Service d'analyse des Systèmes et des
Pratiques d'enseignement (aSPe)
Université de Liège

Les Cahiers des Sciences de l'Éducation

N° 39

Understanding and Addressing Common-Method Bias in International Large-Scale Assessments: The Example of Reading Opportunities-To-Learn in PISA 2009

Svenja Vieluf
German Institute for International Educational Research (DIPF)

Christian Monseur, Ariane Baye, Dominique Lafontaine
University of Liège

2019

ISSN 2030 9643

Table of contents

ABSTRACT **1**

INTRODUCTION **2**

1. VALIDITY IN CROSS-NATIONAL COMPARISONS **3**

2. CROSS-NATIONAL DIFFERENCES IN RESPONSE STYLES **4**

3. HOW TO ADDRESS BIAS OBSERVED IN CROSS-NATIONAL STUDIES **7**

4. AIMS AND RESEARCH QUESTIONS **10**

5. METHOD **12**

5.1. Database and sample 12
5.2. Measures 12
5.3. Analyses 14

6. RESULTS **19**

6.1. Improving the predictive validity of the OTL factors 19
6.2. Understanding the common method dimension 22

7. DISCUSSION **26**

7.1. Limitations and perspectives 32
7.2. Conclusions 33

REFERENCES **34**

Appendix A 41

Abstract

Response style bias linked to Likert scales jeopardizes the validity and cross-cultural equivalence of constructs in comparative large-scale assessments. In our study, within-item multidimensional CFA and IRT models were implemented in order to disentangle common-method bias from the target opportunity-to-learn constructs in PISA 2009. This approach revealed promising results: evidence was found that the within-item multidimensional models improved the validity of the measurement of opportunity-to- learn (OTL) in the PISA context. With this model, stronger positive correlations between target OTL and reading achievement were observed at the student- and school-levels, whereas the common method dimension on which all OTL items loaded had robust negative correlations with reading achievement on both levels in all of the countries except Finland. Most of the variation in the common method dimension appeared at the student level, but some variation was also observed at the school and country levels. Country-level variation in the common method dimension appears to mainly reflect country differences in the tendency to acquiescence; country-level correlations with an acquiescence index computed on broader sets of PISA items were strong while those with an extreme response style index were rather weak. Further, negative correlations with socioeconomic status (SES) and reading ability were observed at the country level, showing that countries scoring high on the common method dimension are the less affluent and the lowest performing ones in reading.

Keywords: Response-style, common method bias, cross-cultural differences, PISA, within-item IRT modelling

Introduction

Since the first wave of the Programme for International Student Assessment (PISA) in the year 2000, international large scale assessments (LSAs) are playing a prominent role for educational monitoring and their policy impact has been growing. Especially considering this impact, it is crucial that not only student achievement, but also the non-cognitive outcomes and contextual variables that could account for differences in achievement between education systems, schools and students are measured in a valid way cross-nationally. Numerous variables related to practices, attitudes, beliefs or conceptions of students, teachers or principals have been assessed with self-report measures in international LSAs. In most cases, Likert-type scales were used and respondents were asked to report their level of agreement or rate the frequency of their activities or behaviors. Meanwhile, it is a well-established fact that those Likert-scales are prone to a number of methodological biases (e.g., Buckley, 2009; He & Van de Vijver, 2015). Individual respondents might be more or less prone to response styles such as acquiescence, extreme responding or social desirability. Additionally, systematic response style differences have been found between social groups and between national cultures (He & Van de Vijver, 2013; Van de Vijver & He, 2014). Whatever their origin, these differences limit measurement validity. Therefore, it is crucial to address response-style biases, either during the instrument development phase, or a posteriori, analyzing the data in a proper way to limit the biasing impact of response styles as much as possible.

The main aim of the present study was to find an effective a posteriori modeling approach in order to address response style bias observed in international large-scale assessments and consequently, to improve the validity of measures of educational practices, beliefs, attitudes, and concepts. Drawing on the PISA 2009 opportunity to learn (OTL) variables as an example for the kind of issues typically raised by Likert-type scales, we explored one possible solution to

address response style bias: within-item multidimensional IRT-models separating content dimensions from one common method dimension. Firstly, we tested whether this approach improves the predictive validity of the OTL dimensions. Secondly, we examined to what extent (in the PISA context) the common method dimension is a country, a group or an individual phenomenon, to what degree cross-national variation in the common method dimension reflects national differences in response styles, and whether cultural or country patterns can be found in the magnitude of common method bias.

1. Validity in Cross-National Comparisons

Valid cross-national comparisons do not only require instruments meeting the traditional standards of construct, content, and criterion validity; the issues of cross-national bias and equivalence also need to be addressed (He & van de Vijver, 2012; Van de Vijver & Poortinga, 1982; van de Vijver & Tanzer, 2004).

Bias refers to a confounding of the measurement in a way that score differences do not correspond to differences in the underlying trait (e.g. Poortinga, 1989). Three forms of bias are typically distinguished: construct, item, and method bias. Construct bias results from cross-national differences in the definition of a construct. An item is biased when members of different groups who have the same position on the construct do not have the same expected score on the item. Method bias refers to differences in sampling, structural features of the instrument, or administration processes that systematically influence the resulting scores of different groups (He & van de Vijver, 2012; Van de Vijver & Poortinga, 1982; van de Vijver & Tanzer, 2004).

Cross-national equivalence represents the flipside of the coin. It is achieved when a score has a similar meaning internationally, i.e. in the absence of bias. Equivalence is not an absolute term; different levels can be distinguished. The

first and most theoretical level is that of conceptual equivalence, which implies that a set of indicators is conceptually adequate for operationalizing a construct in different cultures. Evidence for the second level of functional equivalence is provided when structural equation modelling (SEM) shows a cross-nationally similar factor structure and equivalent relationships with other relevant constructs. However, these two levels are not sufficient for quantitative comparisons to be valid. Comparisons of intracultural score differences and correlations additionally require the constructs to be measured with the same metric (metric invariance) and comparisons of latent means require the same metric and equivalent scale origins (scalar invariance; e.g., Cheung & Rensvold, 1999; Davidov, 2008; Meredith, 1993; Steenkamp & Baumgartner, 1998). Cross-national equivalence and absence of all three forms of bias are a requirement of valid quantitative cross-national comparisons.

2. Cross-national differences in response styles

One form of method bias that often jeopardizes the validity of international surveys is characterized by cross-national differences in response styles. Response styles can be defined as "a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content" (Paulhus, 1991, p 17). The four main response styles are: (1) acquiescence response style (ARS) - a tendency to "endorse items irrespective of the item contents" (Van de Vijver & He, 2014, p. 6); (2) extreme response style (ERS) - the tendency to use the endpoints (or midpoints) of a response scale; (3) non contingent responding - erratic, random answers[1]; and (4) social desirability - the tendency to endorse the most normative or conforming

[1] ARS, ERS and non-contingent responding can result from what Krosnick (1991) described as a "satisficing behaviour", which has the aim of minimizing cognitive efforts by answering in a uniform way to different items belonging to one scale or even to all items in a questionnaire.

answers.

One important question is whether each response style is unique or whether the main response styles share some aspects, i.e., whether individual characteristics influence their response behavior in several different ways at the same time. For instance, He and Van de Vijver (2013) and Van de Vijver and He (2015), using data from a national study in the Netherlands and from OECD's Teaching and Learning International Survey (TALIS), found "support for a general response style factor" and showed that "the different response styles have both shared and unique aspects" (p. 799). ERS and SDR loaded positively and ARS and MRS negatively on this factor, which He and Van de Vijver interpreted as a general preference for amplification versus moderation in communication.

Another relevant aspect in the case of international comparative studies is the question to what extent response styles are individual phenomena or rather found at country level. A large number of studies suggest variation at both the individual (mostly according to personality traits) and the country level (for a synthesis, see Yang, Harkness, Chin, & Villar, 2010). A recent study using multilevel models estimated how much of the variation of one specific response style or of a general response style was explained by the country or by individuals within a country. Lu and Bolt (2015), using seven attitudinal scales of PISA 2006, found an intraclass correlation of 0.31 for the ERS scale, meaning that 31% of the variance was found at the country level and 69% at the individual level. They concluded that, although country-level differences in ERS were detectable, they were relatively small compared to within-country variability in ERS.

Country-level differences appear to reflect cultural traditions of larger regions and are further linked to other national characteristics. For example, ARS seems to be higher in Latin American countries as compared to the USA (e.g. Harzing,

2006; Ross & Mirowsky, 1984). Some studies also found ARS and ERS to be more prevalent in the Mediterranean than in North-western Europe (Harzing, 2006; van Herk, Poortinga, & Verhallen, 2004) and ERS is higher in North America than in East Asian countries (Clarke, 2000; Chen, Lee, & Stevenson, 1995; Heine et al., 2001; Lee & Green, 1991; Takahashi *et al*., 2002). High scores on a common method dimension were further observed in Tunisia, Brazil and Mexico whereas low scores were observed in Korea and Japan (Lie & Turmo, 2005). Moreover, social desirability and a common method dimension were shown to be negatively correlated with achievement at the country-level (Lie & Turmo, 2015; van De Gaer & Adams, 2010; van de Vijver & He, 2014) and with affluence (van de Vijver & He, 2014).

The afore-mentioned individual-, group-, and country-level differences in response styles can bias not only mean score comparisons but also analyses of relationships between scales (e.g. Baumgartner & Steenkamp, 2001). From a content point of view, response styles may be considered more as *nuisance* or as *impression management* (van de Vijver & He, 2014); however, from a measurement point of view, they obviously add noise to the measurement as they bring an additional dimension to the target construct. Response styles are also discussed as one reason for a puzzling observation that has been made in several cross-national studies, i.e. the lack of correspondence between individual- and aggregate country-level correlations of attitudinal constructs with achievement. For example, associations of student motivation, interest, and self-concept with student achievement are positive at the student-level, but negative at the aggregate country level. This so-called *attitude-achievement paradox* has been demonstrated repeatedly in cross-national studies such as PISA and TIMSS and across different subjects, grades, and cohorts (Kyllonen & Bertling, 2013; Lie & Turmo, 2005; Shen & Tam, 2008; van de Gaer, Grisay, Schulz, & Gebhardt, 2012; Lu & Bolt, 2015).

Individual- and group-level correlations describe different phenomena and cannot be used as valid substitutes for each other (Richards, Gottfredson, & Gottfredson, 1990-1991; Robinson, 1950). Hence, there is no theoretical reason why the individual-level and aggregate correlations between two variables should point in the same direction. Nevertheless, the attitude-achievement paradox casts doubt on the validity of cross-national comparisons of student attitudes and motivation, because there is evidence that cross-national differences in scores for scales measuring education-related norms or values do not only represent the constructs of interest. Rather, score differences between countries can also be explained with reference group effects on the one hand (Heine, Lehman, Peng, & Greenholtz, 2002; van de Gaer, Grisay, Schulz, & Gebhardt, 2012), and with cultural differences in self-expression norms (e.g., Harzing, 2006; Kobayashi & Greenwald, 2003; Kurman, 2003) and in response styles (e.g. Buckley, 2009; Peng, Nisbett, & Wong, 1997; Vieluf, Kunter & van de Vijver, 2013) on the other hand. These differences may be interesting from a research point of view, but they are not relevant for educational monitoring

3. How to address bias observed in cross-national studies

When international data are found to be biased because of cross-national differences in response styles, it is – to a certain extent – possible to correct for the bias statistically. This is often done by partialling out measures of response styles according to a number of different approaches (Podsakoff, MacKenzie, Lee, and Podsakoff, 2003).

One approach consists of partialling out an unrelated "marker variable" which is theoretically unrelated to the constructs of interest included in the study, so that correlations observed between this variable and constructs of interest can be assumed to be due to common method bias (Lindell & Brandt, 2000; Lindell & Whitney, 2001). Lindell and Whitney (2001) used the example of predicting

member participation with organizational climate and using marital satisfaction as a "marker variable", which was identified as theoretically unrelated to member participation and then used for adjusting the correlations.

A second approach is determined by partialling out direct measures of response styles. To this end, social desirability is often measured with questionnaire scales, such as the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960) or the Balanced Inventory of Desirable Responding (Paulhus, 1991). Simple indicators of ARS and ERS can be computed by adding the number of agreeing respective extreme responses (e.g., Paulhus, 1991). These indicators of social desirability, ARS and ERS can then be used to "peel off" scales of response styles.

According to a third approach, IRT and CFA techniques are used to correct for either specific response styles or for common method bias. For example, to identify and correct for ERS, Rost, Carstensen and von Davier (1997) suggested a mixed Rasch model approach, Moors (2003) a latent class factor approach, Lu and Bolt (2015) a multilevel multidimensional IRT (MMIRT) approach, and Von Davier and Khorramdel (2013) an approach using binary pseudo items in multidimensional IRT bifactor models. A bidimensional item-factor analysis model can be used for assessing ARS in a balanced set of binary items (Ferrando & Condon, 2009). To examine influences of different types of response styles on response behaviors, van Rosmalen, van Herk, and Groenen (2010) suggested a latent-class bilinear multinomial logit model. Using PISA 2003 data, Lie and Turmo (2005) applied a principal component analysis on the country mean estimates to 11 attitudinal scales (Likert scales) referring to different constructs such as learning strategies, motivation in mathematics, self-concept in mathematics and one school factor construct: they found a general factor explaining 66% of the variance. The authors labelled this factor "superconstruct" and argued that it may be interpreted as a general response

tendency.[2] Common to these techniques is that they try to estimate and distinguish a factor or latent trait that measures substantive content (i.e. the trait or construct to be measured) and a factor or dimension that measures common method bias or specific response styles. The advantages, disadvantages and limitations of the different statistical remedies are discussed in depth in the critical review by Podsakoff, MacKenzie, Lee, and Podsakoff (2003).

Several studies have examined whether these techniques improve the predictive validity of the substantive constructs. Results are mixed. For instance, Lu and Bolt (2015) found that adjusting scores in science related attitudes for ERS only led to small changes in the between-country correlations of these attitudes with science achievement and did not solve the attitudes-achievement paradox. Similarly, van de Vijver and He (2014) observed negligible effects of corrections for social desirability and a general response style factor on effect sizes of cross-cultural differences and rank orders of countries with regard to different constructs included in the TALIS. In contrast, Lie and Turmo (2005) adjusted country mean estimates on attitudinal scales by subtracting the country score on the "superconstruct" they had modelled. The correlations at the country level between achievement and attitudinal variables became much less negative at the country level, and even positive for some constructs. For instance, the correlation for self-concept in mathematics shifted from -0.26 to 0.74. In other words, this quite straightforward adjustment partially ruled out the attitude-achievement paradox. Hence, partialling out response styles does not appear to have large effects, but it might help to improve the validity of scales scores and to solve the attitude-achievement paradox whereby, however, the use of a common method dimension approach appears to be

[2] When multiple constructs are measured with multiple methods, it is also possible to partial out multiple method dimensions (e.g., Podsakoff et al.2003), but this approach is not discussed in the present paper because OTL in reading was measured using only one method in PISA 2006.

more successful than partialling out ERS only.

4. Aims and research questions

The present paper has two aims:

1) The first aim is to address response style bias observed in the PISA 2009 reading OTL and improve the predictive validity of the target constructs. We therefore compare a principal component analysis model and a within-item multidimensional IRT model with regard to their predictive validity.

2) The second aim of the study is to gain a better understanding of what the common method dimension in this modelling approach captures. To this end, firstly we analyzed to what extent it varies at the country, school and individual levels. Secondly we examined whether cross-national variation in the common method dimensions can be explained with different response styles. Thirdly we explored whether typical country patterns can be found in the common method dimension.

The following hypotheses were formulated regarding the first goal (methodological aspects):

1) In PISA, due to the use of Likert-type scales for measuring reading OTL, weak correlations between reading OTL dimensions and achievement will be observed at the individual student level when using principal component analysis (PCA) to model the reading OTL constructs (hypothesis 1a). The attitude-achievement paradox will be observed with weakly positive correlations at the student level, but with negative correlations at the system-level (hypothesis 1b).

2) Using within-item multidimensional IRT modeling to disentangle common

method bias and the OTL target constructs will improve the predictive validity of the OTL measure. Consistent and positive robust links between OTL and reading proficiency are expected as evidence of an improved predictive validity (hypothesis 2a). The attitude-achievement paradox will be fixed (hypothesis 2b).

Regarding the second goal (response style and patterns), we hypothesize that:

1) On the basis of the studies of Lu and Bolt (2015), we expect a small amount of variance of the common method dimension to occur between countries; most of the variance will be observed between individuals within a country, and the proportion of variance at the school level will vary from one country to another (hypothesis 3);

2) Cross-national variation in the common method dimension reflects national differences in response styles; accordingly we expect to find a moderate to strong country-level correlation with indices of response-style such as ARS or ERS (hypothesis 4);

3) On the basis of the literature on cross-cultural differences in response-styles, the following patterns are expected to be found: South American countries and South-East European countries will score higher on the common method dimension, while Asian countries and Scandinavian countries will score lower (hypothesis 5a). More broadly speaking, low-performing and less affluent countries will score higher on the common-method dimension (hypothesis 5b).

5. Method

5.1. Database and sample

The present study is a reanalysis of PISA 2009 Data (2009). PISA 2009 assessed reading, mathematics and science achievement in 34 OECD and 21 partner countries. Representative student samples were drawn in each participating country according to a two-stage design: First, a minimum of 150 schools were selected with a probability proportional to their size. Second, simple random samples of 15-year old students per school were selected across classes, tracks and grades, whereby the target cluster size was 35 students per school, and the minimum recommended total sample size was 4,500 students per country (OECD, 2009). For the present study only the 34 OECD countries were considered with a total sample of 284,806 students.

5.2. Measures

The variables used for the present study are reading achievement and reading OTL.

Reading achievement. Reading achievement in PISA 2009 was measured as part of a two-hour testing session. Several types of texts were used, i.e. continuous (narrative, informative, argumentative texts) and non-continuous ones (containing both texts and maps, graphs, figures, charts). Moreover, three different aspects of reading were measured: retrieve information, interpret text, evaluate and reflect on the content and form of the text. Students' level of reading proficiency was estimated using IRT analysis (five plausible values). Results are reported on a combined scale, three by-aspect scales and two by-type of text subscales (continuous vs. non-continuous).

OTL in reading. In PISA 2009, 17 OTL items were used focusing on the reading materials and activities to which students were exposed in their classes[3] or during their homework. According to Shanahan and Shanahan's typology (2008), the PISA OTL can be considered as mainly capturing "intermediate literacy" skills, namely "literacy skills common to many tasks, including generic comprehension strategies" (p. 44) and some "disciplinary literacy" skills, more specialized in literature. Eight items asked the students how often during the last month different types of texts had been used in classes, i.e. *fiction, poetry, texts that include diagrams or maps*, *texts that include tables or graphs*, *newspaper reports and magazine articles*, *instructions or manuals telling you how to make or do something*, *information about writers or books, advertising material (e.g., advertisements in magazines, posters).* Nine items asked how often students had been required to perform different types of tasks when reading these texts: *explain the cause of events in a text, explain the way characters behave in a text*, *explain the purpose of a text*, *learn about the life of the writer, find information from a graph, diagram or table*, *describe the way the information in a table or graph is organized, explain the connection between different parts of a text (e.g., between a written part and a graph), memorize a text by heart*, *learn about the place of a text in the history of literature*. For the present study, 12 out of the 17 original items were kept.[4]

[3] The question was asked in general ("in your classes"), no reference to specific or disciplinary classes was made.

[4] The items removed from the analyses are: Reading information texts about writers or books, reading poetry, learning about the life of the writer, memorizing a text by heart, learning about the place of a text in the history of literature. Five out of these items target processes or content related to the "disciplinary literacy" (literature). The emphasis on literature in language of instruction classes varies largely from country to country. Consequently, such items were positively correlated with achievement in some countries, especially in some Asian countries, and negatively correlated in other countries. These items appeared to be a source of instability in the multi-group confirmatory factor analysis.

Background variables. The students' socioeconomic background was measured via the PISA economic, social and cultural status index (ESCS). This index encompassed several components: home possessions (including cultural and educational resources), number of books at home, highest parental occupation and highest parental education expressed in total years of schooling see OECD, 2012, p. 312).

5.3. Analyses

To answer the first research question, firstly we implemented a principal component analysis (PCA), secondly we used bifactor confirmatory factor analysis (CFA) and examined cross-cultural invariance of the bifactor CFA model, and thirdly we implemented a within-item tridimensional IRT model. Consequently, we compared the correlations of the PCA factors (that were not corrected for method bias) and the IRT dimensions (that were corrected for method bias) with student achievement. To answer the second research question, firstly we decomposed the variance of the common method factor, secondly we analyzed correlations of this factor with other response style indicators, and thirdly we examined country patterns for this factor.

The list of analyses is detailed hereafter, together with preliminary technical results not related to research questions and hypotheses.

PCA. PCA (without and with varimax rotation) was performed on the 11 items of the OTL scales with each country contributing equally. The results showed that before rotation, all items loaded on a first factor, confirming the existence of one general factor. This factor explained 33% of the variance. After Varimax rotation, two factors of particular interest were extracted, one *Non-continuous* factor and one *Fiction* factor.

Bifactor CFA and cross-national invariance. Within-item dimensionality was first modelled by using a bifactor CFA model, allowing all items to load on one of the two substantive OTL factors (*Fiction* and *Non-continuous*, with the same items loading on each of these factors as in the PCA) and, at the same time, also on a common method factor. The latter factor was meant to capture common method bias such as response styles (ARS, ERS, disacquiescence, intermediate response style, social desirability or satisficing behavior). This model had a good fit (CFI = 0.97; TLI = 0.96; RMSEA = 0.06; WRMR = 13.45).

Next, we tested three levels of cross-national invariance, i.e. configural, metric, and scalar invariance (e.g. Cheung & Rensvold, 2002; Davidov, 2008; Steenkamp & Baumgartner, 1998), for this model. The results (see table 1) show that configural and metric invariance were achieved, but not scalar invariance. A multiple group tridimensional bifactor model, which allowed for variation across countries for the factor loadings, item thresholds, residual variances, factor means, and factor variances had a good fit (CFI = 0.97, TLI = 0.95; RMSEA = 0.06, WRMR = 14.67) according to common criteria (e.g., Hu & Bentler, 1999; Rutkowski & Svetina, 2014). When the factor loadings were restricted to be equal, this model fit dropped only slightly and (ΔCFI = -0.01; ΔRMSEA = 0.00), but adding invariance constraints on the thresholds led to a noticeable drop in model fit (ΔCFI = -0.18 and ΔRMSEA = 0.05), which was above the criteria suggested by Rutkowski and Svetina (2014) for cases with large samples and more than two groups. These results support the validity of cross-national comparisons of the size and strength of intra-group differences and correlations of the two OTL dimensions (which require metric invariance only), but they indicate that factor means should not be compared across countries (because they require scalar invariance).

Table 1. Model Fit for a Multiple Group Confirmatory Factor Analysis Testing Cross-National Invariance of Dimensions 2 and 3

		Model Fit				
Model	Free parameters	$\Delta\chi^2$	CFI	TLI	RMSEA	WRMR
Configural invariance	1,484	-	.97	.95	.06	14.67
Metric invariance	1,268	7,976**	.96	.95	.06	17.64
Scalar invariance	620	122,863**	.78	.81	.11	48.56

Note. CFI = Comparative Fit Index; TLI = Tucker–Lewis Index; RMSEA = Root Mean Square Error of Approximation; WRMR = Weighted root-mean-square residual. *$p < .05$, **$p < .01$.

Within-item tridimensional IRT model. Once the model was identified and its cross-national invariance verified, student scores on the common method dimension and on the two OTL dimensions were generated. These scores could have been exported from MPlus but correlations with achievement and with contextual variables would then have been underestimated; this would also have led to an underestimation of the between-school variance (Monseur & Adams, 2009). To overcome these statistical biases, a within-item IRT tridimensional model with two substantive OTL dimensions and one general method factor was used. More precisely, the eleven OTL items were modeled with a mixed coefficient multinomial model, as described by Adams, Wilson, and Wang (1997) and implemented by Conquest® software (Wu, Adams, & Wilson, 1997). This mixed model describes items using a fixed set of unknown parameters, while the student outcome level is a random effect (Adams, 2002). The eleven OTL items were scaled according to a three-dimensional within-item partial credit model, with all OTL items loading on the first dimension and some loading specifically on dimension 2 (*Fiction*) or on dimension 3 (*Non-*

continuous). This conditional model requires a population model defined by a set of regressors, usually denoted "conditioning variables". In this study, several variables were used as conditioning variables: schools as dummies, reading performance, the PISA ESCS index, gender, the track attended by the student and a few indices such as reading enjoyment, teacher stimulation for reading engagement, use of structuring and scaffolding strategies, memorization, understanding and remembering, and elaboration. Five plausible values were drawn for each student per OTL dimension. This population model ensures that secondary analyses on the common method dimension and on the two OTL dimensions will no longer be biased by the unreliability of the measures, as far as the covariates have been used as conditioning variables. Of course, it does not correct for systematic bias such as response style effects.

OTL items for the two target dimensions were selected according to major features of the PISA framework (OECD, 2009), namely type of text (continuous vs. non-continuous) and aspect (retrieve information, interpret and reflect), drawing on what is known about the effectiveness of reading instructional strategies (Mc Namara, 2007; Pressley, 2000; Rosenshine & Meister, 1997). The four items allocated to the *Non-continuous* dimension are: two items concerning reading of non-continuous texts (*Texts that include diagrams or maps*, and *Texts that include tables or graphs)* and two "retrieve" or "evaluate" items in non-continuous texts (*Find information from a graph, diagram or table* and *Describe the way the information in a table or graph is organized*). The four items allocated to the second dimension are *Reading fiction (novels* and *short stories)* and three "interpret" or "reflect" items also related to fiction texts (*Explain the cause of events in a text, Explain the way characters behave in a text*, and *Explain the purpose of a text)*.

Correlations with reading achievement. Correlations with reading achievement of the three IRT dimensions were then computed at the student, the school and

the country level. At the student- and school-level (Pearson) product-moment correlations were computed. At the country level, product-moment correlations and (Spearman) rank-correlations were computed, because some "outlier" countries influenced the product-moment correlations. Standard errors were computed by using the weight replicates provided in the PISA 2009 database. These correlations with reading achievement were compared with the corresponding correlations of the two OTL factors extracted by the PCA.

Variance decomposition for the common method factor. The question to what extent the common method dimension varies at the country, school and student levels was examined by performing a three-level regression analysis with an empty model.

Correlations of the common method factor with other response style indicators. Two indices of response styles were computed on a set of 41 other PISA items also using Likert scale format (4 points scale from *strongly disagree* to *strongly agree* - 20 items in total, including 7 negative items); and frequency scales/lessons (21 items) ("How often do these things happen in your classroom?" (4-points scale ranging from "*Never or hardly ever"* to "*In all lessons"*).

- An Acquiescence Response Style (ARS) index: the proportion of responses in the two *agree* categories of a scale minus the proportion of responses in the *disagree* categories divided by the number of items (see Van Herk, Poortinga, & Verhallen, 2004).

- An Extreme Response Style (ERS) index: the proportion of responses in the *strongly agree* and *strongly disagree* categories divided by the number of items (ibid.).

Country-level product-moment and rank-correlations of the common method dimension with these two indicators were computed.

Analyzing country patterns. Country means were correlated with indices of the country's average SES and reading achievement. Again, both product-moment and rank-correlations were used.

6. Results

6.1. Improving the predictive validity of the OTL factors

To examine whether correcting for a method factor, improved the predictive validity of the OTL scales and solved the attitude-achievement paradox (hypothesis 1a and 1b) we compared correlations of the two OTL factors with student achievement at different levels of analysis (student, school and country level) before and after correction for method bias. The results are shown in table 2.

Table 2. Average Correlations of the Two OTL Factors with Reading Achievement at the Student School, Country Level Before (PCA Model) and After (Within-Item Multidimensional IRT Model) Partialling Out the General Method Factor

	PCA model		within-item multidimensional IRT model		
Level of analysis	OTL Fiction	OTL Non-continuous	OTL Fiction	OTL Non-continuous	Common method dimension
Student level (Product-moment correlation)	.17	.09	.40	.32	-.24
School level (Product-moment correlation)	.36	.16	.52	.43	-.38
Country level (Product-moment correlation)	-.32	.22	.05	.48	-.50
Country level (rank-correlation)	-.37	.32	-.05	.62	-.28

Without correction for method bias, the two OTL factors correlated positively with reading achievement at the student and school levels, but these correlations were low at the student level (.17 and .09) and even at the school level (.36 and .16). Moreover, the attitude-achievement paradox was observed at the student/school vs. Country level for the *Fiction* factor. That is, a moderate negative correlation with reading achievement was observed at the country

level (-.32), meaning that in the countries in which more opportunities to read and interpret *fiction* texts were offered to students, students performed worse, which is paradoxical. Rank-correlations were somewhat stronger at the country level, due to some outlier countries influencing the correlation, but also paradoxical for the *Fiction* factor (-.37). Hence, response styles like ARS, ERS or social desirability (Yang *et al.*, 2012) were suspected.

When we partialled out a general method factor by using the within-item tridimensional IRT-model, the picture changed: The target OTL dimensions showed, on average, substantial positive correlations with achievement at the student and school levels (.40 for *Fiction* and .32 for *Non-continuous* at the student level and .52 for *Fiction* and .43 for *Non-continuous* at the school level). At the country level the OTL *Fiction* dimension showed a close to zero correlation (.05 product-moment correlation and -.05 rank-correlation) and the OTL *Non continuous* factor showed a strong positive correlation (.48 product-moment and .62 rank-correlation) with reading achievement. Hence, the correlations for both dimensions were stronger and more positive at the student and school-levels than they were without the correction, and the same held true for the country-level correlation for the *Non-continuous* factor. For the *Fiction* factor the country-level correlation moved from negative to zero.

Table 2 also displays correlations of the common method with reading achievement at the student, school and country levels[5]. At the student and school levels, the common method dimension was negatively related to reading achievement (-.24 and -.38), At the country-level, the common method dimension was negatively correlated with reading achievement (-.50). Hence, more method bias was found for the responses of lower achieving students,

[5] Additionally, correlations at the school level by country are displayed in Appendix A, table A1.

schools and countries.

6.2. Understanding the common method dimension

To gain a better understanding of what the common method dimension in the within-item IRT-model captured, its variance was decomposed at different levels of analysis, its country-level correlation with other response style indicators was examined, and country patterns concerning average method factor scores and their relation to other variables were analyzed.

Results of the variance decomposition suggest that the largest proportion of variance lies at the student level, followed by the school and country levels: About 76% of the variance of the common method dimension occurred between students, 15% between schools within countries and 9% between countries. Additionally, we also examined the decomposition of the variance between the student and the school level within each country. As shown in table A2 in Appendix A, the magnitude of the intraclass correlation of the common method dimension showed some variation between countries; ranging from .08 in Finland, .10 in Japan and Korea to .28 in Hungary and .34 in Italy. Hence, common method bias appears to be largely an individual phenomenon, but systematic differences between schools and countries are also observed.

Results of country-level correlations of the common method bias with other response style indicators (see table 3) show that the ARS index was strongly and positively correlated with the common method dimension at the country level (.72 product-moment correlation and .74 rank-correlation), whereas country-level correlations of ERS with the common method dimension were rather weak (.06 product-moment correlation and .18 rank-correlation). Hence, country-level differences in the common method dimension appear to mainly represent differences in the cultural tendency to acquiesce.

Table 3. Country-Level Correlations between Common Method Dimension and Other Indices of Response-Styles

	Common method dimension
ARS (Product-moment correlation correlation)	0.72
ERS (Product-moment correlation correlation)	0.06
ARS (rank-correlation)	0.74
ERS (rank-correlation)	0.18

Countries scoring high on the common method dimension were Mexico, Chile and Turkey, and to a lesser extent, Denmark and Greece; countries scoring low were Iceland, Norway, Sweden, Finland and Italy, Luxemburg, Czech Republic, Japan and Korea.

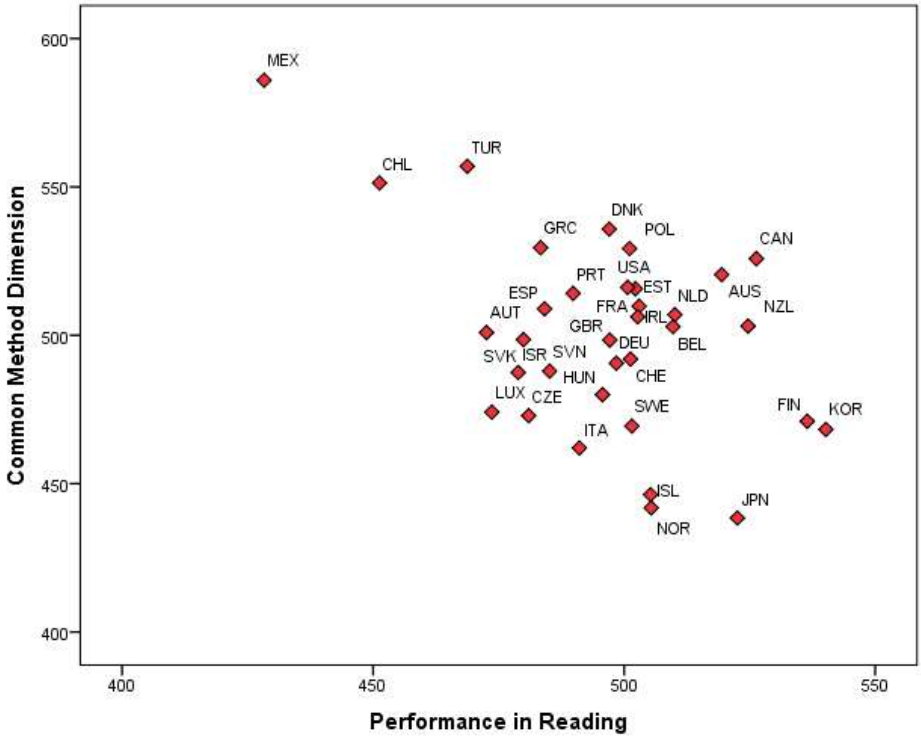

Figure 1. Country-level correlations between the common method dimension and PISA reading achievement.

Figure 1 shows that those countries scoring high on the common methods dimension also tended to have poorer average reading performance (product-moment correlation r_{xy} = -.50; rank-correlation r_{xy} = -.28).

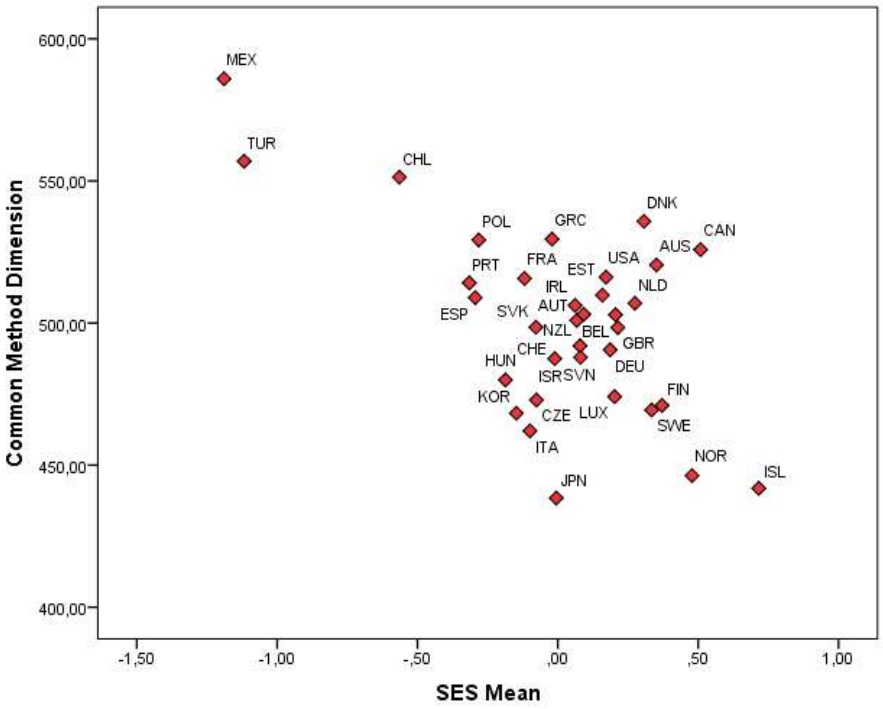

Figure 2. Country-level correlations between the common method dimension and socioeconomic status (average HISEI).

Moreover, figure 2 shows that the less affluent countries scored higher on the common method dimension than the more affluent ones (product-moment correlation r_{xy} = -.60; Rank-correlation r_{xy} = -.31). Analyses of partial correlations further suggest that affluence is more relevant for understanding differences in method bias than reading achievement: under control of the country's reading performance, the effect of socioeconomic status remained strong (r_{xy} = -.42), while the partial correlation with reading proficiency substantially decreased when socioeconomic status was kept under control (r_{xy} = -.17).

7. Discussion

The use of Likert scales in international studies to measure contextual or attitudinal constructs suffers from different limitations and is prone to biases such as cross-cultural response styles. Those biases can be rooted in individuals, groups or cultures (He & Van de Vijver, 2013; Van de Vijver & He, 2014). Whatever their origin, they limit the validity of the measurement of constructs and lead to side effects such as the attitude-achievement paradox (Kyllonen & Bertling, 2013; Lie & Turmo, 2005; van de Gaer, Grisay, Schulz, & Gebhardt, 2012; Lu & Bolt, 2015).

In the present study, several analyses were applied to the PISA 2009 reading OTL scale, firstly to confirm the existence of response style bias, secondly to try to improve the validity of the measurement of OTL dimensions, and thirdly to understand whether common method bias is an individual, a group-level, or a country-level phenomenon, whether country-level differences in this dimension capture cultural differences in response styles, and whether they are related to a country's affluence and achievement level in reading.

The correlations between the two OTL factors extracted by a PCA were weaker at the student and at the school level than could be theoretically expected: hypothesis 1a is thus confirmed. At the country level, in accordance with evidence from numerous international studies, the attitude-achievement paradox was observed for the *Fiction* factor, but was not observed for the *Non-continuous* factor. Countries with higher achievement levels scored lower with regard to their reported OTL for reading fiction in secondary schools, but slightly higher with regard to their reported OTL for reading non-continuous texts. On a theoretical basis, countries where students are exposed to more of both types of OTL would be expected to achieve higher average reading results than countries where students experience fewer OTL. Hypothesis 1b is then partially confirmed.

The use of a mixed coefficient within-item tridimensional IRT model, which aimed at disentangling the common method dimension from the target constructs, gave promising results. The two OTL dimensions (*Fiction* and *Non-continuous text*) were now consistently and positively linked with reading proficiency at the student- and school-levels – in contrast to the results for the PCA model. So, within countries, students enrolled in schools performing better also reported more frequent exposure to beneficial OTL, and within schools those students reporting more OTL also performed better in reading (hypothesis 2a confirmed). In addition, using this model, the attitude-achievement paradox observed for the dimension *Fiction* was less pronounced (hypothesis 2b partly confirmed). The common method dimension on which all items were allocated was significantly and negatively related to achievement at all three levels. Hence, students, schools, and countries performing less well showed a stronger common method bias, explaining why correlations of OTL with achievement were weak when the method bias was not controlled for. Hence, the modeling of a common method dimension by using a within-item tridimensional IRT model consistently improved the predictive validity of the OTL constructs.

The positive results in terms of predictive validity of the within-item tridimensional IRT model are in accordance with the results of Lie and Turmo (2005) who successfully used a general factor ("superconstruct") to adjust the scores of attitudinal scales in PISA 2003, and were able to partially fix the attitude-achievement paradox. Similarly, Van de Gaer (2010) performed a CFA on several attitudinal constructs in PISA 2003 and also integrated a general response style factor into the model. She found that the country-level correlations between attitudes and achievement, originally negative, became positive. Results were less convincing when applying an IRT model for ERS (Lu & Bolt, 2015). Despite detectable variation of ERS across countries, the changes in correlations between attitudes and achievement before and after correction

for ERS did not substantially change.

The difference in the results between Lie and Turmo (2005), van de Gaer (2010) and the current study, on the one hand, and Lu & Bolt (2015), on the other hand, might be due to the fact that the latter study modeled only ERS, while the three other ones used a general response style/common method dimension for adjustment which at least in our case appears to mainly reflect ARS. He and Van de Vijver (2015) also applied an adjustment for social desirability and general response style to the TALIS constructs and found "negligible correction effects in teachers' self-report" (He & Van de Vijver, 2015, p. 283). However, they only examined effects on cross-cultural differences in mean scores, not on the predictive validity of the self-report variables. Altogether, the findings of our and other studies suggest that partialling out common method factors can improve the validity of measures in large scale assessments, but that controlling for ERS alone does not seem to be sufficient.

In accordance with Lu and Bolt (2015), we expected a small amount of variance of the common method dimension to be between countries; most of the variance to be between individuals within schools, and the proportion of variance at the school level varying from one country to another. The results matched well the hypothesis 4. Nine percent of the variance of the common method dimension was between countries, 15% on average between schools within country and 76% on average between students within schools. Our results are coherent with Lu and Bolt's (2015) findings regarding PISA 2006. Applying a two level IRT model for ERS to seven attitudinal scales, they found an intra-class correlation of the ERS of 0.31, and highlighted that "although country level differences in ERS are detectable, they are relatively small compared to within-country variation in ERS" (Lu & Bolt, 2015, p. 15).

Our results also provide evidence that cross-national differences in the common method dimension reflect differences in response styles, especially in

ARS. Strong country-level correlations between the ARS index and the common method dimension were observed. Hypothesis 3 is therefore confirmed. These results are consistent with findings from several other studies (He & Van de Vijver, 2013, 2015; Van de Vijver & He, 2014; van de Gaer, 2010). Using data from a national study in the Netherlands on the one hand and from TALIS on the other hand, "support for a general response style factor" was found and it was shown that "the different response styles have both shared and unique aspects" (He & Van de Vijver, 2013, p. 799). In the cited study, "the general response style explained 28% of the variance, suggesting that there was considerable overlap in response styles, although clearly not all variation was captured" (He & Van de Vijver, 2013, p.797). However, in our study correlations of the ERS index with the common method were much weaker than the ones observed for ARS. This latter result is congruent with findings from Lu and Bolt (2015), showing that ERS has limited consequences in terms of attitudes-achievement paradox in the PISA context.

These systematic cross-national differences in response styles, operationalized as a common method dimension, further appear to be linked to economic factors. In our study, strong negative correlations were observed at the country level between the common method dimension and reading achievement on the one hand, and socioeconomic status on the other hand; low performing and less affluent countries scored higher on the common method factor. Hypothesis 5b is thus confirmed. These results are congruent with results from previous studies (Lie & Turmo, 2005; van de Gaer, Grisay, Schulz, & Gebhardt, 2012; Van de Vijver & He, 2014) which repeatedly showed that less affluent and lower performing countries score higher on the general response style factor. "Findings suggest that in countries with higher levels of economic development and educational achievement, respondents are less inclined to demonstrate the studied response styles than respondents in countries with lower levels of socioeconomic development and educational achievement" (Van de Vijver &

He, 2014, p. 24). As achievement and socioeconomic development are themselves highly correlated, it is difficult to disentangle effects of ability or cognitive aspects versus effects of culture. However, we fully agree with what Yang, Harkness, Chin, and Villar (2010) sustain in their critical review of response styles and culture: "the possible influence of a variety of factors on cognitive processing should be addressed before attributing response behaviours to culture" (p. 219). The "satisficing" behaviour (Krosnick, 1991) is a plausible explanation for several response styles (namely choosing extreme or mid-points) and can reflect the level of motivation or engagement of the respondent, but also his/her reading ability. When a student with very low reading ability is faced with quite demanding questions in terms of reading load, his/her main options are to omit answering or to use some less cognitively demanding pattern for answering – which gives similar results such as response style. Two sets of results in our study provide arguments for a limited role of culture on the response style in the PISA context. First of all, most of the variation of the common method dimension is observed within countries, not between countries. Further, this latter country variability is mainly due to outliers – namely Chile, Mexico and Turkey. In addition, as will be underlined in the limitations section hereafter, a lot of missing answers were observed in non-OECD countries for the OTL variables, located at the end of the cognitive booklets instead of the students' questionnaires as it should be[6]. Obviously, the level of achievement in reading was related to students' level of engagement and perseverance in answering the OTL questions; it is quite likely that in these countries a number of students simply did not reach the OTL questions.

Even though caution is suggested concerning the interpretation of common method bias or response styles in terms of cross-cultural differences, we

[6] The motivation for this unusual allocation of non-cognitive questions in the cognitive section is mainly lack of space in the Student questionnaire, and the fact that the OTL variables were rated low in terms of priority.

checked whether the patterns of differences observed between countries matched the main findings of the literature on cross-cultural differences in response-styles. We expected to find the following patterns: the countries from South America (compared to the North) and possibly South of Europe (compared to northern and western Europe) would score higher on the common method dimension, while Asian countries and Scandinavian countries would score the lowest (hypothesis 5a). This pattern of results was to a large extent observed: the two countries from South America (Mexico and Chile) and Turkey scored the highest on the common method dimension, while Asian countries (Japan and Korea) and some of the Nordic countries (Finland, Norway, Sweden and Iceland, but not Denmark) scored the lowest. No clear pattern was observed for the countries from Southern Europe: some countries scored quite high (Greece, Portugal, Spain), others quite low (Italy). This latter finding is coherent with some previous studies. Van Herk, Poortinga and Verhallen (2004) sustained a north-south divide in Europe, and also observed that the ARS and ERS is especially high in Greece which was also the case in the present study. Nevertheless, Yang, Harkness, Chin, and Villar (2010) also underline that research in Europe "has produced conflicting results" (p. 209) and that finding consistent patterns among European countries is difficult. One possible explanation might be that many European countries, especially the Western ones (Belgium, Germany, France, Luxembourg, the Netherlands, United Kingdom) are nowadays multicultural societies. He and Van de Vijver (2013) investigated cultural differences in response style across several ethnic groups in the Netherlands, and found significant differences in response styles; comparatively, non-Western immigrants (originated from Morocco, Turkey, Surinam and the Netherlands Antilles) showed higher ARS and midpoint responding behavior. There are, hence, limitations regarding the use of countries or sets of countries to investigate cultural differences. In our contemporary multicultural societies, cultural variation is more and more likely to happen within countries.

7.1. Limitations and perspectives

This study has several limitations. It would have been relevant to involve a broader and more diverse set of countries, taking advantage of the non-OECD countries participating in PISA 2009. Only OECD countries were included; the rationale is that the questions about reading OTL were included at the end of the cognitive booklets which resulted in a substantial amount of missing data, especially in the non-OECD countries. Inclusion of a more diverse set of countries could have resulted in more variation in common method explained at the country level.

Even if the modeling approach worked quite well and evidence was found that the within-item tridimensional IRT modeling improved the predictive validity of the OTL constructs and seemed an effective approach to partially rule out the attitude-achievement paradox, the OTL variables are possibly not the most relevant ones to demonstrate how effective the within-item tridimensional modeling approach could be to peel off target constructs from common method bias. Indeed, the OTL items asked the students to rate the frequency of OTL in their classes. A Likert-scale was used with self-report items, but not with items requiring self-evaluation. It has been shown that "response styles are triggered most in questions about personal domains when evaluation apprehension is the strongest" (van Dijk *et al.,* 2009; He & Van de Vijver, 2014). So, on the one hand, common method bias is possibly less strong for these OTL items than it is for self-evaluation measures. On the other hand the potential of the adjustment is possibly underestimated. Consequently, it would be worthwhile to apply the same kind of modeling to other self-reported measures in PISA such as enjoyment of reading or self-concept, anxiety, or interest in mathematics.

7.2. Conclusions

Promising results were gained when implementing the within-item tridimensional IRT model to improve the validity of the PISA 2009 OTL reading scale. The model did not only show a good fit and an acceptable level of cross-national equivalence, "peeling off" the OTL dimensions of the common method dimension also led to an improvement in predictive validity of the OTL scales: Correlations of the two OTL dimensions with reading achievement became stronger and positive and the *attitude-achievement paradox* (opposite signs of correlations at different level of analysis) became less pronounced. We further provided evidence that the common method dimension really measured method bias, mainly ARS, and not some substantive content. Hence, the use of within-item multidimensional IRT models can be considered one promising approach to improving the validity of self-report measures.

References

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

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606. doi:10.1037/0033-2909.88.3.588

Buckley, J. (2009). *Cross-National Response Styles in International Educational Assessments: Evidence from PISA 2006.* Retrieved from https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf

Chen, C., Lee, S., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Sciences, 6*, 170–175. doi:10.1177/0146621697211001

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1–27. doi:10.1177/014920639902500101

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255. doi:10.1207/S15328007SEM0902_5

Crowne, D. P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349–354. doi:10.1037/h0047358

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the European Social Survey. *Survey Research Methods, 2*, 33–46.

Floden, R. E. (2002). The measurement of Opportunity-to-learn. In A. C. Porter & A. Gamoran (Eds.), *Methodological Advances in Cross-National Surveys of Educational achievement* (pp. 231–267). Washington DC: National Academy Press.

Ferrando, P. J., & Condon, L. (2006). Assessing acquiescence in binary responses: IRT-related item-factor-analytic procedures. *Structural Equation Modeling, 13*, 420–439. doi:10.1207/s15328007sem1303_5

Harzing, A.W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management, 6*, 243–266. doi:10.1177/1470595806066332

He, J., & Van de Vijver, F. J. R. (2012). Bias and equivalence in cross-cultural research. *Online readings in Psychology and Culture, 2*. doi:10.9707/2307-0919.1111

He, J., & Van de Vijver, F. J. R. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences, 55,* 794-800. doi:10.1016/j.paid.2013.06.017

He, J., & Van de Vijver, F. J. R. (2015). The value of keeping an open eye for methodological issues in research on resilience and culture. In L. Theron, M. Ungar, & L. Liebenberg (Eds.), *Youth resilience and culture. Commonalities and complexities* (pp. 189-201). New York, NY: Springer.

He, J., & Van de Vijver, F. J. R., Dominguez Espinosa, A., Abubakar, A, Dimitrova, R., Adams, B. G., Aydinli, A., Atitsogbe, K., Alonso-Arbiol, I., Bobowik, M., Fischer, R., Jordanov, V., Mastrotheodoros, S., Neto, F., Ponizovsky, J. Reb, J., Sim, S., Sovet, L., Stefenel, D., Suryani, A. O., Tair, E., & Villieux, A. (2015). Socially Desirable Responding: Enhancement and Denial in 20 Countries. *Cross-Cultural Research, 49*, 227-249. doi:10.1177/1069397114552781

Heine, S. J., Kitayama, S., Lehman, D. R., Takata, T., Ide, E., Leung, C., & Matsumoto, H. (2001). Divergent consequences of success and failure in Japan and North America: An investigation of self-improving motivations and malleable selves. *Journal of Personality and Social Psychology, 81*, 599-615. doi:10.1037/0022-3514.81.4.599

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales: The reference-group problem. *Journal of Personality and Social Psychology, 82*, 903-918. doi:10.1037/0022-3514.82.6.903

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance

structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. doi:10.1080/10705519909540118

Kobayashi, C. & Greenwald, A. J. (2003). Implicit-explicit differences in self-enhancement for Americans and Japanese. *Journal of Cross-Cultural Psychology, 34*, 522–541. doi:10.1177/0022022103257855

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213–236. doi:10.1002/acp.2350050305

Kurman, J. (2003). Why is self-enhancement low in certain collectivist cultures? An investigation of two competing explanations. *Journal of Cross-Cultural Psychology, 34*, 496–510. doi:10.1177/0022022103256474

Kyllonen, P. & Bertling, J. (2013). Innovative questionnaire assessment methods to increase cross-cultural comparability. In L. Rutkowski, M. von Davier, D. Rutkwoshi (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues and Methods of Data Analysis* (pp. 277–285). Boca Raton, FL: Chapman & Hall/CRC.

Lafontaine, D., Baye, A., Vieluf, S., & Monseur, M. (2015). Equity in opportunity-to-learn and achievement in reading: A secondary analysis of PISA 2009 data. *Studies in Educational Evaluation, 47,* 1–11. doi:10.1016/j.stueduc.2015.05.001

Lee, C. & Green, R.T. (1991). Cross-cultural examination of the Fishbein Behavioral Intentions Model. *Journal of International Business Studies, 22*, 289–305. doi:10.1057/palgrave.jibs.8490304

Lie, S., & Turmo, A. (2005). *Cross-country comparability of students' self-reports Evidence from PISA 2003*. Internal Working OECD/PISA document, TAG(0505)11.

Lindell, M. K., & Brandt, C. J. (2000). Climate quality and climate consensus as mediators of the relationship between organizational antecedents and outcomes. *Journal of Applied Psychology, 85*, 331–348. doi:10.1037/0021-9010.85.3.331

Lindell, M. K. & Whitney, D. J. (2001). Accounting for common method variance in cross-sectional research designs. *Journal of Applied Psychology, 86*, 114–

121. doi:10.1037/0021-9010.86.1.114

Lu, Y., & Bolt, D. (2015). Examining the attitude-achievement paradox in PISA using a multilevel multidimensional IRT model for extreme response style. *Large-scale Assessments in Education, 3*. DOI 10.1186/s40536-015-0012-0

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525–542. doi:10.1007/BF02294825

Monseur, C., & Adams, R. J. (2009). Plausible values: how to deal with their limitations. *Journal of Applied Measurement*, *10*, 320–334.

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*, 277–302. doi:10.1023/A:1024472110002

Muthén, B., & Kaplan D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171–189. doi:10.1111/j.2044-8317.1985.tb00832.x

OECD (2009). PISA 2009 Assessment Framework. Key competencies in reading, mathematics and science. Paris: OECD.

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, Inc.

Peng, K., Nisbett, R. E., & Wong, N. Y (1997). Validity problems comparing values cultures and possible solutions. *Psychological Methods, 2,* 329–344. doi:10.1037/1082-989X.2.4.329

PISA 2009 Data (2009). *OECD, PISA 2009 database*. Retrieved from: https://www.oecd.org/pisa/pisaproducts/pisa2009database-downloadabledata.htm.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y. & Podsakoff, N. P. (2003). Common method biases in behavioural research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*, 879–903.

doi:10.1037/0021-9010.88.5.879

Poortinga, Y. H. (1989). Equivalence of cross cultural data: An overview of basic issues. *International Journal of Psychology, 24*, 737–756.

Pressley, M. (2000). What should comprehension instruction be the instruction of? In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of Reading Research*, Vol III (pp. 545–563). Mahwah, NJ: L. Erlbaum.

Richards, J. M., Gottfredson, D. C., & Gottfredson, G. D. (1990/1991). Units of analysis and item statistics for environmental assessment scales. *Current Psychology: Research and Reviews, 9*, 407–413. doi:10.1007/BF02687196

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351–57. doi:10.2307/2087176

Rosenshine, B., & Meister, C. (1997). Cognitive Strategy Instruction in Reading. In A. Stahl & A. Hayes (Eds.), *Instructional Models in Reading* (pp. 85–107). Mahwah, NJ: L. Erlbaum.

Ross, C.E. & Mirowsky, J. (1984). Socially-Desirable Response and Acquiescence in a Cross-Cultural Survey of Mental Health. *Journal of Health and Social Behavior 25*, 189–97. doi:10.2307/2136668

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). Münster, Germany: Waxmann.

Rutkowski, L. & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*, 31–57. doi:10.1177/0013164413498257

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66,* 507–514. doi:10.2139/ssrn.199064

Schmidt, W. H., & Maier, A. (2009). Opportunity to learn. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of Education Policy Research.*

New York, NY: Routledge.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content area literacy. *Harvard Educational Review, 78*, 40–60. doi:10.17763/haer.78.1.v62444321p602101

Shen, C., & Tam, H. P. (2008). The paradoxical relationship between student achievement and self-perception: a cross-national analysis based on three waves of TIMSS data. *Educational Research and Evaluation, 14*, 87–100. doi:10.1080/13803610801896653

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78–90. doi:10.1086/209528

Takahashi, K., Ohara, N., Antonucci, T. C., & Akiyama, H. (2002). Commonalities and differences in close relationships among the Americans and Japanese: A comparison by the individualism/collectivism concept. *International Journal of Behavioral Development, 26*, 453–465. doi:10.1080/01650250143000418

Van de Gaer, E., & Adams, R. (2010). *The modeling of response-style bias: an answer to the attitude-achievement paradox?* Paper presented at the annual conference of the AERA, Denver, Colorado, 30 April-4 May.

Van de Gaer, E., Grisay, A., Schlutz, 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. doi:10.1177/0022022111428083

Van de Vijver, F. J. R., & He, J. (2014). *Report on social desirability, midpoint and extreme responding in TALIS 2013*. OECD Education Working Papers, No. 107. Paris, France: OECD Publishing.

Van de Vijver, F.J.R., & Poortinga, Y. (1982). Cross-Cultural generalization and universality. *Journal of Cross-Cultural Psychology, 13*, 387–408. doi:10.1177/0022002182013004001

Van de Vijver, F.J.R., & Tanzer, N. K. (2004). Bias and equivalence in cross-

cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology, 54*,119–135. doi:10.1016/j.erap.2003.12.004

van Dijk, T.K., Datema, F., Piggen, A.-L. J.H.F., Welten, S.C.M., van de Vijver, F.J.R. (2009). Acquiescence and extremity in cross-national surveys : Domain dependence and country-level correlates. In A. Gari & K. Mylonas (Eds.), *Quod erat demonstrandum: From Herodotus' ethnographic journeys to cross-cultural research*. Athens, Greece: Pedio Books Publishing.

Van Herk, H., Poortinga, Y.H., & Verhallen, T.M.M. (2004). Response styles in rating scales: Evidence of method bias in data from 6 EU countries. *Journal of Cross-Cultural Psychology*, *35*, 346–360. doi:10.1177/0022022104264126

van Rosmalen, J., van Herk, H., Groenen, P. J. F. (2010). Identifying Response Styles: A Latent-Class Bilinear Multinomial Logit Model. *Journal of Marketing Research, 47*, 157–172. doi:10.1509/jmkr.47.1.157

Vieluf, S., Kunter, M., & van de Vijver, F. J. R. (2013). Teacher self-efficacy in cross-national perspective. *Teaching and Teacher Education, 35*, 92–103. doi:10.1016/j.tate.2013.05.006

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 de Ark, D. M. Bolt & C. M. Woods (Eds.), *New developments in quantitative psychology* (pp. 463–487). New York, NY: Springer.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *Conquest®: Multi-Aspect Test Software (computer program manual)*. Camberwell, Vic: Australian Council of Educational Research.

Yang, Y., Harkness, J. A., Chin, T.-Y., & Villar A. (2012). Response styles and culture. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 203–223). USA: John Wiley & sons.

Appendix A

Table A1. School level correlations of the two OTL-dimensions (Non-continuous and Fiction) and the common method dimension with reading achievement

Country	Common method dimension	OTL Fiction	OTL Non-continuous
AUS	-0.21 (0.086)	0.57 (0.045)	0.41 (0.070)
AUT	-0.60 (0.047)	0.70 (0.030)	0.58 (0.037)
BEL	-0.63 (0.040)	0.61 (0.038)	0.49 (0.044)
CAN	-0.10 (0.039)	0.33 (0.042)	0.38 (0.038)
CHE	-0.35 (0.069)	0.61 (0.056)	0.36 (0.105)
CHL	-0.33 (0.071)	0.38 (0.067)	0.36 (0.078)
CZE	-0.55 (0.047)	0.66 (0.041)	0.49 (0.041)
DEU	-0.42 (0.044)	0.61 (0.030)	0.42 (0.041)
DNK	-0.04 (0.063)	0.27 (0.076)	0.25 (0.062)
ESP	-0.36 (0.044)	0.50 (0.039)	0.27 (0.049)
EST	-0.30 (0.058)	0.25 (0.056)	0.28 (0.056)
FIN	0.04 (0.083)	0.18 (0.096)	0.28 (0.072)
FRA	-0.51 (0.057)	0.71 (0.036)	0.73 (0.032)
GBR	-0.37 (0.054)	0.46 (0.069)	0.50 (0.048)
GRC	-0.31 (0.088)	0.66 (0.060)	0.17 (0.080)
HUN	-0.76 (0.038)	0.70 (0.043)	0.67 (0.034)
IRL	-0.29 (0.083)	0.22 (0.084)	0.29 (0.076)
ISL	-0.21 (0.028)	0.34 (0.037)	0.34 (0.027)
ISR	-0.55 (0.051)	0.28 (0.065)	0.38 (0.066)
ITA	-0.59 (0.023)	0.56 (0.024)	0.14 (0.036)

JPN	-0.33 (0.069)	0.65 (0.043)	0.49 (0.046)
KOR	0.17 (0.116)	0.71 (0.057)	0.68 (0.059)
LUX	-0.69 (0.024)	0.89 (0.008)	0.75 (0.011)
MEX	-0.03 (0.042)	0.23 (0.040)	0.13 (0.042)
NLD	-0.59 (0.052)	0.72 (0.035)	0.74 (0.032)
NOR	-0.25 (0.064)	0.42 (0.056)	0.38 (0.059)
NZL	-0.58 (0.061)	0.50 (0.060)	0.58 (0.057)
POL	-0.53 (0.054)	0.58 (0.056)	0.49 (0.063)
PRT	-0.45 (0.065)	0.55 (0.055)	0.47 (0.052)
SVK	-0.62 (0.042)	0.53 (0.052)	0.50 (0.054)
SVN	-0.59 (0.014)	0.79 (0.008)	0.74 (0.010)
SWE	-0.02 (0.082)	0.25 (0.063)	0.18 (0.071)
TUR	-0.37 (0.069)	0.57 (0.049)	0.30 (0.069)
USA	-0.53 (0.067)	0.52 (0.072)	0.44 (0.061)
OECD average	-0.38	0.52	0.43

Table A2. Intraclass correlations (ICCs) for the two OTL-dimensions (Non-continuous and Fiction) and the common method dimension

Country	Common method dimension	OTL Fiction	OTL Non-continuous
AUS	0.12	0.22	0.31
AUT	0.25	0.35	0.60
BEL	0.23	0.32	0.41
CAN	0.14	0.24	0.21
CHE	0.20	0.28	0.43
CHL	0.14	0.24	0.23
CZE	0.23	0.40	0.44
DEU	0.22	0.30	0.40
DNK	0.21	0.23	0.26
ESP	0.12	0.17	0.21
EST	0.15	0.26	0.28
FIN	0.08	0.21	0.29
FRA	0.18	0.27	0.30
GBR	0.18	0.22	0.25
GRC	0.16	0.14	0.17
HUN	0.28	0.34	0.37
IRL	0.15	0.15	0.14
ISL	0.12	0.26	0.23
ISR	0.20	0.26	0.37
ITA	0.34	0.20	0.37
JPN	0.10	0.18	0.30
KOR	0.10	0.24	0.24

LUX	0.13	0.40	0.34
MEX	0.16	0.26	0.45
NLD	0.20	0.35	0.44
NOR	0.17	0.28	0.34
NZL	0.12	0.18	0.26
POL	0.10	0.22	0.18
PRT	0.14	0.13	0.24
SVK	0.27	0.36	0.27
SVN	0.18	0.47	0.50
SWE	0.14	0.30	0.31
TUR	0.12	0.19	0.17
USA	0.15	0.19	0.26
OECD average	0.17	0.26	0.31