Lu and Bolt *Large-scale Assessments in Education* (2015) 3:2
DOI 10.1186/s40536-015-0012-0

Large-scale Assessments in Education
a SpringerOpen Journal

METHODOLOGY Open Access

Examining the attitude-achievement paradox in PISA using a multilevel multidimensional IRT model for extreme response style

Yi Lu[1*] and Daniel M. Bolt[2]

* Correspondence:
yi.lu0519@gmail.com
[1]ACT, Inc., 500 ACT Drive, 52243 Iowa City, IA, USA
Full list of author information is available at the end of the article

Abstract

In this paper, we consider a two-level multidimensional item response model that examines country differences in extreme response style (ERS) as a possible cause for the achievement-attitude paradox in PISA 2006. The model is an extension of Bolt & Newton (2011) that uses response data from seven attitudinal scales to assess response style and to control for its effects in estimating correlations between attitudes and achievement. Despite detectable variability in ERS across countries and detectable biasing effects of ERS on attitudinal scores, our results suggest that the unexpected between-country correlation between attitudes and achievement is not attributable to country differences in ERS. The remaining between-country correlations between mean attitudes and mean achievement once controlling for ERS can be explained by the observation that (1) despite detectable country differences, most variability in ERS occurs within, as opposed to between, countries, and (2) ERS appears to be only weakly correlated with achievement. The methodological approach used in this paper is argued to provide an informative way of studying the effects (or lack thereof) of cross-country variability in response style.

Keywords: Multilevel models; Multidimensional IRT; Extreme response style (ERS); PISA

Background

One objective of cross-cultural assessments such as PISA is to better understand achievement differences across countries. Recent administrations of PISA and TIMSS have included survey instruments that have the potential to inform about cross-cultural differences in student attitudes toward different subject areas. The focus area in PISA 2006 was science, and attitudinal surveys on the assessment considered several different aspects of attitudes related to science (e.g., enjoyment of science, perceived value of science, etc.). The different content areas and numbers of items across the scales studied in this paper are summarized in Table 1. Items on the surveys were answered via self-report using Likert rating scales that had four scale points, ranging from 1 = Strongly Agree to 4 = Strongly Disagree, such that lower overall scores imply a more positive attitude toward science. Table 2 presents example items from two of the attitude scales, the Enjoyment and Value subscales, respectively.

In studying the relationships between attitudes and achievement between and within countries, one frequently occurring observation in both PISA and TIMSS is

Table 1 Seven subscales in PISA 2006 science attitudes survey

Subscales	Number of items
Science Enjoyment (ENJ)	5
Science Value (VAL)	10
Environmental Responsibility (ENV)	7
Usefulness for Science Career (USE)	4
Science in Future A (FUTA)	4
Science in Future B (FUTB)	5
Science Learning (LRN)	6

an achievement-attitude paradox. Specifically, when evaluated at the country level, mean levels of achievement in a subject area correlate with mean attitudes regarding that subject area but in the opposite direction to what is expected. Specifically, the countries that perform best on the subject achievement metric appear on average to have more negative feelings about the subject area. Moreover, such correlational effects are the opposite of what is seen when studying the relationships within countries, where the anticipated positive relationship between attitudes and achievement is regularly observed. Figures 1 and 2 visually present the relationships between science achievement and science enjoyment, within- and between-country, respectively, in PISA 2006. (Note that due to the coding of responses on the attitudinal scales, a negative correlation between survey scores and achievement implies a positive relationship between attitudes and achievement, and vice-versa). Using the country of Australia as an example, as shown in Fig. 1, the more enjoyment students have in learning science, the more likely they are to have higher science achievement. The same pattern is seen within virtually all countries. However, as shown in Fig. 2, the between-country correlation is strongly in the opposite direction. The correlations at the country level for each of the other six science attitude subscales with achievement show a similar pattern.

This apparent paradox has been observed and investigated in a number of previous studies (e.g., Buckley, 2009; Bybee & MaCrae, 2007; Loveless, 2006; Van de Gaer & Adams 2010). Attempts to explain the paradox have typically focused on other differences between countries that may explain how and why countries might differ in their responses to the attitudinal surveys. For example, one explanation is a "big fish little pond effect" (Marsh, Seaton, Trautwein, Ludthke, Han, O'Mara & Craven, 2008), whereby students who attend schools where average ability is lower have higher academic self-concept than students who attend schools where average ability is higher. Another theory attributes the effect to potential response style differences between countries, namely, stylistic differences in how respondents across countries use rating scales when responding to attitudinal survey

Table 2 Examples of PISA 2006 science attitudes assessment

Science Enjoyment

Item 1: I generally have fun when I am learning < broad science > topics

1 = strongly agree 2 = agree 3 = disagree 4 = strongly disagree

Science Value

Item 1: Advances in < broad science and technology > usually improve people's living conditions

1 = strongly agree 2 = agree 3 = disagree 4 = strongly disagree

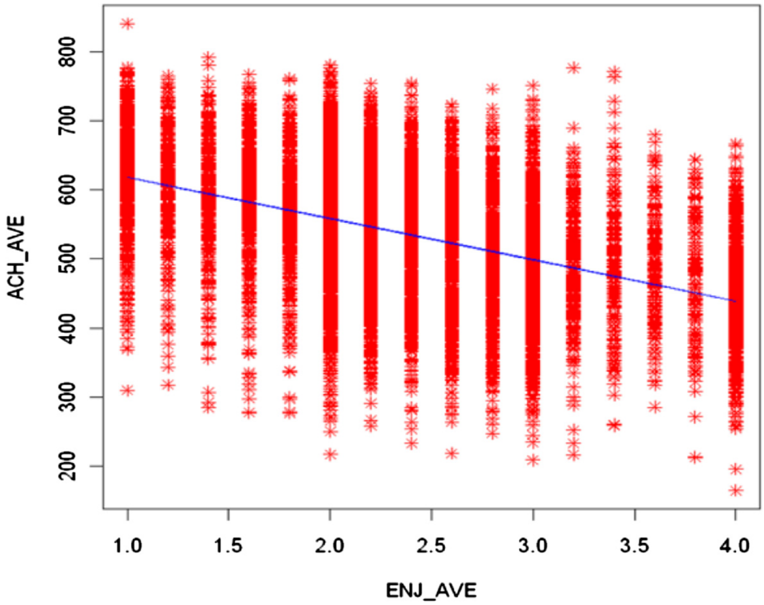

Fig. 1 Scatter plot of science achievement and enjoyment—Australia (corr = − .46)

items (Buckley, 2009). Examples of response style tendencies include extreme response style (ERS, i.e., a tendency to respond using only the extreme endpoints of a rating scale), anti-ERS (i.e., a tendency to avoid the extreme endpoints of a rating scale), acquiescent response style (i.e., a tendency to always agree with items, regardless of content), and many others. The response style theory seems plausible owing to the fact that cross-cultural variability in response styles is well documented (Clarke, 2000; Johnson, Kulesa, Cho & Shavitt, 2005; Van Herk, Poortenga & Verhallen, 2004) and that response styles frequently correlate with

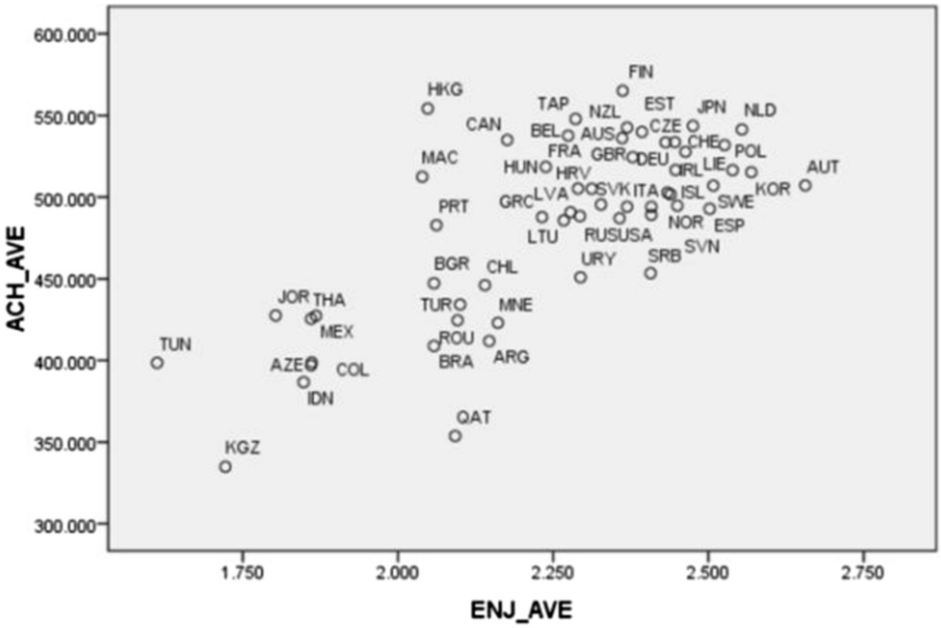

Fig. 2 Scatter plot of science achievement and enjoyment—between countries (corr = .74)

variables related to achievement, such as education and cognitive ability level (Meisenberg & Williams, 2008; Weijters, Geuens & Schillewaert, 2010).

In this study, we formally investigate the possibility that extreme response style (ERS) differences across PISA countries may explain the unexpected correlations between mean achievement and mean attitudes on the country level. ERS refers to the tendency of selecting extreme endpoints of a rating scale regardless of item content, such as 1 = strongly agree or 4 = strongly disagree on a 1–4 Likert scale. The correlations observed between ERS and education in prior work, are in a direction that could in theory explain the peculiar between-country correlation between attitudes and achievement seen in PISA. Specifically, higher ERS tends to be associated with lower education. As the PISA attitudinal scales tend (on average) to elicit "agree" responses across items, ERS will tend to produce more extreme levels of agreement (i.e., scores on the attitudinal scales that are generally biased downward), thus making it possible that more positively reported attitudes would be seen in the context of lower educational achievement.

In the framework of item response theory (IRT), a multidimensional IRT (MIRT) model proposed by Bolt and Johnson (2009) provides a technique for modeling and controlling for the effects of ERS. The MIRT model incorporates response styles as explicit statistical dimensions that influence item responding. As a result, it can account for the simultaneous influence of the substantive and ERS traits on response category selection. Under the MIRT model, the probability that a respondent selects category k on item j, given a substantive trait level θ_1 and an extreme response style trait θ_{ERS} is given as

$$P\left(Y_j = k|\theta_1, \theta_{ERS}\right) = \frac{\exp\ \left(a_{jk1}\theta_1 + a_{jk2}\theta_{ERS} + c_{jk}\right)}{\sum\limits_{h=1}^{K} \exp\ \left(a_{jh1}\theta_1 + a_{jh2}\theta_{ERS} + c_{jh}\right)} \tag{1}$$

where a and c represent category slope and intercept parameters associated with item j. The distinct θ_1 and θ_{ERS} traits reflect the influence each trait has on the propensity toward selecting a given score category. To make the model applicable for modeling response style, fixed value constraints are applied to the category slope parameters across items. For a four-category Likert item for example, the a parameters for θ_1 might be fixed at −3, −1, 1, 3, while the a parameters for θ_{ERS} could be fixed at 1, −1, −1, 1 for all items, so as to allow θ_1 to be interpreted as the substantive trait and θ_{ERS} as an extreme response style trait. The intercept parameters are generally freely estimated across items but subject to a normalization constraint $\sum\limits_{k} c_{jk} = 0$. The latter constraint addresses the lack of identifiability of the individual category intercepts, as the intercept for each category only defines a propensity toward selecting the category relative to the other categories. A related model is presented in Johnson and Bolt (2010).

The model in Equation (1) can be viewed as a multidimensional extension of Bock's (1972) nominal response model and can be estimated using the software package Latent Gold (Vermunt & Magidson, 2005, 2008), among others. Compared with other model-based approaches for ERS, a distinguishing characteristic of the MIRT approach is that it can psychometrically account for how both substantive and ERS traits combine to affect response category selection. Relative to earlier studies on response styles which generally only verified the existence of response styles, the MIRT model is also capable of correcting ERS bias in providing substantive trait estimates.

Bolt and Newton (2011) illustrated the method in relation to PISA 2006 data. Based on the example items in Table 2, Table 3 presents some hypothetical examples of response patterns and the corresponding trait estimates as observed by Bolt & Newton (2011). The response patterns in Table 3 provide examples showing high, moderate, and low estimated levels of θ_{ERS}. Respondents with high estimated θ_{ERS} consistently use extreme categories in their responses, while respondents with low θ_{ERS} predominantly use the intermediate score categories. Respondents with moderate estimated θ_{ERS} use a mix of categories. One advantage of using a statistical modeling approach is that it becomes possible to quantify the biasing effects of ERS, and therefore also make corrections for the effects of bias at the scale score level. Bolt & Johnson (2009) suggest evaluating the biasing effects by examining the expected score on the survey for a hypothetical respondent with the same θ_1 and a θ_{ERS} level of 0, and refer to this expected score as a bias-corrected score. As shown in Table 3, once taking θ_{ERS} into account, the differences between the original and bias-corrected total scores can be quite substantial. Using simulated data based on PISA scales, Bolt and Newton (2011) also showed that the model yields improved estimates of θ_1 when simultaneously estimating θ_{ERS}. The modeling approach provides a general framework within which it may be possible to examine whether controlling for ERS effects changes the unexpected attitude-achievement correlations seen across countries in PISA.

Bolt & Newton (2011) also showed how this general approach can be extended to take into account multiple scales that use the same response format but measure different substantive traits. The extension makes the assumption that the ERS tendency is constant across traits, an assumption that appears plausible based on prior work (Weijters et al., 2010; Wetzel, Carstensen, & Böhnke, 2013) and is also consistent with a theory that extreme response style underlies the paradoxical country-level correlations between achievement and multiple attitudinal scales. In this paper, we extend the Bolt & Newton (2011) approach in additional ways to examine the attitude-achievement paradox. First, we extend the approach to include additional scales. As noted, the PISA 2006 administration included seven such scales using the same four-point rating scale. Second, and more importantly, we generalize the model to include a multilevel structure (i.e., students within countries) and include an external variable (achievement) so as to simultaneously study the covariance structure of the attitudinal scales, achievement, and ERS both within and between countries. This generalized model is discussed in the next section.

It is important to acknowledge that the model in (1) represents just one way in which ERS has been conceptualized in the research literature. As noted, the model emphasizes the simultaneous influence of both the substantive trait and response style on a respondent's selection of extreme versus less extreme categories. Several alternative models (e.g., mixture models—Rost, Carstensen, & von Davier, 1997, or latent class factor models—Moors, 2003) adopt a similar conceptualization but use a latent class representation of the substantive and/or response style traits. Still other methods, including that of Bockenholt

Table 3 Example of item response patterns

Respondent	Item responses	$\hat{\theta}_1$	$\hat{\theta}_{ERS}$	Total score	Bias-corrected total score
1	1414414414	0.84	2.05	53	41
2	2421322131	−0.45	0.00	32	32
3	2332322322	0.34	−1.89	29	35

(2012) and Khorramdel and von Davier (2014), separate the item response into “pseudo items” such that the selection of extreme versus non-extreme categories only reflects a response style tendency. The latter approach is attractive from a measurement standpoint in that it allows for a clearer separation between, and thus better measurement of, the substantive and response style traits; however, it comes at the expense of assuming that the extremity of response is not reflective of the substantive trait, which may not be consistent with prior beliefs in adopting the Likert rating scale format. It is not our intent in this paper to explore differences in these alternative conceptualizations of ERS, although this represents a clear area for further methodological investigation.

The current approach also differs substantially from the approach considered by Buckley (2009), who focused on the paradox in relation to two of the PISA attitudinal subscales, the Enjoyment and Value subscales, in analyzing the PISA 2006 data. Following Baumgartner and Steenkamp (2001), Buckley (2009) used ad hoc measures of response style based on the counts of extreme responses in measuring response style tendencies. Similar to Greenleaf (1992), the indices in Buckley’s analysis were defined from a subset of items across the remaining attitudinal scales so as to yield indices that are more likely indicative of response style than of the substantive attitudinal traits. Further, the correction for response style used by Buckley (2009) entailed a linear regression of attitudinal scale scores on the response style indices (implying linear biasing effects of response style). Importantly, such a correction assumes that the biasing effects of response style are constant across levels of the substantive trait, an assumption that, as seen below, is sharply at odds with the current model based approach. Using this approach, Buckley (2009) found rather substantial changes in the country-level attitude/achievement relationship, such that a strong linear association became replaced by a substantially weaker nonlinear relationship. Buckley (2009) acknowledged the need for additional conceptualizations of response style in order to better understand this result.

As indicated above, the approach in this paper differs significantly from Buckley (2009) both in its measurement of ERS and in the nature of the bias correction applied. Specifically, the model in (1) assumes the substantive trait will demonstrate more influence over the selection of extreme response categories, similar to approaches such as Rost et al. (1997) and Moors (2003). As such, it accounts for the possibility that respondents and/or countries that are selecting a large portion of extreme responses may be doing so in large part because they have extreme levels on the substantive trait. Second, and perhaps more importantly, the current approach accounts for the likely result that the nature of ERS bias will be different depending on the level of the substantive trait. Intuitively, one might expect the bias to be negative when substantive trait levels are low (as ERS leads to selection of more extreme lower scores than otherwise expected) but positive when substantive trait levels are high (as ERS lead to selection of more extreme higher scores). The nature of such biasing effects are looked at in more detail later in the paper.

Methods

In this paper, the methods section is illustrated in terms of proposing a multilevel IRT (MMIRT) model for ERS, applying of the MMIRT model to PISA 2006, and detecting country level bias under the MMIRT.

A multilevel MIRT model for ERS

The multilevel structure of PISA data allows for examination of how a characteristic such as ERS can introduce bias in survey scores observed at both the student as well as country levels. In this context, we can extend the Bolt & Newton (2011) approach to address ERS effects within a multilevel multidimensional IRT (MMIRT) model. Assume country is indexed by *g*, and that *i(g)* denotes a student *i* nested within country *g*. For simplicity of notation, we will denote student by *i*, recognizing that each student belongs to one and only one country *g*. Then the model can be specified as

$$P\left(Y_{ij} = k|\theta_{is}, \theta_{i,ERS}\right) = h\left(a_{jk1}\theta_{is} + a_{jk2}\theta_{i,ERS} + c_{jk}\right) \tag{2}$$

where θ_{is} denotes an attitudinal trait corresponding to the one of seven attitudinal scales assessed by item *j*, which includes science enjoyment (ENJ), science value (VAL), environmental responsibility (ENV), usefulness for science career (USE), science in future A (FUTA), science in future B (FUTB), and science learning (LRN). Further, $h(\cdot)$ is a multinominal logistic function with the same representation as in Equation (1), and the slopes for the substantive and ERS traits are specified as $a_{j1}^{'} = [-3, \ -1, \ 1, \ 3]$ and $a_{j2}^{'} = [1, \ -1, \ -1, \ 1]$, respectively, for all items *j*, reflecting the four category scoring of the PISA attitudinal items. Each attitudinal item is therefore modeled with respect to a substantive trait underlying the scale to which it belongs (denoted as $\theta_s = \theta_{ENJ}$, θ_{VAL} , θ_{ENV} , θ_{USE} , θ_{FUTA} , θ_{FUTB} , or θ_{LRN}) and the extreme response style trait (denoted as θ_{ERS}). We assume θ_{ERS} remains constant for a given respondent across subscales. The advantage of modeling all attitudinal scales simultaneously when making this assumption is that it provides more reliable estimation of θ_{ERS}, and thus better control of ERS with respect to measurement of the substantive traits (see Bolt & Newton, 2011). As for the model in (2), the c_{jk} parameters are freely estimated subject to the constraint $\sum_k c_{jk} = 0$ for all *j*. The second level of the model, the country level, associates with each country a mean vector, μ_g and covariance matrix $\Sigma_g^{(1)}$, representing the mean and covariance matrix across the substantive traits and ERS. Each country mean vector is assumed to be an observation from a multivariate normal distribution, with a multivariate mean of 0 and covariance matrix $\Sigma^{(2)}$, representing the between-country covariance matrix among traits. For simplicity, we assume the within country covariance matrices, $\Sigma_g^{(1)}$, are constant across countries, i.e., $\Sigma_g^{(1)} = \Sigma^{(1)}$ for all g. Although the model permits the introduction of country-level variables as covariates, for all analyses in this paper, such variables are not included.

Applying the MMIRT model to examine the attitude-achievement paradox in PISA 2006

As the primary interest of the current paper is to address the between-country correlation between mean attitudes and mean achievement, we expand the model in the previous section to also include a student-level achievement variable. The PISA 2006 dataset provides student achievement scores in the form of plausible values. Given the complexity of the proposed model, we use the mean of the plausible values as an indicator of student achievement. This approach seemed reasonable in the current application as we are less concerned with the precision of individual country estimates than with the general direction of country-level correlations with attitudes. This simplifying approach was also used in Buckley (2009). However, procedures for analysis using plausible values have been presented (von Davier, Gonzalez & Mislevy, 2009), and we later apply a sensitivity check to examine the likely consequences of this decision, The resulting student achievement variable can then be introduced to the model as a single indicator observed variable, thus expanding each of μ_g, $\Sigma^{(1)}$, and $\Sigma^{(2)}$ to include an additional variable, observed once per student. Our primary interest focuses on those elements of $\Sigma^{(2)}$ that reflect the relationship between the country-level mean achievement, and country-level mean attitudes, as reflected by the seven different attitudinal scales.

The resulting model is still complex, which combined with the large size of the data matrix, make the resulting analyses very computationally demanding. Thus, we simplify the analysis in a couple of fundamental ways, neither of which would appear to substantially influence the findings of interest. First, rather than using the full dataset, we randomly sampled 200 students from each country to create a subsample of the data for analysis. To evaluate whether use of 200 students provided a sufficiently representative sample, we correlated the attitudinal scores and achievement scores for the subsample and compared them to those observed with the full sample. The statistics in Table 4 suggest that the subsample provides a good approximation to the full sample in reproducing approximately the same between-country correlation observed in the full data. Second, to address computational challenges related to the complexity of the model, we adopted a two-stage estimation procedure for the multilevel model. Specifically, in the first stage, we focus on estimating just the category intercepts associated with each item. Using the Latent Gold software (Vermunt & Magidson, 2005, 2008), we estimated the category intercepts for each item (a total of 41 science attitudes items) by fitting separate models to the individual attitudinal subscales. Each of these analyses entailed specification of a two-dimensional model (with one substantive trait, and one ERS trait) in which all items within the subscale tapped both dimensions, with each dimension having category slopes as specified for the model as in Equation 1. The resulting category intercepts, which effectively define the relative propensities toward each score category when both the substantive and ERS traits are at their respective means (in both case 0), were then used in the second stage.

In the second stage, we treated the item category intercept estimates as known and estimated person- and country-level parameters related to the latent traits. The use of fixed, as opposed to estimated, category intercepts was not anticipated to be of substantial consequence to the proposed question of interest, which concerned the direction and magnitude of the between- country correlations observed. This second stage was implemented using Markov chain Monte Carlo (MCMC) methods through the

Table 4 Correlations between science achievement and attitudes

	Correlation w/ACH (full sample)	Correlation w/ACH (subsample)
Enjoy (ENJ)	.73	.74
Value (VAL)	.66	.67
Environment (ENV)	.32	.30
Use (USE)	.52	.47
Future A (FUTA)	.80	.77
Future B (FUTB)	.76	.75
Learning (LRN)	.74	.73

software WinBUGS (Spiegelhalter, Thomas, & Best, 2004). The goal of this analysis was to obtain the within- and between-country covariance estimates, focusing in particular on those elements involving achievement. As noted above, for a given country g, the nine element trait vector for each student i is assumed multivariate normally distributed with mean μ_g and covariance matrix $\Sigma^{(1)}$. Thus, in a fully Bayesian estimation framework, student parameters are sampled as

$$\left(\theta_{i,ENJ}, \theta_{i,VAL}, \ldots, \theta_{i,LEA}, \theta_{i,ERS}, ACH_i\right) \sim MVN\left(\mu_g, \Sigma^{(1)}\right) \tag{3}$$

where $\theta_{i,ENJ}$, $\theta_{i,VAL}$, ... , $\theta_{i,LRN}$ are the student's trait levels on the attitudinal traits, $\theta_{i,ERS}$ is the student's level of ERS, and ACH_i is the student achievement score. In order to make results comparable across countries, we assume the same category intercept parameters apply for all countries. To draw inferences concerning the $\mu_g s$, we assigned a multivariate normal prior to the $\mu_g s$ and denoted the between-country covariance matrix as $\Sigma^{(2)}$, such that,

$$\mu_g \sim MVN\left(\mu, \Sigma^{(2)}\right). \tag{4}$$

We arbitrarily assign $\mu = (0,\ 0, \ldots,\ 0,\ A\bar{C}H)$ for the mean attitudinal latent traits and the mean achievement measure, where the overall grand mean achievement score for the subsample was 4.64. The within- and between-country latent variable covariance matrices are each assigned non-informative inverse Wishart priors, using a scale parameter of 10 and an identity matrix as the scale matrix, i.e.,

$$Inv\left(\Sigma^{(1)}\right) \sim W(\nu, V_\Sigma) \tag{5}$$

$$Inv\left(\Sigma^{(2)}\right) \sim W(\nu, V_\Sigma) \tag{6}$$

To fit the MMIRT model to the PISA subsample in WinBUGS, MCMC runs of 10000 iterations (500 iterations burn in) were used. Such chain lengths appeared adequate according to multiple criteria for evaluating convergence. We used the means of the posterior distributions from the MCMC simulations (expected a posteriori estimates) as estimates of all model parameters.

The accuracy of this general two-stage approach was investigated using simulation analyses as reported in Lu (2012). The simulation analyses were based on sample sizes corresponding to our reduced sample and assuming items with psychometric characteristics that were the same as those included in PISA 2006. Results suggested that the two-stage approach returned estimates of both the within- and between-country covariance matrices that closely resembled the generating matrices. Further details are provided in Lu (2012).

Detecting country level bias due to ERS in attitudinal scale scores

To better understand ERS-related bias on the estimation of country means for the attitudinal traits, we generalized an approach presented by Bolt and Johnson (2009). Relative to Bolt and Johnson, country level bias in total test scores is evaluated with respect to ERS trait distributions (as would correspond to a country) as opposed to individual levels of ERS (as would correspond to individuals). Using the seven-item Environment

(ENV) subscale as an example, the country level expected mean sum score based on the MMIRT item parameter estimates and country level trait distribution (i.e., μ_{ENV}, μ_{ERS}, σ^2_{ENV}, σ^2_{ERS}, $\sigma_{ENV,ERS}$) for a hypothetical country of interest, we define an expected mean score (EMS) for the country as:

$$\begin{aligned} &EMS\left(\mu_{ENV}, \mu_{ERS}, \sigma^2_{ENV}, \sigma^2_{ERS}, \sigma_{ENV,ERS}\right) \\ &= \int_{\theta_{ENV}} \int_{\theta_{ERS}} \sum\nolimits_{j=1}^{7} \sum\nolimits_{k=1}^{4} k * P\left(\mathbf{Y}_j = k|\theta_{ENV}, \theta_{ERS}\right) \; f(\theta_{ENV}, \theta_{ERS}) \; d\theta_{ERS} d\theta_{ENV} \end{aligned} \tag{7}$$

where $P(\mathbf{Y}_j = k|\theta_{ENV}, \theta_{ERS})$ is defined by the MMIRT model in Equation (2), with item category intercepts estimated from Stage 1 of the estimation procedure (shown in Table 5), and $f(\theta_{ENV}, \theta_{ERS})$ is a bivariate normal probability density function. Equation (7) can be estimated through discrete approximation of the bivariate integral. The result is an expected mean score based on both the fixed item characteristics as well as the relevant distributional characteristics of both the Environment subscale and ERS for a hypothetical country.

To quantify bias, we consider a reference country identical in all respects to the hypothetical country above, except now having $\mu_{ERS} = 0$. We apply the same procedure as above to determine an expected mean score on the Environment subscale. Then bias can be estimated as the difference between the two expected scores, i.e.,

$$BIAS = EMS\left(\mu_{ENV}, \mu_{ERS}, \sigma^2_{ENV}, \sigma^2_{ERS}, \sigma_{ENV,ERS}\right) - EMS\left(\mu_{ENV}, \mu_{ERS} = 0, \sigma^2_{ENV}, \sigma^2_{ERS}, \sigma_{ENV,ERS}\right) \tag{8}$$

Although there is naturally no "correct" mean level of ERS, using μ_{ERS} =0 provides a natural reference as 0 corresponds to the overall mean on ERS. Naturally, for a given subscale with fixed category intercepts, it becomes possible to display bias in relation to different levels of μ_{ERS} as a function of μ_{ENV} to better illustrate how between-country variability in response style will introduce bias in the mean scale score estimate of *ENV* for that country.

Finally, in order to evaluate the effects of ERS control on the estimated between-country correlations between achievement and attitudes, we apply the same multilevel model represented in Equations 2–6, but now excluding the ERS trait at both the individual and country levels. The resulting between-country correlations between mean attitudes and mean achievement represent the baseline correlations between these means when ignoring the potential effects of ERS. The same estimation procedure as for the full model was applied. The resulting achievement/attitude correlations provide a more

Table 5 Environment subscale item category intercept estimates

Item	Category			
	1	2	3	4
1	2.29	2.81	0.04	−5.15
2	0.82	1.93	0.84	−3.58
3	0.66	1.84	0.91	−3.41
4	1.45	2.26	0.53	−4.23
5	2.71	2.54	−0.19	−5.06
6	2.94	2.46	−0.31	−5.08
7	1.10	2.08	0.65	−3.83

meaningful reference against which to evaluate the effects of response style control as they are evaluated using a common latent trait metric to that used in the full MMIRT model.

Results and discussion

While various aspects of the results could be discussed in detail, we focus in particular on the estimated within- and between-country covariance matrices, as well as the estimated country level mean vectors observed from the multilevel IRT model. Under the MMIRT model, the results suggest that there is detectable variability in ERS across the PISA countries. Figure 3 displays a plot of the country-level μ_{ERS} estimates and their 95 % credible intervals by country, ordered from least ERS to most ERS. It is apparent from these estimates, for example, that Thailand (THA) displays on average the least amount of ERS (μ_{ERS}= − .69), while Tunisia (TUN) the most (μ_{ERS}=.55). Importantly, a large proportion of countries display non-overlapping credible intervals, suggesting distinctions across countries in the distributions of ERS. Such differences open the possibility that ERS may differentially bias mean scale scores with respect to the attitudinal scales. The between-country variance associated with ERS is estimated at .24 (95 % CI = .16, .34). The meaningfulness of the country-level μ_{ERS} estimates is further supported by the rather strong positive correlation (r = .672, p < .01) between the μ_{ERS} estimates and the country-level mean number of extreme responses (i.e., pseudo “e-items” following Bockenholt, 2012 and Khorramdel & von Davier, 2014). A less than perfect correlation is expected to the extent that the current model assumes extreme responses are in some cases caused by the level of the substantive trait.

As noted above, given these country-level μ_{ERS} estimates, combined with estimates of the item category intercepts, as well as other estimates of the distribution of the

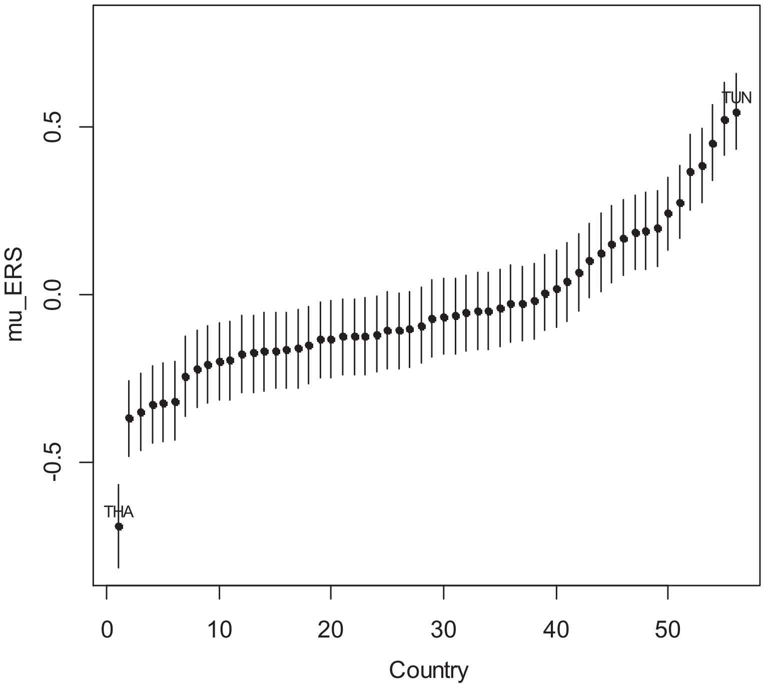

Fig. 3 Plot of mean ERS estimates, MMIRT analysis, PISA 2006 data

substantive and ERS traits for a respective country, we can apply Equation 7 to consider the potential for bias in the mean attitudinal scale score at the country level. Again taking the Environment (ENV) subscale as an example, based on Equation (7) and Equation (8), and the category intercept estimates for the Environment subscale shown in Table 5, Fig. 4 illustrates how ERS introduces bias into the sum score on the country level based on the category intercepts observed for the Environment subscale. Such curves are based also on estimates of the estimated within-country variance of ERS, the within-country variance of Environment, and the within-country covariance of Environment and ERS, all of which are assumed constant across countries, and in this case returned estimates of .47, .54, and .23, respectively.

In Fig. 4, the curves illustrate bias at levels of μ_{ERS}= 1, .5, 0, −.5, and −1, respectively. Figure 4 clearly shows that, for a fixed level of μ_{ERS}, the magnitude and direction of bias varies quite substantially as a function of μ_{ENV}. It should be noted that in Fig. 4 all the curves intersect approximately at a common location, where the bias is basically 0 for all levels of μ_{ERS}. In the current analysis, this common location is approximately μ_{ENV}=1; this is also the μ_{ENV} level where the average expected scores across items is approximately 2.5, the midpoint of the response scale. As μ_{ENV} moves in either direction away from 1, bias is introduced. As shown in Fig. 4, a high ERS mean introduces positive bias on the total score of the Environment subscale as μ_{ENV} moves above 1 and negative bias as μ_{ENV} moves below 1. For example, when μ_{ENV} is larger than 1, countries with high ERS are more likely to select 4 s than 3 s on the items; when μ_{ENV} is smaller

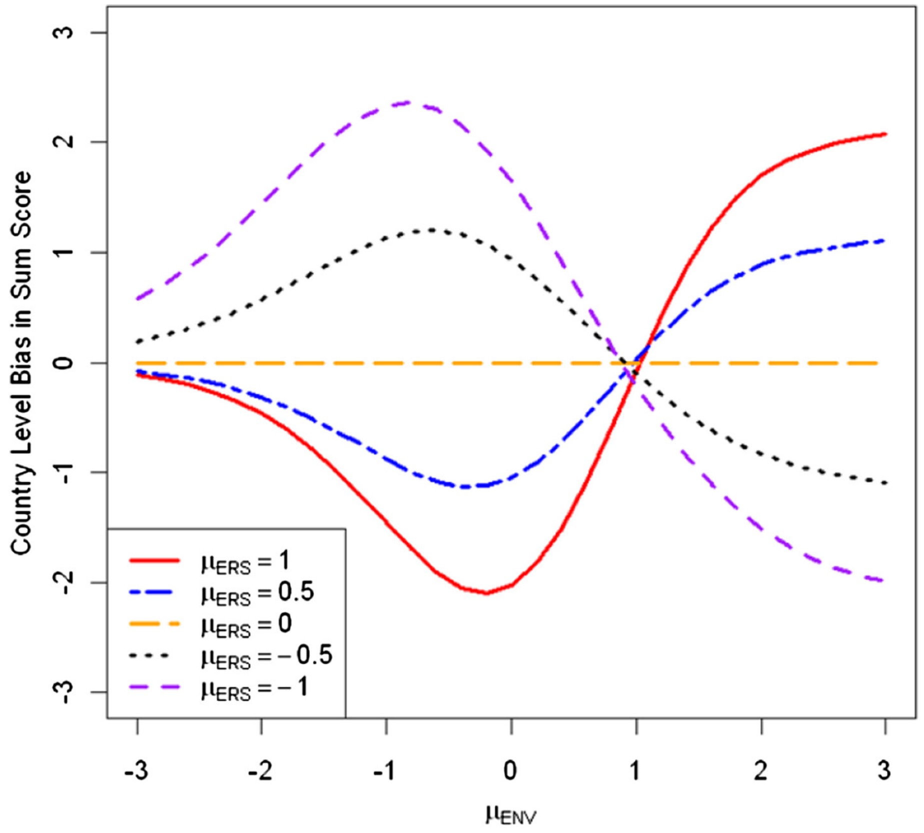

Fig. 4 Plot of country level bias in sum score as a function of mu environment

than 1, high ERS countries are more prone to select 1 s than 2 s on the items. Such curves play a useful role in understanding why controlling for ERS introduces larger changes in the mean estimates of the substantive traits for some countries but not others.

Figure 5 illustrates the implications of the bias correction across all 56 countries. In this plot, the mean Environment (ENV) scale score is shown as the x-axis and the bias-corrected score (i.e., the estimate of μ_{ENV} from the MMIRT analysis) as the y-axis. The countries circled are described for illustrative purposes. Notice for example, that Tunisia (TUN) and Thailand (THA), which show approximately the same mean ENV scale scores, appear quite different with respect to the μ_{ENV} estimates, reflecting the nature of the bias correction applied within the MMIRT model. Interestingly, for Thailand and Tunisia, the μ_{ENV} estimates were -.67 and .06, respectively, both of which according to Fig. 5 represent locations with significant potential for bias in the mean ENV scores due to ERS. In the case of Thailand, the bias should be positive, implying the mean ENV score reflects a more negative attitude than is actually the case, while for Tunisia, the bias should be negative, implying the scale mean reflects a more positive attitude than is actually the case. Many other countries, however, appear rather unaffected by the ERS control. The Netherlands, for example, which showed the highest mean scale score on the ENV scale, is rather moderate in terms of ERS (μ_{ERS} =−.17), and thus retains its relatively high ranking on the μ_{ENV} metric.

We have noted how Buckley (2009) used not only a different definition of response style than is considered in this paper, but also a different method of correcting for bias. At the country level, we find our estimates of μ_{ENJ}, μ_{VAL}, and μ_{ERS} to correlate at levels of .446, .679, and .756, respectively, with Buckley's (2009). Thus, while there is some consistency in how countries are identified with respect to ERS, there are substantial differences from Buckley's approach in terms of the bias correction.

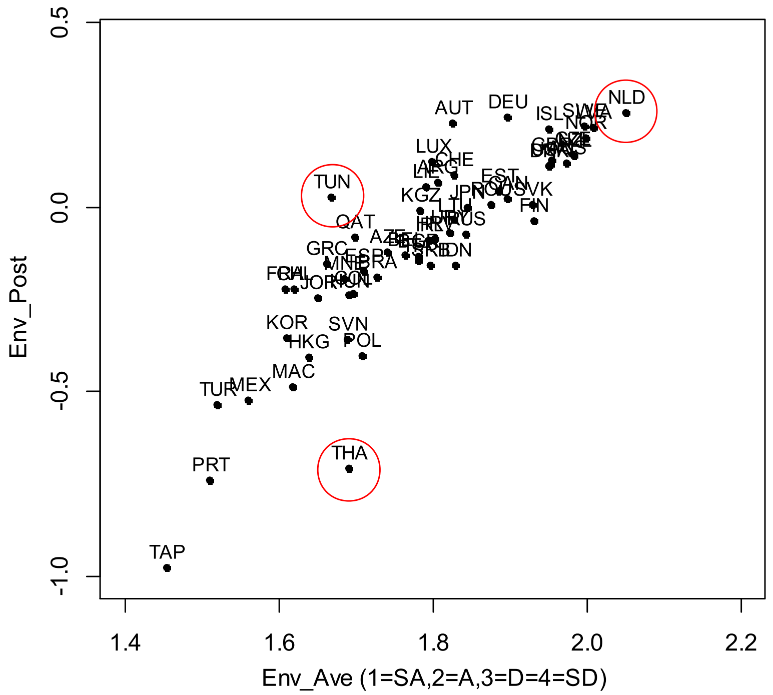

Fig. 5 Scatter plot of country level subscale mean by posterior mean on environment

Examining the estimated within- and between-country covariance matrices

While the above results make clear that there is a potential for meaningful bias correction due to ERS, the primary objective of the paper was to evaluate whether ERS can explain the attitude/achievement paradox. To this end, we are interested in estimates of the within-country ($\Sigma^{(1)}$) and particularly, the between-country ($\Sigma^{(2)}$) covariance matrices from the MMIRT analysis. Table 6 reports these estimates. The variance between countries with respect to the attitudinal traits would appear to vary somewhat across scales, with the largest variability occurring for the Future A (FUTA) scale, and the lowest for the Environment (ENV) scale; moreover the variances within country are also noticeably different across scales, with the same scales occupying the extremes. As expected, the attitudinal scales consistently positively covary with each other, both within and between countries. Moreover, ERS appears rather modestly correlated both with the attitudinal scales (both within- and between-country), but even more importantly, with the achievement variable, both within- and between- countries. In particular, the very weak association between ERS and achievement at the country level suggests it is unlikely to have much of an effect as a control variable in evaluating the attitude/achievement correlations. Finally, the last row/column of each of the within- and between-covariance matrices illustrate the remaining paradox, in that the attitudinal scales consistently show a remaining positive correlation with the achievement metric between countries, but a negative correlation within countries. To better evaluate whether control of ERS nevertheless had some effect on the correlations, Table 7 illustrates the estimated correlations between achievement and attitudes at the between-

Table 6 Estimated between-country and within-country covariance matrices

(a) Between-country covariance matrix

	ENJ	VAL	ENV	USE	FUTA	FUTB	LRN	ERS	ACH
ENJ	.42	.13	.06	.12	.20	.21	.15	.02	.17
VAL	.13	.29	.06	.08	.13	.14	.08	.04	.09
ENV	.06	.06	.25	.03	.04	.07	.01	.03	.02
USE	.11	.08	.03	.31	.09	.14	.10	.03	.07
FUTA	.20	.13	.04	.09	.52	.22	.21	-.00	.24
FUTB	.21	.14	.07	.14	.22	.45	.17	.02	.19
LRN	.15	.08	.01	.10	.21	.17	.42	-.01	.19
ERS	.02	.04	.03	.03	-.00	.02	-.01	.24	-.02
ACH	.17	.09	.02	.07	.24	.19	.19	-.02	.47

(b) Within-country covariance matrix

	ENJ	VAL	ENV	USE	FUTA	FUTB	LRN	ERS	ACH
ENJ	1.55	.64	.34	.59	1.10	.94	.91	.14	-.29
VAL	.64	.60	.31	.44	.55	.58	.45	.21	-.16
ENV	.34	.31	.47	.31	.25	.31	.21	.23	-.16
USE	.59	.44	.31	1.10	.55	.66	.48	.22	-.08
FUTA	1.10	.55	.47	.55	1.67	1.16	.88	.10	-.18
FUTB	.94	.58	.31	.66	1.16	1.54	.84	.16	-.16
LRN	.91	.45	.21	.48	.88	.84	1.47	.07	-.25
ERS	.14	.21	.23	.22	.10	.16	.07	.54	-.01
ACH	-.29	-.16	-.16	-.08	-.18	-.16	-.25	-.01	.71

Table 7 Between-country correlations between attitudes and science achievement, with and without ERS controlled

	w/out ERS control	w/ERS control
Corr(ENJ,ACH)	.44	.39
Corr(VAL,ACH)	.26	.25
Corr(ENV,ACH)	.08	.05
Corr(USE,ACH)	.22	.19
Corr(FUTA,ACH)	.45	.48
Corr(FUTB,ACH)	.41	.42
Corr(LRN,ACH)	.45	.47
Corr(ERS,ACH)		-.07

country level when the ERS trait is included in the model versus when it is not. As seen from the correlation estimates in both columns, the estimates change only minimally. The largest decrease appears to be observed for the Enjoyment subscale, although that change is small, and all of the correlations are in fact positive. As implied above, one clear cause of these minimal effects is the very weak correlation between ERS and achievement, which, while in the anticipated direction, is clearly too small to explain the moderate correlations between attitudes and achievement. In addition, it is apparent from Table 6, that the ERS variable displays more variance within (estimate = .54) than between (estimate = .24) countries, yielding an intraclass correlation estimate of .31. Consequently, an expectation of dramatic changes due to country-level bias correction would seem unlikely given the relatively lower amount of variability in ERS that exists between countries.

Finally, as a way of evaluating the potential consequences of using the average of the plausible values for achievement (in place of the plausible values themselves) we conducted an analysis using the 9-dimensional model including ERS in which the five plausible values were used to account for the uncertainty of the examinee achievement estimates. We observed very similar correlations between achievement and the attitudinal measures in the new analysis, in all cases within .03 of those reported in Table 7. The country-level mean estimates for achievement for the new analysis also correlated .98 with those of the earlier analysis, and the corresponding between- and within-country variances for achievement were unchanged (.47, .71, respectively). So it would appear that the use of the average achievement measure was not consequential with respect to the primary findings of the paper.

Conclusions

Our application of a two-level multidimensional IRT model toward investigating the attitude-achievement paradox in PISA 2006 suggests cross-cultural differences in extreme response style (ERS) are not a likely cause of the paradox. This conclusion is largely based on the observation of nearly identical between-country correlations across attitudes and achievement when controlling for, versus not controlling for, country differences in ERS. Application of the model also provides some indication as to why the between-country correlations are largely unchanged. First, although country level differences in ERS are detectable, they are relatively small compared to within-country variability in ERS. Our intraclass correlation estimate related to ERS is approximately .31, suggesting that even in countries that are extreme in either direction on ERS, there remains a fair amount of within-country heterogeneity. Second, the correlation between achievement and ERS at the country level, while in the expected direction (i.e., higher mean achievement is associated with less ERS), is rather weak (−.07). Consequently, controlling for country effects with respect to ERS is unlikely to result in meaningful effects on correlations with achievement.

Of course, such conclusions may be affected by the nature of the method being used to study and control for the effects of ERS. We find the use of the proposed model, however, to be attractive relative to other approaches that have been used to look at the paradox. Unlike earlier studies (e.g., Buckley, 2009), for example, the proposed model accounts for the biasing effects of ERS in a nonlinear (as opposed to linear fashion), as would seem intuitively to make sense. If ERS is viewed as an effect unrelated to the intended to be measured substantive trait, its biasing effects should in fact be nonlinear, with more positive bias (in terms of the scale score) occurring at higher trait levels, and negative bias at lower trait levels. The example provided in Fig. 4 shows that such nonlinear effects are present at the country level in the same way as they emerge at the individual level (Bolt & Johnson, 2009).

As noted earlier, an alternative approach to modeling ERS presented by Bockenholt (2012) and also applied by Khorramdel & von Davier (2014) provides a competing method that also differs in its definition of ERS and could also be generalized to a multilevel framework. While we did not develop such a generalization in this paper, we can get a sense as to whether this alternative approach will likely yield different results by aggregating the pseudo d-items to the country level and inspecting the correlation with mean country-level achievement. Such correlations likewise were approximately equal to the original uncorrected correlations, ranging from .270 for the Environment subscale to .810 for the Future A subscale, in all cases statistically significant.

While these overall findings seem to lend additional credibility to alternative explanations for the attitude-achievement paradox (e.g., Marsh et al., 2008), there of course also remains a possibility that alternative forms of response style (e.g., ARS, DRS) to those examined in this study still provide an explanation. In a cross-cultural study of response style using related methods to those studied in this paper (Bolt, Lu & Kim, 2014), it was seen that different countries can make differential use rating scales in ways that do not conform to traditionally studied response style types. Thus, there would seem to be value in additional methodological study of cross-cultural differences in response style as possible sources of the attitude-achievement paradox. Along these lines, there may also be value in alternative design considerations for better measurement of response style tendencies. For example, the use of anchoring vignettes (e.g., King, Murray, Salomon & Tandon, 2004) may provide greater value in identifying and controlling for response styles of different kinds and with greater precision than can be achieved using only self-report ratings. Another alternative is the simultaneous administration of other self-report scales involving heterogeneous content, making ad hoc indices of response style less susceptible to influence of the substantive trait. Including reverse-worded items would likely make it easier to detect response styles such as ARS or DRS. Our interest in ERS, specifically, was motivated by several considerations, including (1) the frequent observation in the literature that ERS varies cross-culturally, (2) the tendency for ERS to be associated with educational achievement, and (3) the ability to measure ERS with greater precision than alternative response styles (e.g., ARS, DRS) using the currently available self-report data for PISA. The use of alternative data collection designs would make it easier to measure and thus control for the influence of response style types such as acquiescent response style (ARS), which has also been observed to vary cross-culturally (Johnson et al., 2005).

Competing interests

We have read and understood Large-scale Assessments in Education policy on declaration of interests and declare that we have no competing interests.

Authors' contributions

YL and DB of this research paper have directly participated in the planning, execution, and analysis of this study. All authors read and approved the final manuscript.

Author details

[1]ACT, Inc., 500 ACT Drive, 52243 Iowa City, IA, USA. [2]Department of Educational Psychology, University of Wisconsin-Madison, 1025 West Johnson Street, 3706 Madison, WI, USA.

Received: 23 February 2015 Accepted: 20 July 2015
Published online: 18 August 2015

References

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

Bock, RD. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological methods, 17*(4), 665–678.

Bolt, DM, & Johnson, TR. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement, 33*, 335–352.

Bolt, DM, & Newton, J. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 71*, 814–833.

Bolt, DM, Lu, Y, & Kim, J-S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods, 19*, 528–541.

Buckley, J (2009). *Cross-national response styles in international educational assessment: Evidence from PISA 2006*. NCES Conference on the Program for International Student Assessment: What we can learn from PISA (Downloaded from https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf on May 16, 2010).

Bybee, R, & McCrae, B. (2007). Scientific literacy and student attitudes: Perspectives from PISA 2006 science. *International Journal of Science Education, 33*, 7–26.

Clarke, I. (2000). Extreme response style in cross-cultural research. *Journal of Social Behavior and Personality, 15*, 137–152.

Greenleaf, EA. (1992). Measuring extreme response style. *Public Opinion Quarterly, 56*, 328–361.

Johnson, TR, & Bolt, DM. (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*(1), 92–114.

Johnson, T, Kulesa, P, Cho, YI, & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*, 264–277.

Khorramdel, L, & von Davier, M. (2014). Measuring response styles across the Big Five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research, 49*(2), 161–177.

King, G, Murray, CJL, Salomon, JA, & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of survey research. *American Political Science Review, 98*, 191–207.

Loveless, T. (2006). *The 2006 Brown Center report on American education: How well are American students learning?* Washington, D. C.: Brookings Institution.

Lu, Y. (2012). *A multilevel multidimensional item response theory model to address the role of response style on measurement of attitudes in PISA 2006*. University of Wisconsin, Madison: Doctoral dissertation, 164 pages.

Marsh, HW, Seaton, M, Trautwein, U, Ludtke, O, Hau, K-T, O'Mara, AJ, & Craven, RG. (2008). The big fish little pond effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review, 20*, 319–350.

Meisenberg, G, & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences, 44*, 1539–1550.

Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach. Socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined. *Quality and Quantity, 37*(3), 277–302.

Rost, J, Carstensen, C, & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J Rost & R Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324–332). Munster: Germany: Waxmann.

Spiegelhalter, D, Thomas, A, & Best, N. (2004). *WinBUGS version 1.4*. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health [Computer program].

Van de Gaer, E, & Adams, R. (2010). *The modeling of response style bias: An answer to the attitude-achievement paradox?* Denver, CO: Paper presented at the annual meeting of the American Educational Research Association. http://www.acer.edu.au/files/vandegaer_paper_aera2010.pdf

Van Herk, H, Poortinga, YH, & Verhallen, TMM. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35*, 346–360.

Vermunt, JK, & Magidson, J. (2005). *Technical Guide for Latent Gold 4.0: Basic and Advanced*. Belmont, MA: Statistical Innovations, Inc.

Vermunt, JK, & Magidson, J. (2008). *LG-Syntax User's Guide: Manual for Latent Gold 4.5 Syntax Module*. Belmont, MA: Statistical Innovations, Inc.

von Davier, M, Gonzalez, E, & Mislevy, RJ. (2009). What are plausible values and why are they useful? In M Von Davier & D Hastedt (Eds.), *IERI Monograph Series: Issues and methodologies in large scale assessments* (Vol. 2, pp. 9–36). Hamburg, Germany: IERI Institute.

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

Wetzel, E, Carstensen, CH, & Böhnke, JR. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality, 47*(2), 178–189.