrix reflect real data. They were obtained using the IRT decomposition approach on different scales from PISA 2012. In the majority of the scales, the extreme response style variable was highly negatively correlated with the midpoint response style variable, the latent trait variable was moderately positively correlated with the extreme response style variable, and small, usually negative, correlations between the latent trait and the midpoint response style variable were observed.

Table 4

Correlation matrixes for generated latent variables

Correlation matrix A

	T	M	E
T	1	0.0	0.0
M	0.0	1	0.0
E	0.0	0.0	1

Correlation matrix B

	T	M	E
T	1	-0.2	0.4
M	-0.2	1	-0.8
E	0.4	-0.8	1

Note: T - latent trait; M - midpoint tendency; E - extreme tendency.

Step II - Generating Responses without Response Styles

For the (unidimensional) latent trait, variable responses to 10 simulated items using a 5-point rating scale (5 response categories) were generated. The generalised partial credit model (GPCM; Muraki, 1992), assuming monotonic ordering of the 5 response categories, was used for simulating responses. The generalised partial credit model is an extension of Andrich's multicategory model (Andrich, 1978) and Masters' partial-credit model (Masters, 1982).

For an item scored on a scale *0-m,* GPCM probability for observing response for category $x \in \{0,...,m\}$ is described by:

$$P(\mathbf{x}_i = x \mid \theta_v, \alpha_i, b_{im}) = \frac{\exp\left(a_i\left(x\theta - \sum_{r=0}^{x} b_{ir}\right)\right)}{\sum_{c=0}^{m} \exp\left(a_i\left(c\theta - \sum_{r=0}^{c} b_{ir}\right)\right)} \tag{3}$$

where $b_{i,0}$ =0; and $b_{im.}$ is a threshold parameter for item *i* and response category *m*.

For each item in each simulated dataset, the discrimination parameter "a" was sampled from a uniform distribution; in most scenarios, the range was [0.8-1.2], which represents a limited range of discrimination parameters. Other ranges of discrimination parameters were also used to reflect a scale exposed to the potential problem when data shows only “essential unidimensionality” (Stout, 1987, Nandakumar, 1993) instead of perfect unidimensionality. A small range of simulated discrimination parameters results in observed data where correlations among items are very similar, indicating a situation with one dimension. A wide range of discrimination parameters results in some items having higher discrimination parameters than others. In such situations, observed inter-item correlations are different for each pair of items, as observed correlations between items are proportional to the product of discrimination parameters. Thus, some items are more correlated with one subset of items than with another subset of items violating assumptions of perfect dimensionality.

Thus, even when data is generated from a one-dimensional model, deviations from strict unidimensionality may occur, which can be observed in the item correlation matrix. These deviations are caused by the design of simulation data but are also likely to be found in real data. In such situations, the two-parameter logistic model (2PLM) can partially account for the differences in item correlations through discrimination parameters. Conversely, the Rasch model would suggest a poor fit and indicate the potential need for a multidimensional solution..

Moreover, a broader range of the distribution of slope parameters from 1 indicates larger probabilities that generated scales would have different reliabilities. As different ranges of reliabilities and differences in item discrimination parameters are often found in real-world situation, we introduced those conditions in our simulations.

Thresholds were sampled from the standard normal distribution in a two-step procedure:

1. The difficulty of each item was sampled from the standard normal distribution.
2. Thresholds were sampled from the normal distribution with a mean equal to the difficulty sampled in the first step and a standard deviation of 1. The thresholds were sampled with the following order restriction: $b_{i1} < b_{i2} < b_{i3} < b_{i4}$.

We are referring to the responses generated in this phase as “true responses”, that is, responses that were not affected by any response style or confounding factor. Responses affected by response styles were generated in step III and are referred to as “observed responses”.

Step III - Generating Responses with Response Styles

For each subject and each item, two binary indicator variables were constructed reflecting whether the item is affected by a response styles (1) or not (0). Indicators for midpoint response style and extreme response style were generated independently using the 2-PL model. Three scenarios were simulated:

1. Both the midpoint response style and the extreme response style present
2. Only the midpoint response style present
3. Only the extreme response style present

The difficulty parameter B in the 2-PL model was used for controlling the number of items affected by each response styles. In each scenario, different levels of response styles bias were simulated: 10%, 20%, 30%, 40%, or 50% bias (we use capital "B" to distinguish this parameter from the "b" parameter that was used for generating true responses).

Parameter "A" in the 2-PL model was used for controlling the "consistency" of response styles (we use a capital "A" to distinguish this parameter from the "a" parameter that was used for generating the true responses), referring to whether respondents are consistent in showing response styles or not. The greater the value of A, the greater the consistency of response styles being shown. In low consistency settings, responses affected by response styles are widespread among all respondents, while in high consistency settings, respondents are generally divided into two groups: those affected by response styles and those not.

In the simulations, we used different values for this parameter ranging from 0.5-5.0. Figure 2 illustrates the meaning of consistency in terms of the numbers of average response styles per subject. When the consistency is low (A=0.5), most subjects have at least one affected response (upper left panel). When the consistency is medium (A=1 or A=2), two classes of subjects might be distinguished. The first class consists of subjects with no responses affected by response styles and the second class consists of subjects with at least one response affected by response styles. The second class is described by a considerable variation of a number of affected responses (upper right panel A=1 and lower left panel A=2). If the consistency is very high (A=5; lower right panel), subjects are roughly categorised into two homogenous classes: no items affected by response styles and virtually all items affected by response styles.

After generating indicators of response styles, the true responses were recoded into observed responses. First, midpoint response style was applied. If the indicator variable of midpoint response style related to the item response was equal to 1, the true response was recoded into category 3 (i.e., midpoint category), if the indicator variable was 0, no change was made. Next, the recoding according to extreme response style was applied (if the indicator for midpoint response style was equal to zero). If the indicator variable for extreme response style was 0, nothing was done. If the indicator variable was 1, true responses were recoded:

1. If the true response was 1, no recoding was applied
2. If the true response was 2, the observed response was recoded to 1
3. If the true response was 3, the observed response was recoded to 1 or 5, with a probability of 0.5 for each possibility

4. If the true response was 4, the observed response was recoded to 5 If the true response was 5, no recoding was applied

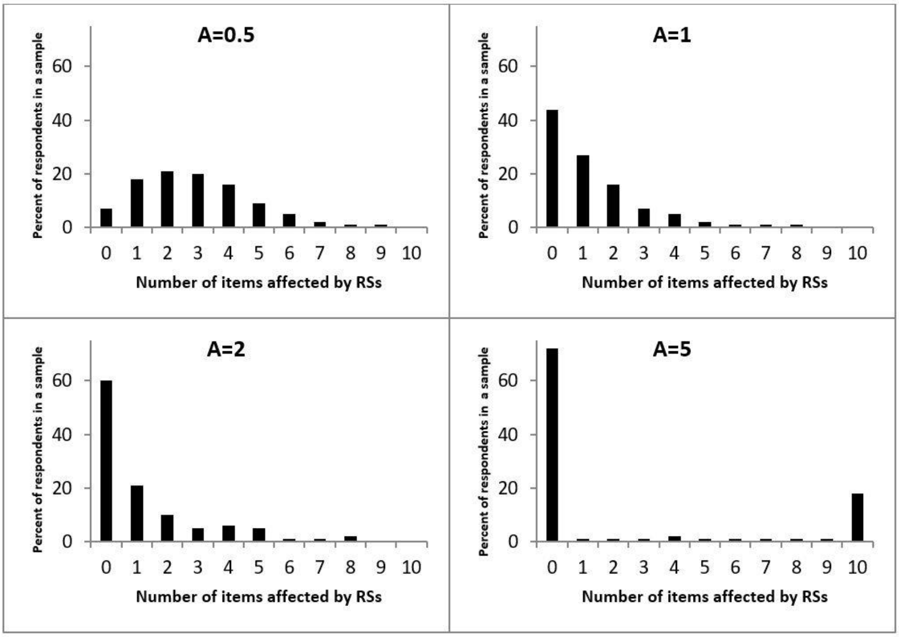

Figure 2.

Three types of consistency.

Step IV - Generating Responses with Response Styles

In this step, binary pseudo items were created using observed responses. Three sets of binary pseudo items were constructed reflecting extreme responses (*e*-items), mid-point responses (*m*-items), and responses representing the latent trait (*d*-items). For a detailed description of the recoding, see section 2.2.

Step V - Modelling

In each sample and scenario, a 1-dimensional (1D) model, a 3-dimensional (3D) model, and a 2-dimensional (2D) model were estimated (seeking to detect extreme response style and midpoint response style), and AIC (Akaike, 1974) and BIC (Schwarz, 1978) measures were computed for model evaluation. In the 1-dimensional model, all binary pseudo items were assigned to one factor representing the latent trait. In the 3-dimensional model, each BPI type was assigned to a different factor representing the latent trait (*d*-items), the midpoint response style (*m*-items), and the extreme response style (*e*-items). In the 2-dimensional model, one factor represented the latent trait and the second factor represented one of the response styles (extreme response style or midpoint response style); to test whether midpoint response style were present, *e*-items and *d*-items were assigned to the trait factor and *m*-items to the midpoint response style factor; to test whether extreme response style is present, *d*-items and *m*-items were assigned to the trait factor and *e*-items to the extreme response style factor (see Table 3). Those five steps were repeated 400 times for each set of simulation conditions.

Results

This section presents the results from the simulation study under different conditions. In the first part of the section, results for datasets without response styles are presented followed by the results of simulations where midpoint response style and extreme response style are present at the same time. Then, the results of simulations are presented where only one response styles type is present (midpoint response style or extreme response style). Several conditions are examined, differentiated by per cent of responses affected by response styles, consistency of response styles, imposed correlation matrix of latent variables, and different range of discrimination parameters sampled for true responses.

Data without response styles and false detection of response styles

First, the described IRT procedure for measuring response styles (see sections 2.2 and 2.3) was applied to data generated without response styles. It was examined whether the 3D model tested against the 1D model could show artificial results (false detection of response styles). If the IRT approach was working properly, no response styles should have been detected when not present in the data. However, there was a weakness in the approach, because to a certain extent, the current IRT method of detecting response styles was depending on the assumption of unidimensionality of the original scale. Perfect unidimensionality would be obtained when discrimination parameters (*a* parameters) are equal across items. If discrimination parameters differ significantly from item to item, statistical methods might show multidimensionality that could contribute to the misfit of the 1D model but could be accommodated by a multidimensional model. This is shown in Table 5, where the *a*-parameter for generating the data is sampled from a uniform distribution with a different range starting with [-0.2-2.2], to [0.8-1.2] and with equal discrimination across all items [1.0].

Table 5 shows the per cent of datasets where the 3D models (accounting for response styles factors) fit relatively better than the 1D models for datasets without response styles (false detection of response styles). When the *a*-parameters are sampled from a broad range [-0.2-2.2], the 3D model fits better than the 1D model in 45% of the cases according to AIC, in 38.5% according to the BIC. When the range of the distribution shrinks, the number of false detections of response styles decreases. With a range of [0.6-1.4], the level of misclassification varies between 4.3% and 8.8%. With a range of [0.8-1.2] for all measures of fit, the rate of false detection is lower than 5%. In an ideal situation (when the *a*-parameters are equal across items) the per centage of a false response styles detection is 4% according to the AIC, 2.5% according to the BIC. Therefore, it seems the method will not show a high false identification rate for unidimensional scales with similar discrimination across items. The BIC is the most conservative measure, that is, it gives the smallest number of false response styles detections even when differences among item discriminations are relatively high [0.4-1.6]. These results suggest that using the BIC measure should be a safer way of assessing response styles compared to the AIC. However, one should keep in mind that when examined scales have a broad range of *a*-parameters, the probability of false identification of response styles will be substantial.

Table 5

Per cent of Simulated Datasets without response styles where the 3D Model Fits Relatively Better than the 1D Model (Different a-Parameters for the GPCM Generation), according to AIC and BIC

Measures of model fit	Item discrimination (a-parameter)						
	[-0.2-2.2]	[0.0-2.0]	[0.2-1.8]	[0.4-1.6]	[0.6-1.4]	[0.8-1.2]	[1.0]
AIC	45.0	30.0	19.5	13.3	8.8	4.0	4.0
BIC	38.5	23.8	14.0	9.8	4.3	2.0	2.5

As described earlier, it should not only be possible to test whether two types of response styles are present in the data – comparing a 1D model with a 3D model – but also to test the hypothesis that only one type of response styles is present. Thus, two types of 2D models were applied to the data as well (see section 2.3). In the 2-dimensional model accounting for extreme response style, 2D(ERS), *d*-items and *m*-items load on one factor and *e*-items on a second factor. In the 2-dimensional model accounting for midpoint response style, 2D(MRS), *d*-items and *e*-items load on one factor and *m*-items on a second factor.

Table 6 shows the per cent of datasets with the relatively best fit for each of the different IRT models (2D, 1D, and 3D models) applied to data without response styles . As before, the aim was to check whether applying the IRT approach for detecting response styles leads to some artificial results (false response styles detection).

Table 6

Per cent of Relatively Best Fitted Models on Data without response styles (Different a-parameters for the GPCM Generation), according to AIC and BIC

Model	AIC						
	Item discrimination (a-parameter)						
	[-0.2-2.2]	[0.0-2.0]	[0.2-1.8]	[0.4-1.6]	[0.6-1.4]	[0.8-1.2]	[1.0]
1D	**26.3**	**34.3**	**49.5**	**62.0**	**73.5**	**78.3**	**79.5**
3D	17.5	14.5	10.8	8.8	5.3	2.8	3.0
2D(ERS)	37.5	31.5	21.5	13.5	10.5	9.5	9.3
2D(MRS)	18.8	19.8	18.3	15.8	10.8	9.5	8.3
Model	BIC						
	Item discrimination (a-parameter)						
	[-0.2-2.2]	[0.0-2.0]	[0.2-1.8]	[0.4-1.6]	[0.6-1.4]	[0.8-1.2]	[1.0]
1D	**29.3**	**41.0**	**56.8**	**71.7**	**79.5**	**85.5**	**86.5**
3D	13.3	10.8	9.5	5.5	3.3	1.5	2.3
2D(ERS)	37.8	29.3	18.3	10.0	8.5	6.5	6.0
2D(MRS)	19.8	19.0	15.5	12.8	8.8	6.5	5.3

Note: In a perfect situation, bolded cells should equal 100.0; ERS - extreme response style; MRS - midpoint response style

It becomes immediately clear that adding models for comparisons leads to a substantial increase in the levels of false response styles detection. When the range of sampled item discrimination parameters is broad ([0.2-1.8] or broader), in more than 33.75% (18.25%+15.50%) of the datasets, both 2-dimensional models (2D(ERS) and 2D(MRS) together) show a better fit according to the BIC and more than 39.75% (21.50%+18.25%) according to the AIC. This is mainly due to good (false) fit of the 2-dimensional model reflecting extreme response style. When the range of discrimination parameters is extremely wide [-0.2-2.2], the 2D(ERS) fits better than the 1D model, which should describe the generated data most adequately. For a reasonably narrow range of sampled discrimination parameters [0.8-1.2], the level of misclassification decreases substantially; however, still in about 19% of datasets according to AIC (9.5%+9.5%) and in 13% according to BIC (6.5%+6.5%), the 2D models fit relatively better than other models (including the 1D model, which ideally should fit the data without response styles best).

Those results suggest that testing only one type of response style might bring false detection when one seeks to distinguish between extreme response style and midpoint response style. Moreover, it appears that the 2D models are more likely to fit the simulated 1-dimensional generated data than the 3D models. This might be easily explained by the fact that a random overrepresentation of one type of true response (extreme or midpoint) is much more likely than a random overrepresentation of both types of true responses at the same time. Simply, it is much more probable that simulated 1-dimensional data without response styles mimics the multivariate distribution typical for one type of response styles rather than for two types of response styles at once. If the discrimination parameters of the investigated scale are relatively high and there is a strong hypothesis for either extreme response style or midpoint response style (so only one of the 2D models is used for testing), testing the hypothesis that one response styles is present using either the 2D(MRS) or the 2D(ERS) model and BIC would be reasonably robust (around 6.5% of error). However, testing for both types of response styles (extreme response style and midpoint response style) using the 2D models would be much more prone to generate false detection (13%), because more than 1 out of 10 times (according to BIC), the 2D model fit would suggest response styles in data without response styles.

Detection of response styles when midpoint response style and extreme response style are both present in the data.

Figure 3 shows the per cent of datasets where the 3D model fits relatively better than the 1D model, with 20% of responses affected by response styles (10% by midpoint response style and 10% by extreme response style) but with different levels of consistency (0.5, 1.0, 1.5, 2.0, 2.5, 5.0.). Results are presented for two situations depicted earlier in Table 4: for uncorrelated response styles (matrix A upper panel) and correlated response styles (matrix B lower panel).

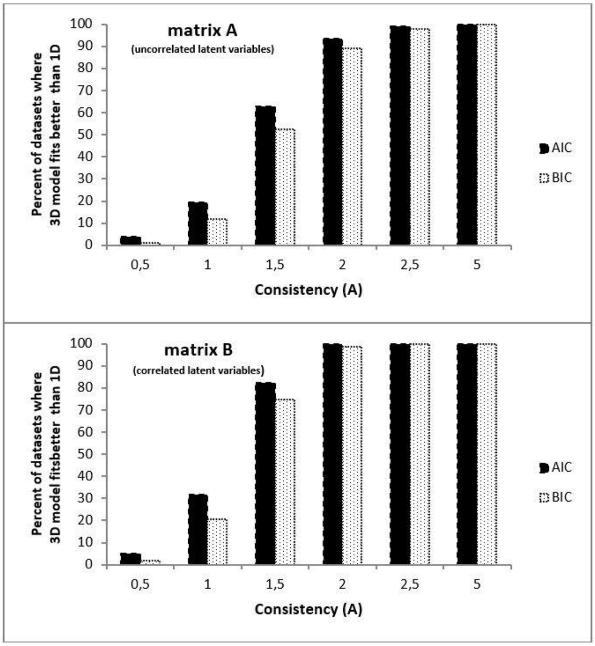

Figure 3.

Per cent of simulated datasets with response styles (20% of responses) where the 3D model fits better than the 1D model (different parameters A of consistency), according to AIC and BIC; uncorrelated response styles (matrix A upper panel) and correlated response styles (matrix B lower panel).

It becomes immediately clear that the greater the consistency, the greater the power for detecting response styles . In scenarios with very low consistency (0.5), response styles are detected only in 1.5 to 5% of the cases (depending on the model fit measures and whether response styles are correlated or not). Increasing the consistency to 1 brings substantial improvement of response styles detection but still only between 12 to 31.75% (depending on the model fit measures and whether response styles are correlated or not). Satisfactory rates of detection (around 90% for uncorrelated response styles and 99% for correlated response styles) are observed for consistency of A=2, and in scenarios with high consistency (A>=2.5), response styles are detected almost without error. This proves that the IRT approach can successfully detect consistent response styles, that is, when response styles are concentrated in a group of respondents rather than widespread among the sample.

Another conclusion that could be drawn from the results depicted in Figure 3 is that correlated response styles (negatively correlated, precisely) are easier to detect. We would expect that, in most situations, only one response styles is affecting the responses of an individual. Switching from one response styles to another might be possible, but would be more expected to occur between different scales (showing one response styles in one scale and another response styles in a different scale) and not within one questionnaire scale. It should also be noted that the AIC have more power for detecting response styles . However, these measures also give more false predictions, as shown in Table 6. As the BIC works reasonably well for consistency A≥2, it seems to be the most balanced measure of fit for detecting response styles .

In Figure 4, the detection of response styles was examined in more detail exploring different consistencies and rates of response styles in response styles datasets using three different consistency levels (A=1, A=2, A=5) and five different levels of response styles (with equal number of extreme response style and midpoint response style in each dataset).

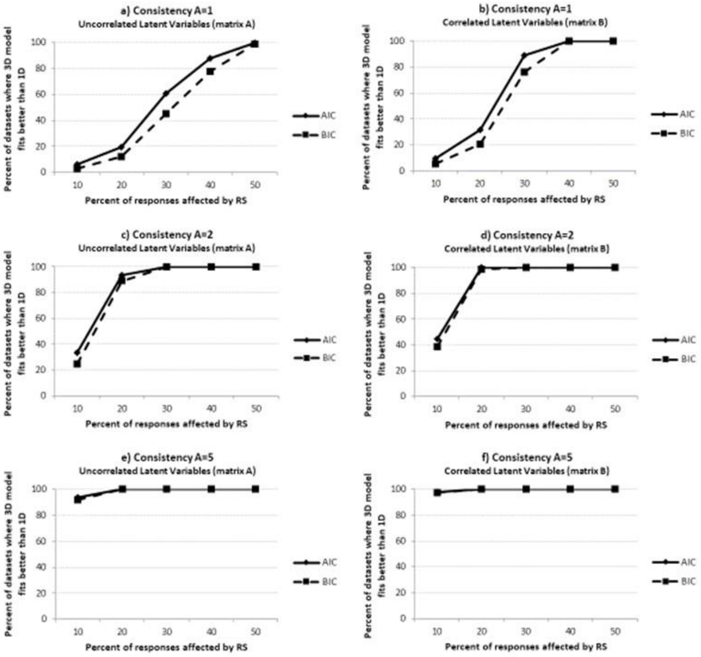

Figure 4.

Per cent of simulated datasets with different response styles rates where the 3D model fits relatively better than the 1D model (different parameters A of consistency, level of response styles and structure of response styles correlation matrix), according to AIC and BIC.

Results presented on Figure 4 show no surprises. The higher the rate of response styles in the data, the higher the ability for detection. Results in Figure 4 also confirm the results presented in Figure 3. The higher the consistency, the higher the detection rate—and it is much easier to detect response styles when they are correlated. We could conclude that the IRT approach is able to detect response styles even when a relatively small number of responses (20%) is affected by response styles if the consistency shown is moderately high (A=2). When the consistency is very high, the IRT approach is able to detect response styles even when only 10% of the responses are affected by response styles. With a low consistency (A=1), a large number of responses (more than 40%) must be affected to detect response styles properly.

In Table 7, the data examined in Figure 4 were used to compare the four IRT models: 1D reflecting no response styles; 3D reflecting a mixture of response styles (extreme response style and midpoint response style); and two 2D models reflecting either extreme response style or midpoint response style as present. In the current data, both response styles are present (midpoint response style and extreme response style) and the number of items affected by the two response styles are equal in each sample, so we would expect the 3D model to fit the datasets relatively best.

The results in Table 7 show that the 2D models fit relatively best in a considerable number of cases (from 15 to 48% depending on the number of responses affected by response styles, the level of consistency, and the type of model fit measure). The 2D(MRS) model accounting for midpoint response style introduces the most bias, indicating the presence of only one response styles when in fact two types of response styles are present. This applies primarily to situations where the consistency is low (A=1) or the per cent of responses affected by response styles is low (10-20%). When the consistency is higher than 2 (A≥2) and the per cent of responses affected by response styles is high (≥30%), the 3D model fits relatively best in more than 90% of the replications. If the level of consistency and per cent of affected response styles is unknown while conducting such analysis, the conclusion must be made that trying to disentangle the type of response styles might bring more harm than benefit. Results suggest that the 2D approach would lead to false conclusions in many situations, indicating that only one response styles is present when in fact a mixture of response styles might be in the data.

Table 7

Per cent of Relatively Best Fitting Models on Data with a Different Rate of Responses Affected by both MRS and ERS, according to AIC and BIC

Measure of fit	Model	Consistency A=1					Consistency A=2					Consistency A=5				
		Per cent of response styles					Per cent of response styles					Per cent of response styles				
		10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
		Uncorrelated Latent Variables (matrix A)														
AIC	1D	61.5	34.8	9.8	1.5	0.0	28.8	1.3	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0
	3D	**2.3**	**6.3**	**28.8**	**48.3**	**67.8**	**16.3**	**70.0**	**92.5**	**97.0**	**99.3**	**74.0**	**99.3**	**99.8**	**100.0**	**100.0**
	2D (ERS)	20.5	24.5	12.8	2.5	0.0	22.8	1.5	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0
	2D (MRS)	15.8	34.5	48.8	47.8	32.3	32.3	27.3	7.5	3.0	0.8	24.3	0.8	0.3	0.0	0.0
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
BIC	1D	75.0	44.5	16.0	2.5	0.0	39.0	2.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0
	3D	**0.8**	**4.3**	**20.5**	**37.3**	**57.8**	**11.8**	**65.5**	**90.8**	**96.5**	**99.0**	**69.8**	**99.0**	**99.8**	**100.0**	**100.0**
	2D (ERS)	13.8	20.5	13.3	3.3	0.0	20.5	1.3	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0
	2D (MRS)	10.5	30.8	50.3	57.0	42.3	28.8	31.3	9.3	3.5	1.0	28.5	1.0	0.3	0.0	0.0
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		Correlated Latent Variables (matrix B)														
AIC	1D	55.5	28.8	1.5	0.0	0.0	22.5	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0
	3D	**3.8**	**15.3**	**52.0**	**84.8**	**99.8**	**25.3**	**93.0**	**99.5**	**100.0**	**100.0**	**86.0**	**97.8**	**99.8**	**100.0**	**100.0**
	2D (ERS)	17.5	16.3	1.8	0.0	0.0	10.8	0.3	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0
	2D (MRS)	23.3	39.8	44.8	15.3	0.3	41.5	6.8	0.1	0.0	0.0	12.5	2.3	0.3	0.0	0.0
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
BIC	1D	69.0	38.5	3.3	0.0	0.0	29.5	0.3	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0
	3D	**2.0**	**11.0**	**41.3**	**75.8**	**99.8**	**21.0**	**90.3**	**99.3**	**99.8**	**100.0**	**83.5**	**97.3**	**99.3**	**100.0**	**100.0**
	2D (ERS)	13.3	14.0	3.0	0.0	0.0	9.3	0.3	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0
	2D (MRS)	15.8	36.5	52.5	24.3	0.3	40.3	9.3	0.8	0.3	0.0	15.0	2.8	0.8	0.0	0.0
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Note: In a perfect situation, where response styles are identified correctly in all datasets, bolded cells should equal 100.00; Note: BIP - ERS - extreme response style; MRS - midpoint response style

Detection of response styles when only one type of response styles is present in the data (midpoint response style or extreme response style).

Finally, in Table 8, analyses on datasets are introduced where only one type of response styles (midpoint response style or extreme response style) was generated (20% or 50% of responses were affected by response styles in these conditions, respectively). Analyses were performed using three levels of consistency (A=1, A=2. A=5) using uncorrelated latent variables (results using correlated latent variables are virtually the same and are not presented here). Bolded cells indicate places where we would expect a high per cent of relatively best fitting models if the IRT approach works correctly.

The model selection based on comparing the fit of four models simultaneously presented in Table 8 indicates the poor performance of the 2D(extreme response style) model in accounting for extreme response style. In most cases, the 2D(ERS) model fits worse than the 3D model even if only extreme response style is present in the dataset. With high and moderately high consistency and a response styles level of 50%, this approach has the ability to detect extreme response style that is essentially zero. The detection of midpoint response style using the 2D(MRS) model works very well, especially when there is a high level of response styles (50%) and a high consistency (A≥2), with the detection rate at almost 100%. A close look at the simulation results reveals when extreme response style are increased in the data, the correlations between the *m*-item-based scale and the true or trait factor (*d*-items) decrease and become more and more negatively correlated (detailed results are available on request). This means that the *m*-items start to load on a separate dimension instead of loading on the true dimension together with the *d*-items. This may explain why the 3D model starts to fit better than the 2D(ERS) model. This finding is most likely an artifact.

In the case of the simulated data with only one response style, in particular those data that were generated by simulating trait differences and extreme response style, there is obviously no need to model a third dimension since the midpoint response style is not present in the data. Therefore, a 3D model will overfit the data, and in particular, will allow for variance in the midpoint response style dimension, while there are no interindividual differences in midpoint response style (as it was never part of the simulation). As a consequence, a 3D model that attempts to estimate correlations with this third – nonexistent - dimension will lead to artifacts, as seen for example in the case of the negative correlations of *m*-item based (or midpoint response style) scale estimates with the other scales. The problem of detecting response styles, in this case, might be solved by the 1D model being compared to the 2D model only (without adding the 3D model). The described problem does not occur when the 2D(MRS) model is compared. Interestingly, when the per cent of midpoint response style is increased in the simulated data, *m*-items start to be a good predictor of midpoint response style, and the 2D(MRS) model fits better than the 3D model. A possible reason could be the difference in missing data in the scoring of *e*-items and *d*-items, while there are no missing data in the *m*-item scoring (see Table 1).

Table 8

Per cent of Best Fitting Models for Data with a Different Rate of Responses Affected by One Type of response styles, according to AIC and BIC, Different Levels of Consistency (Uncorrelated Latent Variables). Results Based on Comparing Four Models Simultaneously

Measure of fit	Model	ERS only			MRS only		
		A=1	A=2	A=5	A=1	A=2	A=5
	20% *responses affected by response styles*						
AIC	1D	30.3	24.8	11.5	14.3	6.8	4.0
	3D	10.5	24.8	44.3	3.0	2.5	7.3
	2D (ERS)	**55.8**	**24.8**	**43.8**	0.3	0.0	0.0
	2D (MRS)	3.5	0.8	0.5	**82.5**	**90.8**	**88.8**
	Total	100.0	100.0	100.0	100.0	100.0	100.0
BIC	1D	36.0	29.8	13.3	20.5	9.3	5.5
	3D	7.3	20.3	38.0	1.5	1.3	4.0
	2D (ERS)	**53.5**	**49.3**	**48.3**	0.3	0.0	0.0
	2D (MRS)	3.3	0.8	0.5	**77.8**	**89.5**	**90.5**
	Total	100.0	100.0	100.0	100.0	100.0	100.0
	50% *responses affected by response styles*						
AIC	1D	0.3	0.0	0.0	0.3	0.0	0.0
	3D	70.5	100.0	100.0	1.3	0.3	0.0
	2D (ERS)	**29.3**	**0.0**	**0.0**	0.3	0.0	0.0
	2D (MRS)	0.0	0.0	0.0	**98.3**	**99.8**	**100.0**
	Total	100.0	100.0	100.0	100.0	100.0	100.0
BIC	1D	0.3	0.0	0.0	0.3	0.0	0.0
	3D	62.5	100.0	100.0	0.0	0.0	0.0
	2D (ERS)	**37.3**	**0.0**	**0.0**	0.3	0.0	0.0
	2D (MRS)	0.0	0.0	0.0	**99.5**	**100.0**	**100.0**
	Total	100.0	100.0	100.0	100.0	100.0	100.0

Note: In a perfect situation, bolded cells should equal 100.0; ERS - extreme response style; MRS - midpoint response style

However, a more appropriate strategy for detecting only one type of response styles would be to compare 1D with 2D models without referring to the 3D model. Such comparisons are presented in Table 9, which describes the same results as Table 8. In this approach, comparisons of 1D and 2D models in most cases reveal the true structure of the data. But satisfactory power (around 90%) is achieved only when the per cent of response styles is high, consistency of response styles is large, or both. Simulations show that with high consistency of (A=5) and even a moderate level of response styles (20%), extreme response style was detected in 88.25% of the simulated datasets and midpoint response style in 96% according to AIC, and with a similar rate of accuracy using BIC, 86.25 and 94.50, respectively. With a high number of response styles, the detection rate is 100%.

Results presented in Table 9 show that in some situations, a comparison of the 3D model with the 1D model might detect when only one type of response styles is present in the data. When the level of response styles is high (50%), the 3D model fits the data better than the 1D model in datasets with extreme response style only; this is true virtually in all cases. When simulated data contained only midpoint response style, the detection of response styles using the 3D model works well but only for low and moderate level of consistency (A=1 and A=2). When midpoint response style was generated with high consistency (A=5) and no extreme response style was present in the data, the capability of detecting response styles using 1D vs. 3D model drops drastically to 32.25% according to AIC and 29.50 according to BIC.

Also, using the strategy based on comparing 1D models against 2D models, the power of detection of extreme response style is smaller than for midpoint response style, at least with 20% of response styles in the data. This issue is directly linked with the problem that in some situations, it is hard to distinguish extreme responses in terms of extreme response style from construct-related responses (*d* items).

Table 9

Per cent of Best Fitting Models(2D or 3D) for Data with a Different Rate of Responses Affected by One Type of response styles, according to AIC and BIC, Different Levels of Consistency (50% of Responses Affected by response styles). Results Based on Comparing 2D Models against 1D Model, and 3D against 1D Model Separately

Meas-ure of fit	Model	ERS only			MRS only		
		A=1	A=2	A=5	A=1	A=2	A=5
	20% *responses affected by response styles*						
AIC	1D vs. 3D	26.0	49.0	72.8	25.5	40.5	54.8
	1D vs. 2D(ERS)	**68.0**	**74.0**	**88.3**	14.0	12.0	16.0
	1D vs. 2D(MRS)	20.0	24.0	28.0	**85.8**	**93.3**	**96.0**
BIC	1D vs. 3D	19.5	41.5	68.0	15.5	29.0	46.5
	1D vs. 2D(ERS)	**61.8**	**69.5**	**86.3**	7.3	7.8	9.8
	1D vs. 2D(MRS)	14.0	17.0	20.8	**79.5**	**90.8**	**94.5**
	50% *responses affected by response styles*						
AIC	1D vs. 3D	96.0	100.0	100.0	100.0	100.0	32.3
	1D vs. 2D(ERS)	**99.8**	**100.0**	**100.0**	24.8	2.0	24.3
	1D vs. 2D(MRS)	54.8	96.0	71.5	**99.8**	**100.0**	**100.0**
BIC	1D vs. 3D	94.0	100.0	100.0	99.3	100.0	29.5
	1D vs. 2D(ERS)	**99.8**	**100.0**	**100.0**	10.0	1.3	18.3
	1D vs. 2D(MRS)	43.8	94.3	70.3	**99.8**	**100.0**	**100.0**

Note: In a perfect situation, bolded cells should equal 100.00; Note: ERS - extreme response style; MRS - midpoint response style

Summary of Findings and Conclusions

Results of the simulation study clearly show that the presented IRT approach is a valid and efficient tool for detecting response styles when both extreme and midpoint response styles are present in the data. When the discrimination of the items is roughly equal in the questionnaire scale, there is almost no false detection of response styles. The power of the method is reasonably good as well. For correlated response styles showing high (A=5) or moderately high (A=2) consistency, response styles were detected in more than 90% of replications in cases where 20% of data responses were affected by response styles, and more than 99% were detected in cases where 30% of data responses were affected by response styles (both according to the BIC measure).

Results clearly show an advantage of using the BIC measures for detecting response styles when applying the presented IRT approach. BIC is a more robust measure for detecting response styles than AIC. This measure of model fit combined with modelling response styles as multidimensional factors provide the smallest number of false response styles detections and reasonable power when the consistency of response styles is moderate.

The situation is more complicated when only one type of response styles is present in the data. On the one hand, 1D vs. 3D model comparisons in most situations lack the power to detect only one type of response styles. On the other hand, comparisons involving 2D models result in a high rate of false detection (up to 19% in some situations). According to those results, we would recommend using 1D vs. 3D model comparisons as the primary test and additional model comparisons (2D versus 1D models) as additional sources of information. In most situations, the 3D model would indicate the presence of both response styles. However, in some situations it might indicate a high level of only one type of response styles. In such situations, 2D models might be helpful. For instance, if the 3D model fits the data better than the 1D model, the 2D(ERS) model also fits the data better than the 1D model, but the 2D(MRS) fits worse than the 1D model. We would conclude that response styles in the investigated data are mainly driven by extreme response style.

Empirical Study: Evaluation Using Data from PISA 2012

The simulation study brings necessary but limited proof of the validity of the presented IRT approach for response styles detection. To complete the picture we used empirical data to examine the validity of the approach and show its usefulness in real data analysis. To do so, we defined three measures: two for low test-taking motivation and one for low ability. The low test-taking motivations are (a) overclaiming in a selected background questionnaire (BQ) scale and (b) the number of omitted responses in the cognitive assessment. The low-ability measure is the plausible values obtained from the cognitive assessment in the reading domain. The idea behind this study is that students with low test-taking motivation and students with problems reading and understanding items from the BQ might be more likely to show response styles. The IRT approach for measuring response styles was applied to the whole student sample and to subsamples, with the student sample divided by the three measures of motivation and ability.

Sample and Instruments

The data used in the empirical study come from PISA. It has been conducted in cycles every three years since 2000 to monitor students' ability to use their knowledge and skills to meet real-life challenges and provide trend measures over time. In each cycle, one of the three domains is featured as the major domain, while the others serve as minor domains. In addition to the cognitive assessment, PISA measures noncognitive scales and variables through BQs (student, parent, and school questionnaires). The data used in the current study come from the student questionnaire of the PISA 2012 main study when mathematics was the main domain.

The PISA 2012 survey was conducted in 34 Organisation for Economic Co-operation and Development (OECD) countries and 31 partner countries and economies on students enrolled in lower-secondary or upper-secondary institutions and aged between 15 years, 3 months, and 16 years, 2 months. The sample was stratified and two-stage, meaning schools were sampled first, and students were then sampled within those schools (see Organisation for Economic Co-operation and Development, 2014, and www.oecd.org/pisa for full documentation on the PISA coverage and technical standards). For this analysis we used a sample consisting of English-speaking students from English-speaking countries only. The sample consists of n=51,836 students with 49.72% female (n=25,771) and 50.28% male (n= 26,065).

Paper-based assessments were used in PISA 2012 and lasted 2 hours. Cognitive test items were a mixture of questions requiring students to construct their own responses and multiple choice. The BQ was administered to all participating students. The questionnaire collected a range of information on students' households, resources available in the home, parental and family circumstances, and the practices and influences that may be related to academic success in specific subjects. In PISA 2012, there were four scales with a 5-point rating scale, meaning they were suitable for the decomposition of rating data into binary pseudo items (see section 2.2). For the empirical example, we have chosen the longest scale, comprising 13 questions measuring "Familiarity with Math Concepts" (the remaining 3 scales are substantially shorter: two comprise 5 questions and one comprises 4 questions). In this scale, students were asked about their familiarity with certain math concepts like exponential function, divisor, quadratic function, and so on. The rate of missing data was very small, not exceeding 1.3% for any item. We excluded all items with missing data from the analysis. A detailed description of the Familiarity with Math Concepts scale is presented in Table 10 and in (OECD, 2014).

Table 10

The "Familiarity with Math Concepts" Scale in PISA 2012

Number and name of the selected scale	*Scale 62* *Familiarity with Math Concepts*
Question	Thinking about mathematical concepts: how familiar are you with the following terms?
Category 0	Never heard of it
Category 1	Heard of it once or twice
Category 2	Heard of it a few times
Category 3	Heard of it often
Category 4	Know it well, understand the concept
N of Items	13

Measures of Low Test-Taking Motivation

In addition to items that illustrate real mathematical concepts, the scale for Familiarity with Math Concepts was equipped with three so-called "overclaiming" items. With these items, respondents were asked about nonexistent mathematical concepts – "proper number," "subjunctive scaling," and "declarative fraction" – that could be used to measure students' attention. We assume that students who confirm knowing nonexistent mathematical contents have either less knowledge or low test-taking motivation, meaning they might show response styles.

We used two additional measures of motivation in order to examine the validity of the IRT approach. The first is based on the ratio of omitted responses to all test items in the cognitive assessment. We assume that a large number of omitted responses could be a sign of low motivation. Students who show low motivation by skipping a lot of items are assumed to show response styles instead of giving reliable and valid responses when answering the BQ questions. With that said, we do note a distinction between students who omit items quickly without reading them or spending time on them – a possible sign of low motivation – and students who omit items after reading them – a possible sign of low ability. Since PISA 2012 was administered as a paper-based (instead of a computer-based) assessment, there is no measurement of time that could be used to differentiate rapid responses along these lines. However, even in the case of low ability (especially in the cognitive domain of reading)[3] we are assuming a higher probability of response styles in the BQ.

Analyses and Results

We conducted our analysis on the whole sample and subsamples divided three ways: by the number of omitted responses in the cognitive assessment, overclaiming scores, and reading test scores (we used the first plausible value as an indicator of reading proficiency). For each indicator of motivation/ability, the sample was divided into five subgroups: low, medium-low, medium, medium-high, and high, as shown in Table 11.

Table 11

Subsamples used in the analysis

	Motivation/Ability				
Criterion /group	Low	Medium Low	Medium	Medium High	High
Over-claiming	4th quartile of persons who are overclaiming	3rd quartile of persons who are overclaiming	2nd quartile of persons who are overclaiming	1st quartile of persons who are overclaiming	No nonexistent concepts identified
Omitted Responses	4th quintile of persons with missing data	3rd quintile of persons with missing data	2nd quintile of persons with missing data	1st quintile of persons with missing data	No missing data in test
Reading Ability	1st quintile of reading results	2nd quintile of reading results	3rd quintile of reading results	4th quintile of reading results	5th quintile of reading results

First, considering the simulation study's findings, only 3D and 1D models were estimated and compared to each other using the BIC (results are presented in the upper panel of Figure 5). The analysis conducted on the full sample indicates the existence of response styles. The BIC for the 3D model is substantially smaller than for the 1D model (3D: 1140525; 1D: 1184943), showing a better fit. In a second step, we compared the 3D and 1D models in each subgroup (shown in Figure 5); the vertical axis

[3] We have chosen reading as most relevant ability in the context of this analysis. We checked results for other domains and they stay essentially the same.

shows how much less per cent the BIC for the 3D model was than for the 1D. With a smaller BIC meaning a better model fit, the figures shows the relative fit of the 3D model, implying the existence of response styles.

Results show that in subgroups where we suspect low motivation/ability, the fit of the 3D model is substantially better than in subgroups where motivation/ability is presumed higher. These results confirm our hypothesis and, thus, the utility of the IRT approach.

Similar analyses were conducted comparing the 1D vs. 2D(ERS) model (middle panel of Figure 5) and comparing the 1D vs. 2D(MRS) model. All three sets of model comparisons show that both extreme and midpoint response styles are present in the data and that both are similarly related with criterion variables, with the difference that midpoint response style is less related to the number of omitted responses than extreme response style.

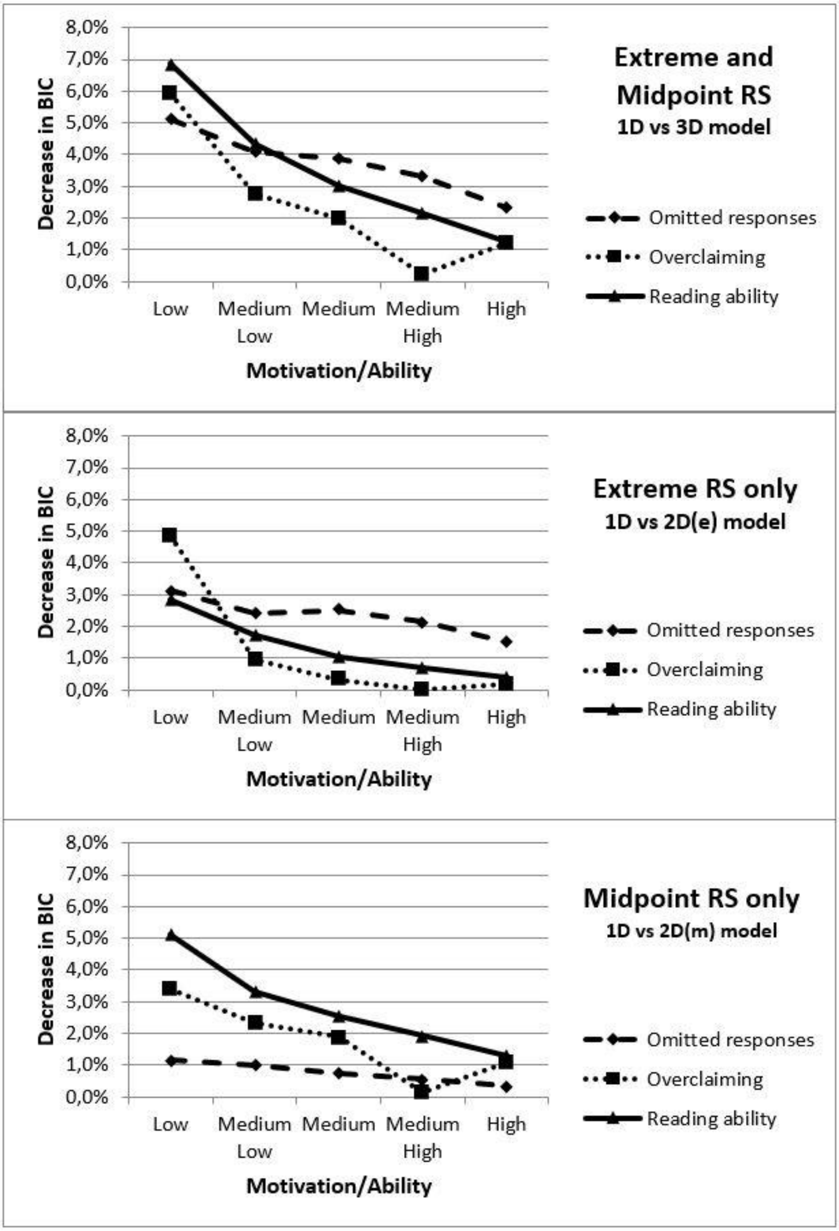

Figure 5.

Comparison of relative fit of the 1D model versus either the 3D model or the 2D models in subpopulations with different motivation/attention/ability. Per centages indicate how often the 2D or the 3D model fit better than the 1D model.

Discussion

Two studies are presented to examine the validity and power of a new IRT approach to measure and correct for response styles with a focus on the extreme and midpoint response styles. The approach was introduced by Böckenholt (2012) and is based on the decomposition of rating data into binary pseudo items representing different response subprocesses in relation to the offered response categories of a rating scale. After the data are decomposed, IRT models are applied to the resulting binary pseudo items to test whether they are measures of response styles or construct-related responses. In the current study, the 2-PL model is used.

The first study is a simulation study in which data are simulated that either show no response styles (construct-related responses), both extreme response style and midpoint response style, or only one of the two examined response styles (extreme response style or midpoint response style). Construct-related responses and response styles were simulated with regard to a 5-point rating scale. The simulated response styles data consider different levels of response styles (10%, 20%, 30%, 40%, 50%) and different consistencies (A=1, A=2, A=5). Items for the latent trait (construct) were simulated showing either a high psychometric quality (similar item discrimination or slope parameter) or low psychometric quality (different item discrimination or slope parameter).

The second study is an empirical study based on data from the Familiarity with Math Concepts scale selected from the PISA 2012 student BQ; the scale was administered with a 5-point rating scale. The rating data in both studies were recoded into three different kinds of binary pseudo items as described by Khorramdel and von Davier (2014) and von Davier and Khorramdel (2013): *e*-items as a possible measure of extreme response style considering only extreme responses, *m*-items as a possible measure of midpoint response style considering only responses to the midpoint of the scale, and *d*-items as a measure of the latent trait not biased by response styles considering moderate and extremely positive responses (after recoding negatively worded items).

Findings

In the simulation study, a 1D model with all three types of binary pseudo items loading on the same factor (the trait or construct) was compared to a 3D model with each of the three binary pseudo items loading on a separate factor accounting for both extreme response style and midpoint response style. Both models were additionally compared to two 2D models, one accounting for the latent trait and extreme response style only (2D(ERS)) and one accounting for the latent trait and midpoint response style only (2D(MRS)). In the 2D(ERS) model, *d*-items and *m*-items were assigned to the trait factor while *e*-items were assigned to the extreme response style factor. In the 2D(MRS) model, *d*-items and *e*-items were assigned to the trait factor, while *m*-items were assigned to the midpoint response style factor. Results show that with scale having similar item slope parameters, the detection rate of response styles is very high if both response styles are present in the data (extreme response style and midpoint response style). It is also shown that the more the responses are affected by response styles and the higher the consistency of response styles that is evident, the higher the response styles detection rate using the presented IRT approach. In addition, it is easier to detect response styles when they are correlated compared to uncorrelated response styles, although the differences in most situations are not very high.

When only one type of response styles is present in the data (extreme response style or midpoint response style), findings show that model comparisons have to be performed carefully. 3D vs. 1D model comparisons often lack power for detecting only one response styles in such situations.. Using 2D models for detecting response styles might lead to a relatively high false detection rate when using the 3D model in the comparison simultaneously. Because there are not three factors underlying the simulated data but two (one response styles factor and the trait factor), a 3D model will overfit the data, confounding the fit of the 2D model.

The problem disappears when only the 2D model is compared to the 1D model (without comparing both models to the 3D model). However, we are not recommending using 1D vs. 2D models comparisons as a main strategy for detecting response styles. The results of our simulation study provide support for the following sequential steps that should be used to differentiate between types of response styles:

1)Perform a 1D versus a 3D model comparison. If the 1D model fits best - stop. Most likely, no response styles are present. If the 3D model fits better than the 1D model, the presence of response styles can be assumed. To differentiate between types of response styles, proceed with 2):.

2) Perform a 2D (extreme response style) versus 1D model comparison and a 2D (midpoint response style) versus 1D model comparison. If both comparisons indicate that a 2D model fits better, both types of response styles are most likely present. If only one type of 2D model fits better than the 1D model, most likely only one type of response styles is present. If the 1D model fits better than each of the 2D models, results should be treated as inconclusive.

In other words, 2D models might be helpful as an additional tool for recognising which of the response styles types are present in the data (extreme response style or midpoint response style).

Another finding of the simulation study is that the BIC as an overall model fit seems to be a more robust measure than the AIC for detecting response styles with the current IRT approach. Using the BIC measure minimalises false positives while keeping reasonable high power when the consistency of response styles is high and/or the per cent of responses affected is substantial.

In the empirical study, it was decided to estimate 1D and 3D models based only on the findings of the simulation study, using the BIC for model comparison. Analyses were conducted for English-speaking students from English-speaking countries. The sample was divided into different subsamples based on three variables: overclaiming and omitted responses (in the cognitive part of the assessment) as measures of test-taking motivation and the cognitive test scores for the domain of reading. Analyses were applied to the whole sample and to the different subgroups. It can be shown that response styles exist in the data since the 3D model shows a better fit than the 1D model in the BQ scale of Familiarity with Math Concepts. As expected, subgroups with suspected lower test-taking motivation (higher scores in overclaiming and higher omitted response rates) and with lower reading ability scores show a substantially better fit of the 3D model than in subgroups with higher motivation and ability.

Limitations and Further Research

The presented method works well only when the range of discrimination parameters is narrow. This is not the case for all operational datasets. Therefore, it is important to interpret the results carefully. We would not recommend using the presented method on scales that show considerable differences in the range of estimated discrimination parameters because it would increase the risk of a high false detection rate.

As in most simulation studies, not all possible scenarios and situations were simulated and examined. However, in our opinion, we provided the most relevant ones. Future research might address additional combinations of variables and additional scenarios to examine the power and validity of the examined IRT approach using binary pseudo items. The empirical data used in this study are based on a paper-based assessment. Thus, no measure of response times is available to differentiate between rapid omitted responses as a sign of low motivation and responses with higher response times as a sign of a lack of ability. Further research could use data from a computer-based assessment to incorporate response times and other variables available in such assessments (e.g., such process data as the number of actions in each item). It would also be interesting to use mixture IRT models and compare this approach to using multiple known subgroups (Khorramdel, von Davier, Pokropek 2019).

Further research should also focus on constructing methods that could detect situations with only one type of response styles with more power and less false detection than the described approach of the 2D models. At least two approaches might achieve this: expanding the approach described in this paper by additional recoding schemes that would be appropriate to only one type of response styles, or expanding the model framework using other models such as HYBRID models (Yamamoto, 1989) or mixture IRT models (Rost, 1991).

Finally, one potential area of investigation could be a comparative analysis of various methodologies like Henninger's sum-to-zero constraints (2018), Tutz's finite mixtures approach (2018), and our extended IRTree model. However, such a comparative analysis is challenging due to the different theoretical assumptions about response processes underpinning these approaches. These distinctions necessitate careful consideration of their unique perspectives on how respondents engage with rating scales.

Moreover, for a meaningful comparison, there would be a need to refer to some external criteria that could serve as a standard for comparison across models with different assumptions. Consequently, future research efforts in this area would benefit from identifying such criteria to effectively compare these methodologies, further enriching our understanding of response patterns in survey-based research.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19:716-723. 716–723. https://doi.org/10.1109/TAC.1974.1100705

Andrich D. (1978) A rating scale formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Asparouhov, T., & Muthén, B. (2020) IRT in Mplus. *Version 2. Technical report. [https://www.statmodel.com/download/MplusIRT.pdf]*

Bachman, J. G., and O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black–white differences in response styles. *Public Opinion Quarterly*, 48:491–509. https://doi.org/10.1086/268845

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

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M.T., and Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment* 14(4):317-335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison–WesleyBöckenholt, U. (2012). Modelling multiple response processes in judgment and choice. *Psychological Methods* 17:665-678. https://doi.org/10.1037/a0028111

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological methods*, *22*(1), 69.

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

Buckley, J. (2009) *Cross-national response styles in international educational assessments: Evidence from PISA 2006*. NCES Conference on the Program for International Student Assessment: What we can learn from PISA, Washington, D.C. Retrieved from http://edsurveys.rti.org/PISA/

Chen, C., Lee, S. Y., and Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science* 6:170–175. https://doi.org/10.1111/j.1467-9280.1995.tb00327.x

De Jong, M. G., Steenkamp, J.-B. E. M. Fox, J.-P., and Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation." *Journal of Marketing Research* 45:104-115. https://doi.org/10.1509/jmkr.45.1.104

Dolnicar, S. and Grun, B. (2009). Response style contamination of student evaluation data. *Journal of Marketing Education* 31:160-172. https://doi.org/10.1177/0273475309335267

Greene, W. H. (2012). *Econometric analysis,* 7th ed. Upper Saddle River, NJ: Prentice Hall.

Hamamura, T., Heine, S. J., and Paulhus, D. L. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual Differences* 44:932-942. https://doi.org/10.1016/j.paid.2007.10.034

Henninger, M. (2021), A Novel Partial Credit Extension Using Varying Thresholds to Account for Response Tendencies. Journal of Educational Measurement, 58: 104-129. https://doi.org/10.1111/jedm.12268

Hui, C. H., and Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology* 20:296–309. https://doi.org/10.1177/0022022189203004

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

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

Khorramdel, L., von Davier, M., & Pokropek, A. (2019). Combining mixture distribution and multidimensional IRTree models for the measurement of extreme response styles. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12179

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

Muraki, E. (1992). A generalised partial credit model: Application of an EM algorithm. *Applied Psychological Measurement* 16:159-177.

Nandakumar, R. (1993). Assessing essential one-dimensionality of real data. *Applied Psychological Measurement* 17(1):29-38. https://www.jstor.org/stable/1435095

Nunnally, J. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.

Organisation for Economic Co-operation and Development 2014. *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems (Volume V), PISA.* Paris, France: OECD Publishing. http://dx.doi.org/10.1787/9789264208070-en

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, and L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Plieninger, H. (2017). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*(1), 32-53. https://doi.org/10.1177/0013164416636655

Plieninger, H., and Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles *Educational and Psychological Measurement* 20:1–25. https://doi.org/10.1177/0013164413514998

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen and Lydiche (Expanded Edition, Chicago, University of Chicago Press, 1980).

Robie, C., Brown, D. J., and Beaty, J. C. (2007). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology* 21:489-509. https://doi.org/10.1007/s10869-007-9038-9

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

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* [*Textbook test theory – test construction*] (2nd ed.). Bern, Switzerland: HuberSchwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6:461-464.

Spratto, E. M., Leventhal, B. C., & Bandalos, D. L. (2021). Seeing the forest and the trees: Comparison of two IRTree models to investigate the impact of full versus endpoint-only response option labeling. *Educational and Psychological Measurement, 81*(1), 39-60.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika* 52:589-617. https://doi.org/10.1007/BF02294821

van Herk, H., Poortinga, Y. H., and Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology* 35:346-360. https://doi.org/10.1177/0022022104264126

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

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

von Davier, M. (2008). The mixture general diagnostic model. In G. R. Hancock and K. M. Samuelson (Eds.), *Advances in latent variable mixture models*. Charlotte, NC: Information Age Publishing.

von Davier, M.(2010a). Why sum scores may not tell us all about test takers. In L. L. Wang (Ed.): Special issue on Quantitative Research Methodology. *Newborn and Infant Nursing Reviews* 10:27-36.

von Davier, M. (2010b). Hierarchical mixtures of diagnostic models. *Psychological test and assessment modelling* 52:8-28.

von Davier, M., and Khorramdel, L. (2013). Differentiating response styles and construct related responses: A new IRT approach using bifactor and second-order models. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, D.M., and C. M. Woods (Eds.) *New developments in quantitative psychology: Presentations from the 77th Annual Psychometric Society Meeting* (pp. 463-488), New York, NY: Springer.

von Davier, M., Rost, R., and Carstensen, C. H. (2007). Introduction: Extending the Rasch model. In M. von Davier and C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 1-12). New York, NY: Springer.

Tutz, G., Schauberger, G., & Berger, M. (2018). Response Styles in the Partial Credit Model. Applied Psychological Measurement, 42(6), 407–427. https://doi.org/10.1177/0146621617748322

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

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

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*(2), 304-324.

Yamamoto, K. (1989). *A Hybrid model of IRT and latent class models* (Research Report Series No. RR-89-41). Princeton, NJ: Educational Testing Service.