Psychological Test and Assessment Modeling, Volume 59, 2017 (1), 71-92

Recent IRT approaches to test and correct for response styles in PISA background questionnaire data: a feasibility study

Lale Khorramdel[1]*, Matthias von Davier*[2]*, Jonas P. Bertling*[3]*,*
Richard D. Roberts[4] *& Patrick C. Kyllonen*[3]

Abstract

A relatively new item response theory (IRT) approach (Böckenholt, 2012) and its multidimensional extension (Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013) to test and correct for response styles was applied to international large-scale assessment data – the Programme for International Student Assessment 2012 field trial – for the first time. The responses of n = 17,552 students at age 15 from 63 different countries to the two personality scales of openness and perseverance (student questionnaire) were examined, and bias from an extreme response style (ERS) and midpoint response style (MRS) was found. The aim of the study is not to report country level results but to look at the potential of this methodology to test for and correct response style bias in an international context. It is shown that personality scales corrected for response styles can lead to more valid test scores, addressing the "paradoxical relationship" phenomenon of negative correlations between personality scales and cognitive proficiencies. ERS correlates negatively with the cognitive domains of mathematics and problem solving on the country mean level, while MRS shows positive correlations.

Key words: Bifactor model, large-scale assessment, multidimensional item response theory (MIRT), rating scale, response style

[1] *Correspondence concerning this article may be addressed to*: Lale Khorramdel, PhD, Educational Testing Service, Princeton, NJ 08541, USA; email: lkhorramdel@ets.org

[2] National Board of Medical Examiners, Princeton, USA

[3] Educational Testing Service, Princeton, USA

[4] Professional Examination Service, Princeton, USA

Introduction

The current study is designed to provide an example of how to measure and correct for response styles (RS) in international large-scale assessment data using a multidimensional item response theory (MIRT) approach. The aim is to obtain more valid measures of noncognitive constructs. The primary purpose of international large-scale assessments such as the Programme for International Student Assessment (PISA) is to compare student achievement levels across educational systems in different countries, while a secondary purpose is to compare student responses on background questionnaires measuring noncognitive constructs. Such background questionnaires are administered in addition to the main cognitive assessment to gather more information about differences in cognitive proficiencies and provide additional information that can be used in scaling the cognitive assessment (e.g., different variables for dividing the sample to examine differential item functioning).

Another aim is to use this additional information for generating plausible values (von Davier, Gonzalez, & Mislevy, 2009; von Davier, Sinharay, Oranje, & Beaton, 2006), which are multiple imputations computed by combining the cognitive information (IRT scaling) and noncognitive information (principal components) through regression analyses into a posterior distribution – a process called population modeling. Plausible values are supposed to be more reliable measures than just the cognitive IRT-based proficiency estimations (von Davier, Gonzalez, & Mislevy, 2009), because not enough cognitive items can be administered in large-scale assessments (due to time limits on the assessment) to provide reliable measures based on the cognitive items alone. For these reasons, a valid measurement of noncognitive variables in large-scale assessment background questionnaires is important.

Noncognitive constructs are assessed through self-ratings, often using a rating or Likert-type scale as response format (where the test taker marks the degree of agreement with a statement on a scale providing a certain number of response options). Different problems can occur with self-ratings that jeopardize measurement validity. In high-stakes assessments (such as personnel selection), self-ratings can be biased by intentional response distortion (faking good), while so called response styles (RS) can occur in low-stakes assessments (such as international large-scale assessments) where the test results have no consequences for the test takers.

The issue with response styles

RS are defined as respondents' tendencies to give construct-irrelevant responses (Paulhus, 1991; Rost, 2004) to rating or Likert-type scales that harm the validity of survey data (Baumgartner & Steenkamp, 2001; De Jong, Steenkamp, Fox, & Baumgartner, 2008; Dolnicar & Grun, 2009; Weijters, Schillewaert, & Geuens, 2008) and the dimensionality of the measurement (Rost, 2004) because the responses are not related to the intended measurement construct. Depending on the number of response categories, different response styles may be identified (e.g., tendency to choose the scale midpoint, tendency to choose extreme responses, acquiescence, and so on). Consider an example where two respondents (A and B) respond to eight items of a personality scale measuring extraversion, using a rating scale with five response options (coded with 0 to 4). Respondent A chooses the extreme responses on the rating scale showing a response pattern of 0-4-0-4-0-4-0-4, while respondent B chooses only the midpoint of the rating scale displaying a response pattern of 2-2-2-2-2-2-2-2. Respondent A and B would receive the same raw score of 16 and the same level of extraversion would be assumed for both respondents if we looked at the raw score alone. The question is whether these two scores really mean the same. Do both indicate an average level of extraversion, or do they rather measure two different response styles (an extreme and a midpoint response style)? However, RS are not just random responses to single items. They are assumed to be largely stable individual characteristics within questionnaire administrations (Nunnally, 1967; Javaras & Ripley, 2007) and even across longitudinal survey data (Weijters, Geuens, & Schillewaert, 2010).

There could be different reasons for RS in low-stakes assessments. RS could be the result of low test-taking motivation or low acceptance of the assessment. They could also be the result of not understanding the question due to low test-taker reading proficiency, or due to poorly written items or complex item stems (ambiguous, inconsistent, too complex, etc.). Fatigue effects toward the end of the questionnaire, which is typically given after a challenging test of skills, also could result in RS.

Gender differences (De Jong, Steenkamp, Fox, & Baumgartner, 2008; Weijters, Geuens, & Schillewaert, 2010) and cultural differences with regard to RS have been found as well (Bachman & O'Malley, 1984; Buckley, 2009; Bolt & Newton, 2011; Chen, Lee, & Stevenson, 1995; Dolnicar & Grun, 2009; Hamamura, Heine, & Paulhus, 2008; Hui & Triandis, 1989; van Herk, Poortinga, & Verhallen, 2004). Women have shown a tendency to give more extreme responses than men, and respondents from different cultures have shown tendencies to use different types of response styles as well (e.g., respondents from Asian countries lean towards midpoint responses in contrast to respondents of European heritage). In addition to different groups having a tendency to show different response styles, some respondents might give valid responses and show response styles (e.g., due to a fatigue effect). Hence, response styles do not bias test results in a consistent way across groups of respondents and may be responsible for cultural and other group differences that have been found in prior studies. This is especially a problem in large-scale assessments (Buckley, 2009) that aim to compare individuals from different countries.

Findings that could be related to cross-cultural differences in RS are unexpected correlations between cognitive proficiency measures and noncognitive scales (e.g., attitudes). In PISA (2003, 2006), for example, a negative correlation across all countries (based on country-means) between mathematics self-concept and mathematics achievement was found, while the mean of within-country correlations was positive. This so-called "paradoxical relationship" phenomenon (Van de Gaer, Grisay, Schulz, & Gebhardt, 2012) was observed for a number of scales (e.g., mathematics interest, attitudes toward school), across different subjects and grades, and also in other international comparative studies, such as Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading and Literacy Study (PIRLS) (cf. Van de Gaer, Grisay, Schulz, & Gebhardt, 2012).

Thus, not testing and correcting for RS in large-scale assessments can lead to potentially meaningless group comparisons and artifacts when examining the relation between non-cognitive variables and cognitive proficiency scores.

IRT Approach to Measure and Correct for RS

Using IRT approaches to measure and correct for RS provides the opportunity for examining RS on the respondent (trait) level and the item level (Bolt & Johnson, 2009; De Jong, Steenkamp, Fox, & Baumgartner, 2008). The IRT approach applied in the current study is a multidimensional extension (Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013) of an approach proposed by Böckenholt (2012). The approach was successfully applied to empirical data (Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013) and validated using extraneous RS criteria, academic grades, and the relationship between self-concept and reading performance to prove its usefulness (Plieninger & Meiser, 2014). Likert type rating data are decomposed into multiple response subprocesses (binary pseudo items; BPIs) that can be used to separate RS from construct-related responses by applying simple-structure MIRT models. The advantages are that data can be tested for RS and that the approach provides a data structure that is easy to handle and with clear-cut interpretations. Böckenholt (2012) applies this approach to a single questionnaire scale where the construct of interest (the questionnaire scale) is modeled as a unidimensional factor and compared to different multidimensional response style factors.

Von Davier and Khorramdel (2013) and Khorramdel and von Davier (2014) extend this approach to questionnaires consisting of multiple scales (constructs; e.g. Five-Factor personality inventories) modeling different RS (separately from each other) as unidimensional factors that are tested against multidimensional constructs of interest (the questionnaire scales). This extended approach provides the opportunity to test whether a single RS – if present in the data – can be modeled as a unidimensional measure showing a consistent pattern across all questionnaire scales. It was shown that this extended approach can provide more detailed information about the dimensionality of RS measures. Furthermore, it was shown that RS measures are not always unidimensional but may be confounded with trait-related (or construct-related) responses (cf. Bolt & Johnson, 2009; Bolt & Newton, 2011; Johnson & Bolt, 2010). This problem can be addressed by using IRT models such as the bifactor model (von Davier & Khorramdel, 2013).

In the following, the example of a five-point rating scale is used to illustrate the difference of the approaches. Let's assume we are testing data for an extreme response style (ERS) and a midpoint response style (MRS), and decompose the rating data into three different BPIs similar to von Davier and Khorramdel (2013) and Khorramdel and von Davier (2014). One comprises responses in extreme response categories only (BPIs *e*; responses to extreme categories are coded as 1, the remainder as 0 or a missing value; e.g., 1-0-missing-0-1). One comprises responses to the middle category only (BPIs *m*; responses in the middle category are coded as 1, the remainder as 0; i.e., 0-0-1-0-0). One comprises construct-related responses corrected for RS (BPIs *d*; categories that are assumed to be biased by response styles are not used for scoring; negative scored items are rescored before being scored for BPIs; e.g., 0-0-missing-1-1). In cases where the middle category of the rating scale is chosen, BPIs *e* and *d* receive a missing value code. With this type of scoring, no dependencies are implied between BPIs *e, d,* and *m* (cf. Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013).

Table 1 shows how the three BPIs could be modeled using the Böckenholt (2012) approach on a single questionnaire scale. Table 2 shows how each single BPI can be modeled using the extended approach in the case of a questionnaire with five scales, using the example of *e*-items as a possible measure for ERS.

In Table 1, a one-dimensional IRT model is compared to a three-dimensional IRT model testing whether the BPI data can best be described by the questionnaire scale as a unidimensional factor, or by the questionnaire scale (corrected for ERS and MRS) and two RS factors (ERS and MRS) as multidimensional factors. In Table 2, a one-dimensional IRT model is compared to a five-dimensional IRT model to test whether BPIs *e* can best be described by the five questionnaire scales as multidimensional factors or by an ERS as a unidimensional factor. The same procedure can be applied to BPIs *m* to test whether a unidimensional MRS does exist in the data, and to BPIs *d*. The extended approach illustrated in Table 2 tests whether different BPIs as possible measures for RS are unidimen-

Table 1:
Loading Matrix for a Single Questionnaire Scale (Böckenholt Approach).

BPIs	1-Dimensional Model			3-Dimensional Model		
	Construct-factor	ERS factor	MRS factor	Construct-factor	ERS factor	MRS factor
d:	1	0	0	1	0	0
e:	1	0	0	0	1	0
m:	1	0	0	0	0	1

Table 2:
Loading Matrix for BPIs *e* for Multiple Questionnaire Scales (Extended Approach).

BPIs *e*	1-Dimensional Model					5-Dimensional Model				
	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5
e_1	1	0	0	0	0	1	0	0	0	0
e_2	1	0	0	0	0	0	1	0	0	0
e_3	1	0	0	0	0	0	0	1	0	0
e_4	1	0	0	0	0	0	0	0	1	0
e_5	1	0	0	0	0	0	0	0	0	1

sional measures across different questionnaire contents (cf. Khorramdel & von Davier, 2014); if this is not the case, a bifactor model can be applied to test whether the BPIs are measuring both an RS and (at least partly) the construct of interest (cf. von Davier & Khorramdel, 2013).

Thus, the information obtained with the extended approach adds to the information the Böckenholt approach provides. However, both approaches can be used to test if there are RS in the data or not, and to find a scoring method that is not affected (or affected less) by certain RS. The aim is to achieve a more comparable assessment of noncognitive constructs, more meaningful scores, and a more meaningful assessment of group differences.

The multidimensional extension of Böckenholt's approach so far was tested on two different sets of personality data (measuring the Big Five constructs of personality), showing promising results. Applying the approach on rating data based on items of the International Personality Item Pool (IPIP; Goldberg et al., 2006), analyses showed that RS can be measured as unidimensional factors after excluding items deviant from the one-dimensional model (Khorramdel & von Davier, 2014). Applying the approach on rating data based on items of the NEO Five-Factory Inventory (NEO-FFI; Borkenau & Ostendorf, 2008) analyses illustrated that a bifactor model can be used to detect the amount of variance explained by RS when BPIs are not a purely unidimensional measure of RS but measure both RS and construct-related responses (von Davier & Khorramdel, 2014). Both studies showed that a more valid measurement of the personality scales could be achieved after correcting for RS (BPIs *d*).

Because the approach appeared to work well for personality data, the current study aims to explore its usefulness in international large-scale assessment data and to explore the relationship between RS in noncognitive data and cognitive proficiencies. Moreover, it is of interest whether the correlations between noncognitive scales corrected for RS and cognitive scores can be improved in terms of the "paradoxical relationship" phenomenon (Van de Gaer, Grisay, Schulz, & Gebhardt, 2012). For this, data from the PISA 2012 field trial coming from two personality scales of the background questionnaire were used: the scales of perseverance and openness. Both scales are assumed to correlate positively with the cognitive domains.

Method

Similar to von Davier and Khorramdel (2013) and Khorramdel and von Davier (2014), the rating-scale response data were decomposed into BPIs to reflect RS and construct- or trait-related responses, and then modeled by applying unidimensional and MIRT models. The BPIs represent multiple nested response (sub)processes with regard to the response options of the rating scale and were modeled to present three latent variables per questionnaire scale: the target of measurement (trait-related responses), the tendency to use ERS, and the tendency to choose the middle response category. As two personality scales (perseverance, openness) were examined in the current study, and because prior research indicates that RS are consistent behavioral patterns, we assumed for most models that ERS and MRS are best represented by a variable where each describes RS as a unidimensional factor across the two personality scales. In addition to simple-structure IRT models, a bifactor model was applied to the BPIs to account for the possibility that they are measuring both RS- and trait-related responses.

All IRT models applied in this study are based on the two-parameter logistic (2PL) model (Birnbaum, 1968) and were estimated by applying the mixture general diagnostic modeling framework (MGDM; von Davier, 2008, 2010), which allows specification of a discrete mixture model with a hierarchical component (von Davier, 2010) using the software *mdltm* (von Davier, 2005) for multidimensional discrete latent traits models. The software provides marginal maximum likelihood estimates (MML) obtained using customary expectation-maximization methods (EM), with optional acceleration.

Simple-Structure Unidimensional and Multidimensional IRT Models

In a first step, simple-structure unidimensional and multidimensional IRT models based on the 2PL model (Birnbaum, 1968) were estimated. In addition to the Rasch model (Rasch, 1960) – which 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} (respondent's trait level) – the 2PL model postulates an item discrimination parameter α_i . For unidimensional scales, the model equation is defined as:

$$\mathrm{P}\ (\mathrm{x{=}1}|\theta_{\mathrm{v}}, \beta_i, \alpha_i) = \frac{\exp(\alpha_i(\theta_v - \beta_i))}{1 + \exp(\alpha_i(\theta_v - \beta_i))} \tag{1}$$

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

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

$$\mathrm{P}\ (\mathrm{x{=}1}|\theta_{\mathrm{vk}}, \beta_{ix}, \alpha_i) = \frac{\exp(\alpha_i(x\theta_{vk} - \beta_{ix}))}{1 + \sum_{y=1}^{m_i} \exp(\alpha_i(y\theta_{vk} - \beta_{iy}))} \tag{2}$$

Bifactor IRT model

In a second step, a bifactor model for binary data (Gibbons & Hedeker, 1992) was applied to the rating data. Each item measures a general dimension and one out of *K* specific dimensions. The general dimension represents the latent variable of central interest and accounts for the covariance among all items. The specific dimensions are integrated to account for additional dependencies (unique coherency) among particular groups of items. Statistical independence is assumed between all responses that are conditionally dependent on the general dimension and the specific dimensions. The latent variables typically are assumed to be normally distributed. The model equation for binary data can be written as follows:

$$P(y\,|\,\theta)=\prod_{i=1}^{I}P(y_{i(k)}\,|\theta_g,+\theta_k) \tag{3}$$

with y as vector of all binary scored responses, $y_{i(k)}$ as response on item i ($i=1,...,I$) in dimension k ($k=1,...,K$), θ_k as dimension-specific variable, and θ_g as general latent variable that is common to all items with $\theta=(\theta_g,\theta_1,...,\theta_k,...,\theta_K)$.

$\pi_i=P(y_{i(k)}=1\,|\,\theta_g,\theta_k)$ is related to a linear function of the latent variables through a (probit or logit) link function $g(\cdot)$:

$$g(\pi_i)=\alpha_{ig}\theta_g+\alpha_{ik}\theta_k+\beta_i \tag{4}$$

with β_i as intercept parameter for item i, and α_{ig} and α_{ik} as slopes or loadings of item i on the general and specific latent variables. Figure 1 shows an illustration of a bifactor model for the RS scores examined in the current study.

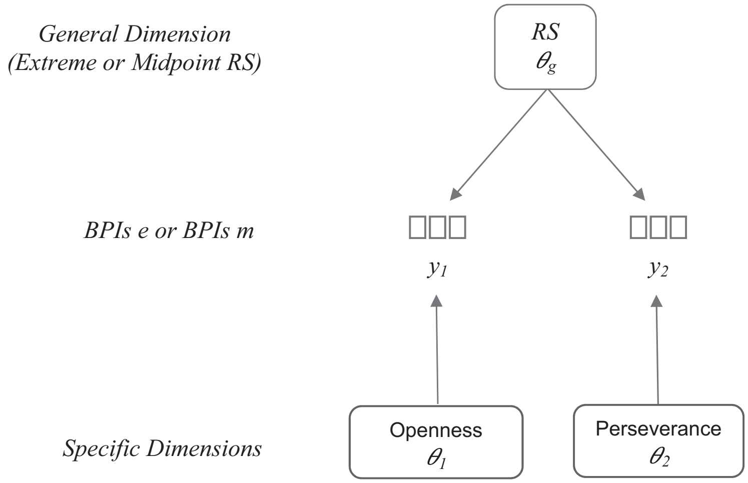

Figure 1:
Illustration of a bifactor model for BPIs *e* or BPIs *m* with regard to the personality scales of openness and perseverance (Note: arrows represent conditional dependencies).

Description of the dataset – PISA

The data used in this study come from PISA, a major international academic student survey of 15-year-old school populations (students in grade 7 or higher) in the domains of mathematics, reading, and science (sometimes accompanied by additional cognitive domains of interest such as problem solving). PISA has taken place in cycles every three years since 2000 with the aims of monitoring students' ability to use their knowledge and skills for meeting real-life challenges and to provide trend measures over time (e.g. OECD, 2014). In each cycle, one of the three domains is featured as the major domain and consists of trend and new items, while the others serve as minor domains and consist of trend items only. The item development for each domain is based on a framework provided by the Organisation for Economic Co-operation and Development (OECD). Each cycle consists of a field trial and a main survey, with the field trial serving as preparation for the main survey by testing the new items and survey procedures (technical platform, scoring, administration, and so on). In addition to the cognitive assessment of the three domains, PISA measures noncognitive scales and variables with background questionnaires (student, parent, and school questionnaires).

Sample

The current study is based on the data from the PISA 2012 field trial with mathematics as the major domain – more precisely, on the data from the student questionnaire for the personality scales of perseverance and openness. The sample for these two scales consists of n = 17,552 students at age 15 from 63 different countries, with 50.6% female (n = 8,884) and 49.4% male (n = 8,668) students.

Instrument

It was decided to take the two self-description scales of perseverance and openness from the PISA 2012 field trial student questionnaire because the multidimensional extension of Böckenholt's approach has shown to be promising when used on personality data and because these scales use a Likert-type scale as response format. "Perseverance" consists of 11 items measuring students' perseverance in situations in which they encounter cognitive challenges ("When confronted with a problem I give up easily."). "Openness" consists of 15 items assessing students' openness to problem solving ("I like to solve complex problems."). Both scales are measures of general drive and motivation, and both use a five-point Likert-type scale with five possible response options (not at all like me; not much like me; somewhat like me; mostly like me; very much like me). See more information about the two scales in the OECD (2014) report. The Cronbach's Alpha reliabilities and the IRT-based marginal reliabilities for Perseverance based on the original items are .73 and .78 respectively; the ones for Openness are .80 and .85. The Cronbach's Alpha reliabilities and the IRT-based marginal reliabilities for Perseverance based on the BPIs *d* (*d*-items; see more information in the next sections) are .77 and .65

respectively; the ones for Openness are .83 and .64. The estimated skill distribution correlation (latent correlations) between Perseverance and Openness (obtained from a two-dimensional 2PL model; see the method section) based on the original items is .69, the correlation based on the BPIs *d* is .65.

Procedure and design (BPIs)

Before the rating data were decomposed into BPIs, missing responses were coded as missing values, and negatively worded items were recoded (nine items) so that endorsement on the recoded negative items and the positively phrased items all indicated higher levels of the trait. Then, the five-category responses to the 26 items were decomposed assuming three latent variables with regard to the response process (cf. Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013). Thus, every questionnaire item was recoded into three different kinds of BPIs (see Table 3): one considering extreme positive and negative responses (*e*-items), one accounting for responses to the middle category (*m*-items), and one considering only positive (extreme and nonextreme) responses (*d*-items).

The score composed of *e*-items constitutes a possible measure for ERS, and the score composed of *m*-items a possible measure for MRS. Scale-wise scores based on *d*-items in turn aim to model the trait-relevant responses that are not biased by ERS and MRS (if ERS and MRS can be identified in the data using *e*-items and *m*-items). If the middle category of the rating scale was chosen, BPIs *e* and *d* received a missing value code. The reason is that with this type of scoring, no dependencies are implied between BPIs *e, d,* and *m* (cf. Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013).

Table 3:
Example for Coding Binary Pseudo Items (BPI).

Original Item (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

Hypotheses

We assume that in the case of distinct measurements of ERS and MRS in the data – that is, BPIs *e* and *m* are unidimensional measures of RS – one-dimensional simple-structure IRT models (where BPIs are modeled to load on one factor only) would fit these items across the two personality scales better than two-dimensional simple-structure IRT models (where BPIs are modeled to load on the two personality dimensions). Moreover, we would expect substantial correlations between the two personality scales based on BPIs *e* and BPIs *m,* indicating high consistency across the scales.

However, if this is not the case, we assume that a bifactor IRT model may indicate whether BPIs *e* and BPIs *m* are indicators of either RS or personality dimensions. The corresponding RS would then be defined as the general factor and the two personality dimensions as specific factors. In the case of items with higher loadings on RS, most variance should be explained by the general (RS) factor. In the case of items with higher loadings on their respective specific factors (personality dimensions), the general factor should explain less variance than each of the specific factors.

Moreover, if there is evidence of ERS and MRS in the data, scores for the two personality dimensions based on *d*-items should not be (much) affected by ERS and MRS, and thus should be a better measurement of the two personality dimensions than scores based on the original five-point rating scale items. More specifically, we assume that correlations between the two personality dimensions based on BPIs *d* should be lower compared to those based on the original score. We also assume that scores based on BPIs *d* show an impact on the problem of negative between-country correlations between the personality dimensions (perseverance, openness) and the cognitive dimensions (mathematics, problem solving), with the negative correlations disappearing and being replaced by positive correlations.

Results

To test whether the described IRT approach (Böckenholt, 2012) and its multidimensional extension (Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013) can be useful in international large-scale assessment data, both approaches were applied to data from the PISA 2012 field trial. It was examined whether BPIs *e* (extreme responses) and BPIs *m* (midpoint responses) are measurements of RS, and whether BPIs *d* are a less biased measurement of the two personality dimensions. Simple-structure IRT and MIRT models with either unidimensional or multidimensional RS factors (cf. Khorramdel & von Davier, 2014), as well as a bifactor IRT model to account for items with loadings on ERS/MRS and with loadings on one of the personality dimensions (cf. von Davier & Khorramdel, 2013) were estimated. For model evaluation, the Akaike information criterion (AIC; Akaike, 1974), and the Bayesian information criterion (BIC; Schwarz, 1978) were used.

Furthermore, the influence of RS on comparing different countries was examined by investigating the relationship between ERS, MRS, and personality scores based on BPIs *d*, as well as based on the original items and the cognitive domains of mathematics and problem solving. For this purpose, correlations between the different variables and multiple linear regression analyses were calculated.

Simple-structure IRT and MIRT models

IRT models with multidimensional RS factors (Böckenholt's approach). In a first step, Böckenholt's approach was used to examine if there are response styles (ERS, MRS) in the rating data that can be differentiated from the two personality dimensions. Thus, two-, three-, and four-dimensional IRT models (all based on the 2PL model) were estimated and compared to one another. BPIs were assigned by type (*e, m, d*) to the dimensions in the three-dimensional model, while in the two-dimensional model, all BPI types were assigned to the two personality dimensions. For the four-dimensional model, BPIs of type *d* were assigned to the two personality dimensions (or factors), but BPIs *e* and *m* to a third and fourth (RS) factor, respectively.

Results show that the three-dimensional model fits the data better than the two-dimensional model, and that the four-dimensional model fits the data best. Thus, it can be assumed that BPIs *d* are measuring the two personality dimensions, and that BPIs *e* and *m* are measuring ERS and MRS. Detailed results are given in Table 4.

IRT models with unidimensional RS factors (multidimensional extension of Böckenholt's approach). As the analysis with multidimensional RS factors shows that RS can be found in the data, further analyses at the BPI level with unidimensional RS factors – the multidimensional extension of Böckenholt's approach – were computed. A one-dimensional IRT model was estimated with either BPIs *e* or *m* assigned to one (RS) factor, and two-dimensional IRT models with BPIs *e* or *m* assigned to the two personality factors.

Comparing the results of the one-dimensional models with those of the two-dimensional models shows that BPIs *m* are a unidimensional measure of RS, as the one-dimensional model fits the data better than the two-dimensional model. But this does not apply to BPIs *e*, where the two-dimensional model shows a better model fit than the one-dimensional model. Detailed model-fit statistics are given in Table 5.

Still, it cannot be assumed that BPIs *e* are pure measures of the two personality dimensions because the differences between the model fit indexes (AIC, BIC) of the two- and one-dimensional models are not large.

Therefore, we decided to test the hypothesis that the BPIs *e* might measure both RS and construct-related responses using a bifactor IRT model.

Table 4:
Results of the 2-, 3-, and 4-Dimensional Simple-Structure IRT Models with Multidimensional RS Factors, Including All BPI Types.

All 5 Scales, Items: e,d,m	4-D Model	3-D Model	2-D Model
AIC Index	1219953.19	1221426.55	1257257.28
BIC Index	1221251.27	1222693.54	1258500.94
log-penalty (model based, per item)	0.551	0.552	0.568

Table 5:
Results of the 1- and 2-Dimensional Simple-Structure IRT Models with Unidimensional RS Factors, and the Bifactor Model Separately for each BPI Type.

	2-D Model	1-D Model	Bifactor Model
All 2 scales, e-items:			
AIC Index	383635.54	384303.48	382973.81
BIC Index	384101.91	384738.76	383634.51
log-penalty (model based, per item)	0.587	0.588	0.586
All 2 scales, m-items:			
AIC Index	509355.76	509112.71	508782.17
BIC Index	509822.14	509548.00	509442.87
log-penalty (model based, per item)	0.562	0.562	0.562
All 2 scales, d-items:			
AIC Index	334325.92	336618.72	-----
BIC Index	334792.30	337054.00	-----
log-penalty (model based, per item)	0.511	0.515	-----

Bifactor IRT model

To examine if BPIs *e* have loadings on both RS and personality dimensions, and to examine how much variance is explained by each factor, we computed a bifactor IRT model (based on the 2PL model, with the assumption of factor independence). This model allows items to load on two factors at the same time. BPIs *e* were assigned to the two personality factors and one RS factor. In the bifactor model, the general dimension (here the ERS factor) reflects the covariance among items, while the independent specific dimensions (here the two personality factors) reflect the unique coherency among particular groups of items. Items depend directly on the general dimension.

Results (see Table 5) show that the bifactor IRT model fits the BPIs *e* better than the two-dimensional simple-structure IRT model. Table 6 gives an overview of the amount of variability in the general vs. the specific factors, given that the slope parameters were normalized for each of the factors. This comparison gives insight into how much of the total respondent-based variance in the bifactor IRT model is explained by each factor. It can be seen that the largest variance falls to the general (or ERS) factor (ERS: 1.936, perseverance: 0.342, openness: 0.263) indicating that the BPIs *e* are mainly affected by the ERS factor, and to a lesser extent to the domain specific residual factors. That is, it appears that the ERS factor (defined as the common source of response variance across domains) explains more of the response variance than each of the domain-specific residual factors. In addition, the ERS factor shows a higher IRT-based marginal reliability than the two specific factors (ERS: .763, perseverance: .132, openness: .068).

A similar finding can be reported for BPIs *m* (see Table 5 and 7): a bifactor model fits the data relatively better than the one-dimensional model. Again, the general (or MRS) factor explains more variance than the specific factors (MRS: 0.677, perseverance: 0.163, openness: 0.129) and shows a higher IRT-based reliability (MRS: .692, perseverance: .067, openness: .087).

Table 6:
Estimated Variances (SD^2) of the Specific Factors (Perseverance, Openness) and the General Factor (ERS Factor for e-items, MRS Factor for m-items) according to the Bifactor Model.

Bifactor Model	Perseverance	Openness	ERS/MRS
e-items, SD^2	0.342	0.263	1.936
m-items, SD^2	0.163	0.129	0.677

Dimensionality of BPIs *d*

Because it could be shown that ERS and MRS exist in the data, BPIs *d* might be a better measurement of the two personality scales than the original scored items considering BPIs *d* are not (or at least much less) biased by ERS and MRS. To test this hypothesis, the following analyses were conducted.

A two-dimensional model (BPIs were assigned to the two personality factors) was compared to a one-dimensional model (BPIs were assigned to one factor only), both based on BPIs *d* (again based on the 2PL model), to examine whether BPIs *d* are an adequate

Table 7:
Estimated Intercorrelations of the Score Distributions of the Personality Dimensions According to the 2-Dimensional Simple-Structure IRT model for BPIs *e* (Extreme Responses), BPIs *c* (Midpoint Responses), BPIs *d* (Construct-Related Responses), and for the Original Items (Original 5-point Likert Scale)

	Perseverance
e-items (extreme)	
Openness	0.72
m-items (trait)	
Openness	0.63
d-items (midpoint)	
Openness	0.65
original NEO-FFI items (5-point rating scale)	
Openness	0.69

measurement of the two personality scales. Moreover, scale intercorrelations among the two personality scales based on BPIs *d* were calculated and compared to the scale intercorrelations based on the original items. Results (see Table 5) show that BPIs *d* are relatively better fitted with the two-dimensional model than the one-dimensional model. In addition, the scale intercorrelations based on BPIs *d* are slightly lower than the scale intercorrelations based on the original items (see Table 7). The two personality dimensions are supposed to be two different distinct measures so that lower correlations between these scales would be preferable. In this regard, BPIs *d* show to be a more appropriate measure for the two personality dimensions than the original scored items.

Validation: correlations between BPI factors and cognitive domains

As stated earlier in this paper, one important aim of international large-scale assessments is to provide fair comparisons of test results between different countries. This is not possible if the reported scales are biased by RS. Therefore, it was examined whether correlations between noncognitive scales and cognitive scores can be improved in terms of the "paradoxical relationship" phenomenon (Van de Gaer, Grisay, Schulz, & Gebhardt, 2012) by using BPIs *d* (scores corrected for ERS and MRS) instead of the originally scored Likert-type items. For this, the personality scale scores (IRT person parameters) for openness and perseverance were estimated based on both BPIs *d* and the original Likert-type items. Then, the correlations between the different personality scale scores and the scores (IRT person parameters) of the two cognitive domains of mathematics and problem solving were calculated. Moreover, the relationship between RS measures (ERS, MRS) and the two cognitive domains was of interest. The following Pearson correlations were computed between the RS and personality variables and the cognitive domains:

1. Correlations across countries based on estimates of country averages (see Table 8).
2. Average of the within-country correlations based on individual test scores (person parameters) for each variable (see Table 9).

The mean correlations across countries (see Table 8) show medium to relatively high negative correlations (-.57 to -.66) for personality scales based on the original Likert-type items, but only low to medium negative correlations (-.16 to -.36) for scales based on BPIs *d*. The average of within-country correlations (see Table 9) show low but positive correlations (.10 to .26) for personality scales based on both BPIs *d* and original Likert-type items. Note that the single within-country correlations are not reported here because the sample sizes per country of the field trial data were small in most cases, not allowing stable measures on the within-country level; therefore, only the average of all within-country correlations is used for interpretation and discussion. Moreover, Table 8 illustrates that the ERS measure shows a medium negative correlation with the cognitive domains (-.54 and -.59), while the MRS measure shows a medium positive correlation (.61 and .63). The correlations of RS measures with the cognitive domains are smaller when looking at the single within-country correlations – the average of within-country correlations between both ERS and MRS and the cognitive domains are close to zero and negative (-.07 to -.03; see Table 9) – but using the country-level correlations elevates the problem, resulting in higher correlations.

Decomposing the rating data into RS factors and BPIs *d* is shifting some of the problem away from the personality measures (BPIs *d*) into the RS measures, resulting in lower negative correlations between personality scales and cognitive domains. These findings show clearly that cultural differences in RS are affecting country-level measures and that a score corrected for RS can lead to a potentially more meaningful comparison.

Table 8:
Mean Correlations across Countries: Correlations between the Country Based Means of the Person Parameters of Problem Solving/Mathematics and Scales Based on BPIs and Original Scored Items.

	Extreme RS	Midpoint RS	Perseverance *d*-items	Openness *d*-items	Perseverance original items	Openness original items
***Mathematics** – correlation based on country-means*	-.59	.63	-.16	-.21	-.57	-.58
Sig. (2-tailed)	.00	.00	.22	.11	.00	.00
***Problem Solving** – correlation based on country-means*	-.54	.61	-.24	-.36	-.65	-.66
Sig. (2-tailed)	.00	.00	.14	.02	.00	.00

Table 9:
Average of Within-Country Correlations between Cognitive Domains and Scales Based on BPIs and Original Scored Items.

	Extreme RS	Midpoint RS	Perseverance d-items	Openness d-items	Perseverance original items	Openness original items
***Mathematics** - Average of country-based correlations*	-0.03	-0.07	0.10	0.16	0.21	0.26
***Problem Solving** - Average of country-based correlations*	-0.03	-0.04	0.16	0.21	0.18	0.23

Note: All values are based on individual person parameters obtained from the IRT calibration

Discussion

A relatively new IRT approach (Böckenholt, 2012) and its multidimensional extension (Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013) to test and correct data for different RS were tested on an international large-scale assessment dataset for the first time. The two approaches showed promising results when used on data measuring the Big Five personality constructs in prior studies (Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013) and were applied to data coming from the student questionnaire of the PISA 2012 field trial. In the current study, we focused on the two personality scales of openness and perseverance. Response styles – or construct-irrelevant responses – as a possible result of fatigue effects, low test-taking motivation, or lack of understanding the questions are especially a problem in low-stakes assessments and studies using large-scale assessments. They can harm the validity of the measurement and lead to biased survey results and group comparisons. The examined IRT approaches aim to provide fair and valid measures of noncognitive constructs when rating scales are used as response formats. Openness and perseverance were both assessed using a five-point Likert-type response scale. Using the IRT approaches, it was examined whether ERS and MRS are present in the data, and how measures of ERS and MRS relate to the cognitive constructs for mathematics and problem solving.

Similar to an IRT approach proposed by Böckenholt (2012), the rating data for openness and perseverance were decomposed into BPIs and then modeled with different IRT models (all based on the 2PL model). The rating data were decomposed into three different BPIs: *e* accounting for extreme responses, *m* accounting for responses to the midpoint of the rating scale, and *d* accounting for response categories not biased or (at least) less biased by a possible ERS or MRS.

Findings for the application of the IRT approach and its multidimensional extension to the PISA 2012 field trial data

First, according to Böckenholt's approach, BPIs as possible measures for RS were modeled as multidimensional factors in a multidimensional IRT model and compared to an IRT model with one factor for each of the personality scales. A three-dimensional 2PL model (BPIs *e* with loadings on one ERS factor, BPIs *m* with loadings on one MRS factor, and a third factor for BPIs *d*) was compared to a two-dimensional 2PL model (all three BPIs with loadings on two factors representing the two personality constructs) and a four-dimensional 2PL model (BPIs *e* with loadings on one ERS factor, BPIs *m* with loadings on one MRS factor, and BPIs *d* with loadings on two factors for the two personality constructs). Results show that the four-dimensional model fits the BPI data best, indicating that BPIs *e* and *m* are measures of ERS and MRS, respectively, and that BPIs *d* are measures of the two personality constructs. According to these findings, it can be assumed that ERS and MRS both exist in the PISA student questionnaire data (at least for the two examined personality scales).

Second, the multidimensional extension of Böckenholt's approach (Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013) was applied to the BPI data. BPIs were examined separately by modeling them as unidimensional RS measures and comparing them to multidimensional factors representing the two personality constructs. For each BPI type, a one-dimensional 2PL model (with BPIs loading on one factor representing a RS) was compared to a two-dimensional 2PL model (with BPIs loading on two factors representing the two personality scales). Results show a unidimensional measure of MRS as BPIs *m* are better fitted with a one- than a two-dimensional model, while this is not true for BPIs *e*, which are slightly better fitted with a two-dimensional model. This could lead to the assumption that the data are biased by MRS but not by ERS. However, the difference of the model fit criteria between the one- and the two-dimensional model is small and the correlations between openness and perseverance are lower when based on BPIs *d* (correcting for MRS and ERS) than when based on the original Likert-type items (not correcting for ERS and MRS). Therefore, the hypothesis was tested that BPIs *e* might measure both trait-related responses and an ERS.

To test whether BPIs *e* have loadings on the two personality factors and an ERS factor, a bifactor IRT model (based on the 2PL model, with the assumption of factor independence) – allowing each item to have loadings on two factors, a general and a specific factor, at the same time – was applied to BPIs *e* and compared to the one- and two-dimensional models. Results show that the bifactor model fits the data relatively best, supporting this hypothesis. Moreover, ERS as the general factor seems to explain more variance than the personality (or specific) factors in the bifactor model, also showing a much higher IRT-based marginal reliability. These results indicate that there is a clear bias of ERS in the data, but measurement is not straightforward. The same results could be found for BPIs *m*: The bifactor model showed the relatively best model fit, with the MRS factor explaining more variance and showing a much higher reliability than the two personality factors. Thus, BPIs *e* and *m* show to be mainly measures of ERS and MRS but also seem to comprise construct- (personality) related responses to a small extent. A possible explanation could be that there are different groups of respondents, with some showing RS and others not.

Because the student questionnaire data seem to be biased by MRS and ERS, BPIs *d* might be a less biased measurement of openness and perseverance than the original Likert-type items. It could be shown that BPIs *d* are better fitted with a two-dimensional 2PL model than with a two-dimensional 2PL model. Moreover, as noted above, the correlation between openness- and perseverance-based BPIs *d* are lower than the correlation based on the original Likert-type items. Lower correlations between the two personality scales are desirable because they indicate that the two scales are two different (rather distinct) measures.

Validation Findings: correlations between BPI factors and cognitive domains

To validate the IRT approaches and their application to the PISA data, correlations between the RS factors and personality scales with the cognitive domains of mathematics and problem solving were calculated. Results show that a cultural bias of RS on country-mean correlations can be assumed because medium to large negative correlations between personality scales based on Likert-type items occur when mean country correlations are calculated, while low positive correlations can be seen on the within-country level. Moreover, while ERS and MRS measures show very low correlations (close to zero) with the cognitive domains on the within-country level, they show much higher correlations on the country-mean level, leading to the assumption of a cultural RS bias that gets aggregated when performing analyses across all countries. It could further be illustrated that using BPIs *d* instead of the original Likert-type items can lead to less biased correlations between the personality scales and the cognitive domains on the country-mean level, making progress in terms of the "paradoxical relationship" phenomenon (Van de Gaer, Grisay, Schulz, & Gebhardt, 2012) to some extent, but not fully because correlations remain negative. Most of the bias seems to be moved to the RS measures that are separated from the personality measures by decomposing the rating data into the BPIs.

Limitations and recommendations

The current study was a first application of an IRT approach using BPIs to test and correct for RS in international large-scale assessment data (feasibility study). The proposed approach showed promising results with regard to detecting and correcting for RS bias in noncognitive rating data, producing more reliable and valid measurements. However, it also showed that the measurement of RS is not always straightforward. Depending on the sample and assessment characteristics, different types of RS may be present. Hence, a certain type of RS should never just be assumed. Each data set has to be tested, using different models, to determine whether RS are present before attempting a correction. Moreover, the application of the proposed approach to other international studies should be considered to examine whether issues might arise that may make extensions of the current approach necessary. As the PISA field test always consists of smaller samples than the main study, it would be interesting to look at main study data as well. It should also be investigated to see if the findings of the current study can be generalized to other noncognitive constructs measured with rating or Likert-type scales and whether similar correlations of ERS and MRS with other cognitive domains can be found. Moreover, models that account for latent classes of respondents with different response patterns could optimize the application of the BPI approach to measure and correct for RS.

Another limitation of the current study is the use of data from the field trial instead of data from the main survey, which were unavailable at the time of analyses. It would be interesting to look at results with data from the main survey, which always uses larger samples. Since the current study is a feasibility study and does not aim to make country comparisons, the field trial data should be sufficient. Nevertheless, it has to be stressed that the field trial data only provide proxies instead of real proficiencies, and that the smaller samples are not representative for the countries.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Bachman, J. G., & 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.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38,* 143-156.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M., & Novick, M. R. (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison–Wesley.

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

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.

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

Borkenau, P. & Ostendorf, F. (2008). *NEO-FFI – NEO-Fünf-Faktoren-Inventar nach Costa und McCrae* [*NEO-FFI – NEO-Five-Factor-Inventory following Costa and McCrae*]. Handanweisung. (2. Aufl.). Göttingen, Germany: Hogrefe.

Buckley, J. (2009, June). *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., & 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.

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

Dolnicar, S. & Grun, B. (2009). Response style contamination of student evaluation data. *Journal of Marketing Education, 31*, 160-172.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423-436.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Hamamura, T., Heine, S. J., & Paulhus, D. L. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual differences, 44*, 932-942.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*, 296–309.

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

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

Khorramdel, L., & 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.

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

OECD (2014), *PISA 2012 Results: Creative Problem Solving: Students' Skills in Tackling Real-Life Problems* (Vol. V). Paris: Author. doi:10.1787/9789264208070-en

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

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement, 20, 1–25*.

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

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* [*Textbook test theory – test construction*] (2nd edition). Bern, Switzerland: Huber.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of statistics, 6*, 461-464.

Van de Gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and self-confidence across countries. *Journal of Cross-Cultural Psychology*, *43,* 1205-1228.

van Herk, H., Poortinga, Y. H., & 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.

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 & K. M. Samuelson (Eds.), *Advances in Latent Variable Mixture Models*. Information Age Publishing.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8-28.

von Davier, M., Gonzalez, E. & Mislevy, R. (2009) What are plausible values and why are they useful? In *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments, Vol. 2*. Retrieved from IERI website: http://www.ierinstitute.org/IERI_Monograph_Volume_02_Chapter_01.pdf

von Davier, M. & 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, & 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., & Carstensen, C. H. (2007). Introduction: Extending the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models: Extensions and applications* (pp. 1-12). New York, NY: Springer.

von Davier, M. Sinharay, S., Oranje, A. & Beaton, A. (2006) Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao and S. Sinharay (Eds.), *Handbook of Statistics (Vol. 26): Psychometrics*. Amsterdam, Netherlands: Elsevier.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods, 15*, 96-110.

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