CROSS-CULTURAL COMPARATIVE QUESTIONNAIRE ISSUES

Kyllonen, P., Burrus, J., Roberts, R. and Van de gaer, E.

Paper for the PISA 2012 Questionnaire Expert Group Meeting, Hong Kong, February 2010

Please do not cite without the permission of the authors.

Consortium:

Australian Council for Educational Research (ACER, Australia)

cApStAn Linguistic Quality Control (Belgium)

Deutsches Institut für Internationale Pädagogische Forschung (DIPF, Germany)

Educational Testing Service (ETS, USA)

Institutt for Lærerutdanning og Skoleutvikling (ILS, Norway)

Leibniz - Institute for Science and Mathematics Education (IPN, Germany)

National Institute for Educational Policy Research (NIER, Japan)

The Tao Initiative: CRP - Henri Tudor and Université de Luxembourg - EMACS (Luxembourg)

Unité d'analyse des systèmes et des pratiques d'enseignement (aSPe, Belgium)

Westat (USA)

TABLE OF CONTENTS

CROSS-CULTURAL COMPARATIVE QUESTIONNAIRE ISSUES

BACKGROUND

1. One of the major challenges of an international study such as PISA is the cross-cultural validity and applicability of all instruments. In this context, a phenomenon has been of concern which has continued to appear across all PISA cycles whereby for a number of attitudinal student context constructs have shown to be linked to performance in unexpected ways. More specifically, at the between-country level, countries that demonstrate higher performance in a subject show less positive attitudes towards that subject whereas more positive attitudes are recorded for lower-performing countries.
2. In this context, it should be noted that the PGB has asked Eugene Owen to oversee how the issue will be addressed in the PISA2012 cycle. In response, an overarching proposal has been drafted outlining three ways to approach the issue, namely through a) secondary analyses of PISA2003 data, b) exploration of new item types in the PISA2012 field trial and c) specification of an analytical design and reporting of data from the main study in 2012.
3. While the proposal is an overarching outline, this document has been prepared in order to obtain the advice of QEG members on the details of how to proceed on this important issue. To this end, the phenomenon is briefly presented first, followed by a summary of previous analyses of the issue, with a focus on PISA data. This is followed by a discussion of new item formats for attitudinal measures. Finally, suggestions and their applications regarding how the issue might be addressed through models and adjustments at the data analysis stage will be presented.

THE PHENOMENON

4. Figure 1 illustrates this phenomenon using data from PISA2003, the previous cycle in which mathematics was the major domain. The figure shows a negative correlation between interest in mathematics and mathematics achievement at the country level. At the one end, countries such as Finland, Japan and Korea with relatively high achievement display low interest in mathematics. At the other end, countries such as Brazil, Indonesia and Tunisia with relatively lower performance in mathematics have high levels of interest in mathematics.

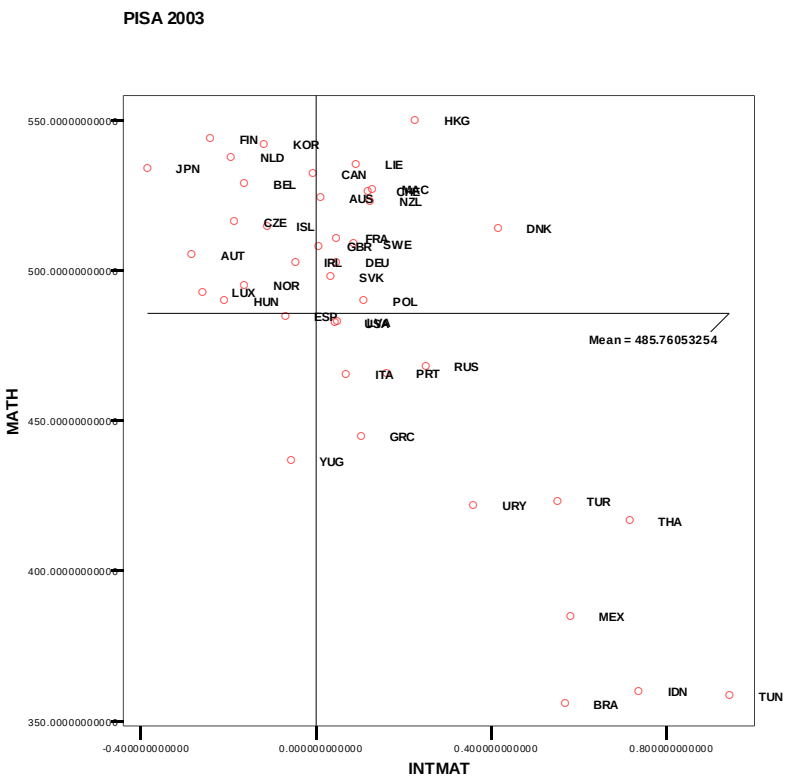


Figure 1 Between-country level correlations between interest in mathematics and mathematics performance in PISA 2003

5. At the between-student within-country level, in contrast, the expected positive correlation reflecting greater interest in mathematics of higher performing mathematics students is found in most countries which is in line with motivational theories such as the expectancy-value theory (Atkinson, 1957; Eccles et al., 1983).
6. An examination of Figure 1 could lead to the conclusion that the higher students perform in mathematics, the lower their interest in mathematics. Such an incorrect conclusion where results of analyses at the group level (here countries) are used to draw inferences to the individual within that group (here students) has been termed an "ecological fallacy" (Robinson, 1950).
7. While this difference in the relationship with achievement at the between-country level and the between-student within-country level applies to some of the student context constructs, it does not apply to all. Table 1 and Figure 2 provide further information about which constructs are affected in what way, again using PISA 2003 data.

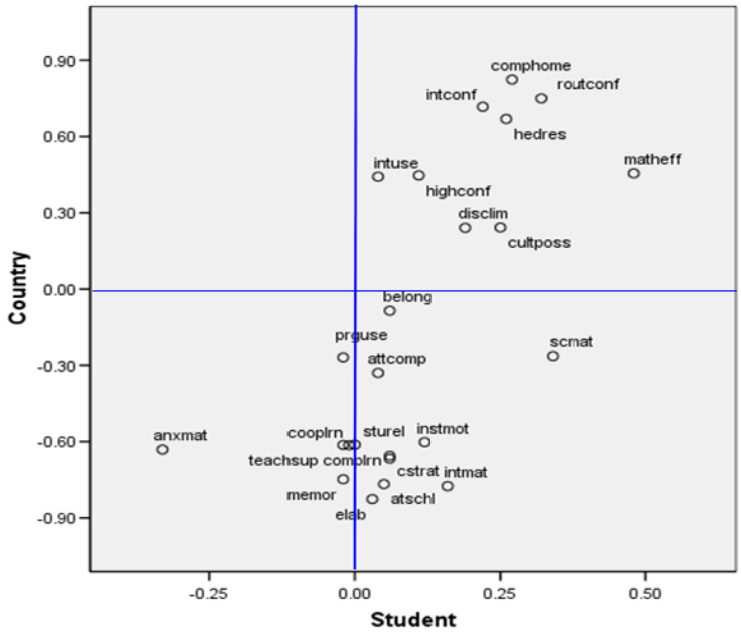


Figure 2 Relationship of correlations between interest in mathematics and mathematics achievement at the between-student within-country level and the between-country level in PISA 2003

8. Figure 2 shows three groups of student context constructs according to their between-country and between-student within-country correlation with mathematics achievement. The first group in the upper right hand quadrant shows constructs for which the correlation with achievement is positive at both levels. Hence, these constructs do not display the phenomenon as students in higher performing countries tend to provide positive responses while at the same time higher performing students tend to provide positive responses. The construct in the second group in the lower left-hand quadrant also does not display the phenomenon in that students in lower performing countries tend to provide negative responses and poorer performing students tend to provide negative responses. Constructs in the third group in the lower right-hand quadrant, however, display the phenomenon. For these constructs, students in lower performing countries tend to provide positive responses whereas, within countries, higher performing students tend to provide positive responses. This phenomenon, while presented for PISA2003 data has been demonstrated also in PISA2000 and 2006 (Van de Gaer & Han, 2009).

Table 1 Correlations between student context constructs and mathematics achievement at the between-student within-country and the between-country level, PISA 2003

Construct name	Construct label	B'ween country corr.	Between-student within-country correlation	Item format (4 or 5 point Likert scale)
Student background				
HEDRES	Educational resources at home	0.67	0.26	Tick box if resource at home
CULTPOSS	Cultural possessions at home	0.25	0.25	Tick box if possession at home
COMPHOME	Possession of a computer at home	0.77	0.28	Tick box if computer at home
Self-concept				
SCMAT	**Mathematics self-concept**	**-0.20**	**0.34**	**Str. agree – Str. disagree**
Self-efficacy				
MATHEFF	Mathematics self-efficacy	0.45	0.49	Very confident – Not at all confident
ICT self-efficacy				
HIGHCONF	ICT: Confidence in high-level tasks	0.38	0.10	I can do this very well by myself– I don't know what this means
INTCONF	ICT: Confidence in internet tasks	0.67	0.21	I can do this very well by myself – I don't know what this means
ROUTCONF	ICT: Confidence in routine tasks	0.63	0.31	I can do this very well by myself – I don't know what this means
Interest and motivation				
ATSCHL	**Attitudes towards school**	**-0.72**	**0.05**	**Str. agree – Str. disagree**
ATTCOMP	ICT: Attitudes towards computers	-0.04	0.03	Str. agree – Str. disagree
INSTMOT	**Instrumental motivation in mathematics**	**-0.57**	**0.13**	**Str. agree – Str. disagree**
INTMAT	**Interest in mathematics**	**-0.74**	**0.16**	**Str. agree – Str. disagree**
Learning strategies				
CSTRAT	**Control Strategies**	**-0.57**	**0.06**	**Str. agree – Str. disagree**
ELAB	**Elaboration Strategies**	**-0.80**	**0.02**	**Str. agree – Str. disagree**
MEMOR	Memorisation Strategies	-0.77	-0.02	Str. agree – Str. disagree
School climate				
BELONG	Sense of belonging at school	-0.07	0.06	Str. agree – Str. disagree
DISCLIM	Disciplinary climate in math lessons	0.20	0.19	Every lesson – never or hardly ever
STUREL	**Student-teacher relations at school**	**-0.58**	**0.02**	**Str. agree – Str. disagree**
TEACHSUP	Teacher support in math lessons	-0.59	-0.01	Every lesson – never or hardly ever
Engagement in learning activities				
INTUSE	ICT: Internet/entertainment use	0.39	0.04	Almost every day - Never
PRGUSE	ICT: Programs/software use	-0.28	-0.02	Almost every day - Never
Other attitudes				
ANXMAT	Mathematics anxiety	-0.57	-0.33	Str. agree – Str. disagree
COMPLRN	**Competitive learning**	**-0.65**	**0.07**	**Str. agree – Str. disagree**
COOPLRN	Co-operative learning	-0.51	-0.01	Str. agree – Str. disagree

Note: Constructs displaying positive correlations with achievement at the between-student within-country level and negative correlations with achievement at the between-country level are given in bold.

9. Several observations can be made from Figure 2 and Table 1 regarding the constructs that show the phenomenon and those that do not show it:

- None of the constructs related to ICT show the phenomenon. This could stem from two things: First ICT is linked to wealth (at the between student within country as well as the between country level). Second, the reference points (Computers, software, internet) are more concrete than for other scales. This also applies to HEDRES, COMPHOME and CULPOSS where students have to indicate whether or not specific items are available at home. Cross-cultural differences have been shown to be smaller where referents are more concrete.
- Whereas mathematics self concept shows the phenomenon, mathematics self-efficacy as well as the ICT efficacy measures do not show the phenomenon. Marsh, Trautwein, Lüdtke, and Köller (2008) noted that the phenomenon which they call "Big-fish-little-pond-effect (BFLPE) is very specific to academic self-concept and has not been found for self-esteem, self-efficacy, and other non-academic aspects of self-concept measures. BFLPE in the educational context is the observation that a student will have a lower academic self concept in an academically selective school than in a non-selective school. They argue that what is critical for the BFPLE to occur is whether the wording of the items measuring a certain construct invoke social comparison and do not include a specific criterion or an absolute frame of reference (i.e., self-efficacy).
- Incidentally, multiple group confirmatory factor analysis investigating the cross-cultural validity of the scales in Table 1 reported by Vieluf, Lee & Kyllonen (2009) identified two of the scales that show the phenomenon to a greater (ATSCHL) and to a lesser (MEMOR) extent as having not even metric invariance across countries.

PREVIOUS ANALYSES OF THE PHENOMENON

10. Cross-cultural differences in response styles are considered to be a serious source of bias in international surveys using Likert items. Several types of response styles have been described (e.g. Greenleaf, 1992, Clarke, 2000; Johnson & al., 2005, Thomas & al., 2008). All of them can make it difficult to distinguish authentic cultural differences from "stylistic" biases in respondent behaviour (Van de Vijver & Poortinga, 1997; van Hemert, Poortinga & van de Vijver, 2007).

    Extreme response style refers to the case of persons who tend to select the answer categories at the extreme sides of the Likert scale (e.g. , Strongly agree / Strongly Disagree; or Never / Always) rather than the intermediate responses.

    Intermediate response style refers, by contrast, to the case of people who tend to avoid the extreme answer categories, and to make most frequently use of the middle ones. A special case of middle response style is **Central response style** (also called Midpoint response style), where people tend to prefer the neutral category (e.g. Neither agree nor disagree), if one is provided.

    Acquiescent response style refers to people who tend to agree with most or all statements, independently of their content. Such respondents give 'positive' answers (e.g. Agree / Strongly Agree; or Often / Always) even to sets of items with opposite meanings (*blind agreement*, sometimes called "Yeah-saying").

    Disagreement response style refers to the opposite: the case of individuals who select negative answers for most of the items (e.g. Disagree / Strongly disagree; Never / Almost never), regardless of content.

    Previous analyses of the attitudinal constructs used in PISA2006 (Buckley, 2008) and PISA2003 (see second part of Appendix A) have provided evidence of cross-national differences in these response styles.

11. Proposed explanations of differences in response styles include the assumption of frame-of-reference effects whereby responses to attitude (or other) questions might differ systematically depending on which frame of reference (either across countries or across sub-groups within countries) is applied. These frames-of-reference include so-called "cultural macro values", the Big-Fish-Little Pond Effect" and social desirability. After a brief definition, results of analyses exploring the applicability of these explanations using PISA data are provided.

    (i). **Cultural macro values**. Several authors have provided empirical evidence which suggests that countries and cultures differ systematically with respect to their value-orientation. Thus, Inglehart et al. (2004) show

that countries can be classified on a two-way axis, one ranging from traditional values to secular rational values and the other from survival values to self-expression values. Schwartz (2006) divides the circular structure of human values into four quadrants with self-enhancement and self-transcendence in opposite quadrants and conservation and openness to change in opposite quadrants. Hofstede (2001) compares countries in terms of power distance, individualism, masculinity, uncertainty avoidance and long-term orientation. Also prevalent in the literature (Triandis et al., 1988) is the distinction between countries that are more interdependent/collectivist and those that are more independent and individualistic. Positive self-concept and self-enhancement are encouraged in individualistic cultures, while in collectivist cultures, children are taught to be modest which means that they must avoid showing their superiority. This is clearly a country-level factor - although a cultural bias towards "humility" could also exist at within-country level, affecting immigrant students or minority groups belonging to communities that are predominantly "collectivist", in societies where the majority of the population has an "individualistic" culture. If this explanation was applicable, one would observe lower self-concept mean scores in Asian countries, and perhaps in Islamic countries compared to other countries with predominantly individualistic cultures. At the within-country level, African American or indigenous minorities or immigrant students from countries with "collectivist" cultures can also be expected to report lower levels of self-concept than other students in Western societies. It might also explain the greater preparedness of respondents in more independent cultures to choose extreme response behaviour compared with a greater propensity to choose more moderate response options in cultures with a more interdependent outlook.

(ii) **Big Fish-Little Pond Effect** (BFLPE). A person's self-perception tends to be relative rather than absolute. Students will tend to compare their competency with that of other students in their environment, so a student with average competency will consider him/herself "good in maths" or good in sciences" if he/she is enrolled in a poorly achieving class or school, while the same student would report low self-concept if enrolled in a high-achieving class or school. This clearly suppressive effect could probably explain why very little between-school or between-class variance in self-concept scale scores is usually observed in most participating countries.

The BFLP hypothesis could also explain why the behaviour of the self-efficacy scale is so different from that of the self-concept scale. Self-efficacy items ask the students to what extent they feel confident that they will succeed in doing specific science tasks – thus the scale is much less likely than the self-concept scale to elicit "relative" rather than "absolute" perceptions. The BFLP factor could have suppressive effects on the *country-level* correlations between self-concept and performance for the following reasons:

- The between-school variance is much lower in some comprehensive school systems (e.g. Scandinavian countries) than in tracked systems (e.g. German-speaking countries). One may expect that the suppressive effect of BFLP is larger in the latter than in the former.
- The way teachers assess the proficiency of individual students can vary from country to country and either increase or moderate the classroom or school aggregation effects. Teacher marks are probably the most critical information used by students to construct their self-concept in a given domain. In school systems where teachers tend to refer to classroom norms when grading the students' work, their marks will probably increase the students' propensity to compare themselves with their classmates. By contrast, in school systems where the marks are usually based on curricular standards that are common to all schools, the information provided to students will be more consistent with their absolute (rather than relative) proficiency, thus decreasing the BFLP effect on their self-concept.
- Even in cases when the teacher marks are referred to curricular standards, the standards themselves may vary across countries (or within countries across different study programmes). Depending on the extent to which the teacher marks reflect the "true" position of the students in relation to these standards, students with a same level of proficiency may consider themselves as "poorer" when they attend a more demanding curriculum than when they attend a less demanding curriculum.

(iii) **Social desirability** according to Holtgraves' (2004, p. 161) "refers to a tendency to respond in self-report items in a manner that makes the respondent look good rather than to respond in an accurate and truthful manner". This term contains possible interpretations of *meaningfully acquiescent behaviour* whereby some people tend (either consciously or unconsciously) towards selecting the responses that will reflect those attitudes that are considered as most 'acceptable' in their society. In studies where the survey instrument contains no negative statements concerning the target construct, compliance cannot be distinguished from blind acquiescence, which may explain why the terms *acquiescence* and *social desirability* often seem to be used as near synonyms in the literature. While SD is frequently examined within countries, the above differences in cultural macro values might also lead to systematic differences in the propensity towards SD across countries. In addition, respondents' SD behaviour may be different in different cultures in that more collectivist cultures this would lead to the intermediate (or middle) response

option whereas in more individualistic countries such as USA, SD might trigger more extreme response options, to be seen as different. Several scales (e.g. the Edwards Social Desirability Scale, Edwards, 1957; the Balanced Inventory of Desirable Responding (BIDR), Paulhus, 1984; Marlowe-Crowne Social Desirability Scale (MCDS), Crowne & Marlowe, 1960) have been developed to measure SD in order to control for this tendency in subsequent analyses (Diekmann, 2003; Seitz, 1977), although a number of authors have questioned the validity and reliability of these scales (Leite & Beretvas, 2005; Moorman & Podsakoff, 1992; Paulhus & Van Selst, 1990; Paulhus, 1991).

12. One of the currently proposed techniques to measure SD is the overclaiming technique as a measure of self-enhancement that reduces concerns about the legitimacy of subject responses. It is accomplished by asking test takers to identify things that they recognize. The catch is that the scale contains a percent of illegitimate (i.e., non-existent) people, places, or things (Paulhus et al., 2003). A measure of accuracy (indicating knowledge of legitimate items) acts as a self-report measure of knowledge and is moderately to strongly correlated with measures of cognitive ability. A measure of bias demonstrates moderate correlations with other measures of self-enhancement. Even when participants are warned about the basic nature of the measure, it continues to operate as intended and has demonstrated initial utility in detecting when people are intentionally distorting their responses (i.e., when directed to do so in a laboratory). Some other approaches including unlikely virtues (Hough et al., 1990) or idiosyncratic items types (Kuncel, 2010) might also be useful for detecting social desirability. In a review of research into questionnaire design Lietz (in press) found evidence that question lead-ins such as "What do you believe other people think about.." for attitudes and "Do you happen to know…" or "Have you had time…" for knowledge questions might reduce the propensity of respondents to give socially desirable answers. All of these techniques, however, would have to be tested for their cross-cultural applicability in PISA.

Analyses exploring the cultural macro values explanation

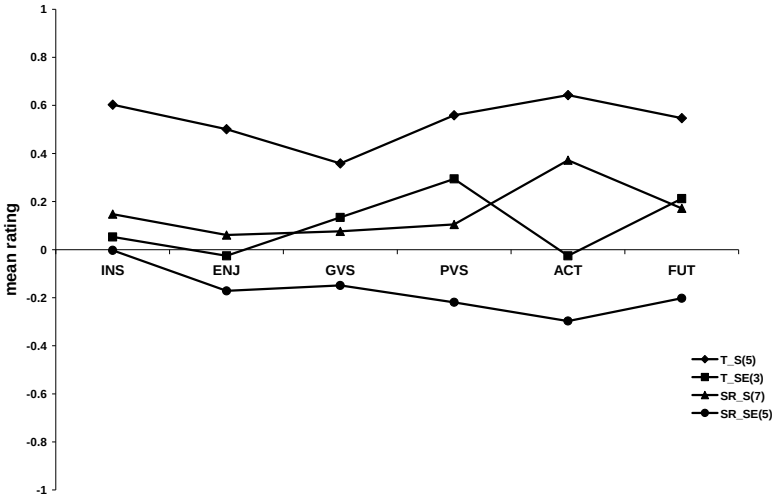
13. Using data from PISA2006, Ainley and Ainley (under review) explored the hypothesis that if macro cultural values were related to attitudes towards science different profiles across the constructs measuring attitudes towards science (i.e. general interest in science, enjoyment of science, general value of science, personal value of science, science-related activities and future oriented motivation to learn science) would emerge for groups of countries with different value orientations. Therefore, they selected the following four groups of countries which differed in terms of traditional versus secular-rational values on the one axis and survival versus self-expression values on the other axis (see Inglehardt et al. 2004).

    - Brazil, Chile, Columbia, Mexico and Turkey from the traditional/survival quadrant,
    - Argentina, Ireland and USA from the traditional/self-expression quadrant,
    - Bulgaria, Czech Republic, Estonia, Latvia, Lithuania, Russia and South Korea from the secular-rational/survival quadrant, and
    - Denmark, Germany, Japan, Norway and Sweden from the secular-rational /self-expression quadrant.

    Using each country's weighted mean scale score (WLE) an overall mean was calculated for each of the interest in science variables for the four cultural contrast groups (see Figure 2).

    The five countries representing the traditional/survival values had the highest mean ratings on each of the interest in science attitude whereas the countries representing the secular-rational/self-expression values showed exactly the opposite pattern with the lowest mean scores on all constructs. The contrasting patterns of mean ratings for these groups of countries confirmed the prediction that responses to the science attitude constructs reflected the degree to which science and technology were embedded within the cultural fabric. It was argued that the comparison was between a set of countries where science and technology were an assumed part of everyday life and figured prominently in students' perceptions of their future careers and lifestyles and countries where science and technology have only more recently become possible career and lifestyle opportunities for young people. To some extent the differences in these profiles were seen to contrast science as a cultural known and science as a developing cultural opportunity. In addition, the high ratings for the traditional/survival countries did not simply represent response bias as there was clear discrimination between ratings on some of the variables. For example, general value and personal value of science scales were subsets of items within the same question and the mean rating for general value of science was substantially lower than the mean rating for personal value of science. The order of science achievement for the groups of countries was the reverse of their level on the interest profiles. The secular-rational/self-expression countries had the highest average science achievement (521.46) whereas the

traditional/survival values countries had the lowest (402.49) with the secular-rational/survival (489.50) and the traditional/self-expression (476.94) in between.



Notes: Scales: INS = General interest in science; ENJ = Enjoyment of science; GVS = General value of science; PVS = Personal value of science; ACT = Science-related activities; FUT = Future-oriented motivation to learn science. Groups: T_S = Traditional/survival (Brazil, Chile, Columbia, Mexico, Turkey); T_SE = Traditional/self-expression (Argentina, Ireland, USA); SR_S = Secular-rational/survival (Bulgaria, Czech Republic, Estonia, Latvia, Lithuania, Russia, South Korea; SR_SE =Secular-rational /self-expression (Denmark, Germany, Japan, Norway, Sweden).

14. Van de gaer et al. (2009) applied multilevel analyses to examine the unexpected relationship between science achievement and science self-concept using PISA 2006 data. Results showed that educational standards and curricula that were measured by the quality of educational resources and the educational index were higher in higher performing countries and that higher performing countries showed lower self-concept scores. Moreover, the results showed that the East-Asian countries contributed significantly to the negative relationship between self-concept and achievement as these countries were among the highest scoring countries in PISA but are also among the lowest scoring in self-concept. In the literature, the tendency to score around the midpoint of the Likert scale on items tapping positive emotions among East-Asian countries is called modesty bias (Heine et al., 2001, 2002). The findings confirmed previous research because after controlling for country mean achievement, SES, and educational standards East-Asian countries still showed lower levels of self-concept than the other countries. Japan and Korea among the East-Asian countries showed very low level of self-concept. More important, once this modesty bias was taken into account, the remaining variance between countries in self-concept could be explained by the educational index, a measure that represents not only the educational standard but also the educational attainment of a country.

Analyses exploring the big-fish-little-pond-effect explanation

15. In a three-level analysis of data from 26 of the 32 countries that participated in PISA 2000 (Marsh and Hau, 2003), individual student achievement (at level1), average school achievement (at level2) and country (level3) were employed as predictors of student academic self concept. Results showed the universal applicability of the BFLPE and a significant effect of country on self-concept. This was argued to be expected in that the suppressive effect of BLFP was larger in tracked systems with greater between-school variance than in countries with comprehensive school systems with lower between-school variance.

16. Maybe not exactly exploring the BFLPE, an analysis of PISA2006 data (ACER internal) demonstrated that the smaller correlations between self-concept in science and science performance between students in lower performing countries was mainly due to the smaller within country variance for self-concept and science achievement in the lower performing countries. Together, these variables explained 67 per cent of the variance in (see Table 2).

Table 2 Variance in the magnitude of between-student within-country level correlations between science performance and science self-concept predicted by the within-school variance of both variables and the dichotomous variable *OECD* vs partner country, PISA2006

N= 56 countries

Regression Model	Predictors	R^2
A	OECD (OECD=1; Partner=0)	0.407
B	Science proficiency within-school variance	0.559
C	Self-concept within-school variance	0.435
AB	OECD + Science proficiency within-school variance	0.694
AC	OECD + Self-concept within-school variance	0.578
BC	Science within-school variance + Self-concept within-school variance	0.668
ABC	OECD + Science within-school variance + Self-concept within-school variance	0.756

Note: Dependent variable is between-student within-country correlations between *Science proficiency estimates* and *Self-concept in Science;* range -0.23 in KGZ to 0.47 in ISL).

Analyses exploring the social desirability explanation

17. Several ways were explored to calculate indicators of social desirability (ACER internal) using PISA2006 data:

    Two cross-country indicators of SD:

    a) Number of items out of 51 attitude items that were "anomalous" in a country (i.e. the mean science score of the group of students who chose the most positive response category was lower than the mean score of the group of students who chose the less positive response option. Of the 29 OECD countries showed that only Mexico had more than one third (=17 items) anomalous items. Of the 27 partner countries, 15 countries had more than one third anomalous items.
    b) Cases when the student reported that he/she invested more effort in the PISA test than he/she would have invested "*if the marks from the test were counted as school marks*". Of the 29 OECD countries, 3 countries (Greece, Mexico, Turkey) had more than 15% of students reporting putting more effort into PISA than other school tests. Of the 27 partner countries, 16 countries had more than one 15% of students reporting putting more effort into PISA than other school tests.

    Seven additional **within-country SD indicators** were calculated

    TOPICS: As most students across the world indicated "not sure" as regards these science topics and there was a positive correlation with the science score, students who had indicated being certain re these topics were considered giving SD responses.
    SELFENH: Self-concept minus Self-efficacy: Positive values indicating SD.
    IMPSCIE: Students who reported no science courses this year but >2 hrs studying in regular science courses
    IMPHWK: Students who reported no science courses this year but >2 hrs science homework
    IMPLEARN: Students who reported no science courses this year but reported that they had learnt at school about science topics.
    MAXSCIAC: Sum of extreme positive responses (very often) to 4 science activities.
    INVEFFOR: Students who reported putting more effort into PISA than usual school tests.

    Correlations for each of these indicators with the first plausible value in science (PV1) were calculated. MAXSCIA and INVEFFOR had non-significant or positive correlations in OECD countries and mainly negative correlations in partner countries while the remaining indicators showed negative correlations in all countries. In addition, within-country factor analysis of these seven indicators showed widely varying factor structures across countries. In addition, in only 19 countries could an "SD factor" be extracted in that this factor had non-negligible negative correlations with proficiency estimates. It also tended to correlate negatively with ESCS and home background variable.

    In conclusion, none of these indicators could be regarded as being indicators of SD that could be applied consistently across countries.

WAYS TO ADDRESS THE PHENOMENON

18. Three approaches, although intertwined, to address this phenomenon can be identified. First, the phenomenon can be considered to reflect genuine differences between countries whereby some countries or

cultural groups might have more positive attitudes regardless of the fact that the related actual context or outcome of interest is worse than in other countries. Second, it can be regarded as a measurement issue in that the measures or item types employed accentuate differences in response styles between countries and cultural groups. Therefore, the currently used measures are far from being optimal indicators of attitudes in cross-cultural research and measures that are less affected by different response styles have to be found. Third, it is considered that this phenomenon can be adjusted for through the application of different methods in the analyses of attitudinal data. Each of these approaches is discussed below.

THE PHENOMENON IS GENUINE

19. According to this approach, differences between countries in terms of attitudes and outlook are genuine in that, there are broad differences between countries on some basic cultural values and orientations (e.g. Dekker and Fisher, 2008; Hofstede, 2001; Inglehart, Basanez, Dıez-Medrano, Halman and Luijkx, 2004; Iyengar and Lepper, 1999; Schwartz and Sagiv, 1995). Thus, some countries or cultures are more positive, hopeful or optimistic than others for various reasons. Probably related to the idea of changing frames of reference (or changes in reference groups as proposed in some of the analyses described above), people in some countries tend to be more positive regardless of the fact that, actually, their country is worse off than others in terms of gross domestic product (GDP), human development index (HDI) or purchasing power parity (PPP). In a sense, it is argued that the correct moderator variable(s) have to be identified for the expected relationship to emerge. For the phenomenon under review, this means that the relationship between, for example, interest in mathematics and mathematics achievement depends on the level of GDP, HDI or value orientations.

20. Results of the analyses of PISA 2006 data by Ainley and Ainley (under review) reported above provide support for such an approach where attitudes to science differ between countries depending on a systematic difference in viewing science as generally beneficial and lifting the standard of living or as potentially dangerous and detrimental. Similarly, studies of employee satisfaction (e.g. Johnson, Kulesa, Cho and Shavitt, 2008) have repeatedly demonstrated higher level of satisfaction in less developed countries where the actual work conditions are worse than in countries in which employees show lower work satisfaction despite the fact that their actual conditions are comparatively pleasant. In the educational context, systematic differences may arise as a consequence of different proportions of students enrolled at different stages of schooling across countries. Finally, Heyneman-Loxley (1982) demonstrated that the effects of many school and teacher variables on academic achievement were systematically different for high and low income countries. This gave rise to the so-called Heyneman-Loxley effect which stated that school quality had a greater impact on student achievement in countries that were less developed economically than other countries.

21. The problem with this approach is that, ultimately, it implies that measures of attitudes are idiosyncratic and that, for example, only comparisons of attitudes within a country or a set of countries which are similar on the moderator variable either at one time point or shifts in level over time are warranted rather than comparisons of attitudes across countries. Such a stance could question the usefulness of measuring attitudes in a cross-national enterprise such as PISA. Alternatively, guidelines might be developed for secondary analysts of PISA data as to the appropriate comparison countries for different research questions.

THE PHENOMENON REQUIRES DIFFERENT WAYS OF MEASUREMENT

22. A second approach to the phenomenon argues that given that previous measures of attitudes triggered systematic differences in response styles across countries, ways of measuring attitudes that are less prone to such differences should be identified and trialled. In addition to reducing or eliminating response style effects, different items types might also capture the constructs of interest more accurately and more validly. The latter is suggested as some of the proposed item formats have been shown to increment over and above self-assessment in regression analyses predicting outcomes. The purpose of this section is to discuss alternative ways of measuring attitudes and to provide some sample items for consideration.

23. To align the sample items with constructs that are especially relevant in PISA, where possible, examples that match constructs assessed in the context questionnaire for PISA2003 (see Table 1) are provided, the previous cycle in which mathematics was a major domain, though it is acknowledged that not all constructs will necessarily be included in PISA2012.

24. Non-cognitive characteristics can be assessed in many different ways: for example, self-assessments,

interviews, and behavioural observations. A commonly used classification of assessment methods is Block's (1971) LOTS system of organizing assessment techniques: L = life data (e.g., biodata, transcripts); O = observer data (e.g., behavioural observations, open-ended text); T = test data (situational judgment tests, conditional reasoning tests); S = self-report (standard self-report inventories).

25. However, we organize assessments in a two-by-two table, by source (self or other) and type (ratings or performance), as shown in Table 3. This is not an exhaustive list of all possible new assessment types, but represents those that might appropriately be implemented in PISA2012.

Table 3 Source × type organization of assessment methods

Self		Others	
Ratings	Performance	Ratings	Performance
Self-assessments: Forced Choice Bayesian Truth Serum Vignettes Biodata	Situational Judgement (SJT) Conditional Reasoning	Others-ratings Other SJT Other Biodata	Transcripts

26. Different assessments do not always give the same score on a construct. For example, self-ratings of cognitive ability correlate at r = 0.25 (average over 55 studies) with actual cognitive test performance (Mabe & West, 1982), although this can be increased with certain methodological procedures.
27. It is not necessarily the case that one method is better than the other. Sometimes, two measures independently predict outcome criteria, each adding variance to the other. For example, Bratko, Chamorro-Premuzic, and Saks (2006) found that both self-reported and peer-rated conscientiousness predicted school performance (controlling for intelligence) independent of one another. In another study, MacCann, Minsky, Ventura, and Roberts (2010) found self- and parent-reports (mostly mother-reports) on conscientiousness, respectively incremented grades by 7% and 16% (controlling for intelligence).

Self-assessments

28. Self assessments are the most widely used approach for capturing students' non-cognitive characteristics, certainly in PISA. Self-assessments usually ask individuals to describe themselves by answering a series of standardized questions. The answer format is mostly a Likert-type rating scale, but other formats may also be used (such as Yes-No or open answer). Typically, questions assessing the same construct are aggregated; this aggregated score serves as an indicator of the relevant non-cognitive attribute.
29. Many issues need to be taken into account when developing a psychometrically sound questionnaire, and there is a large literature on such issues (e.g., number of points on a scale, scale point labels, neutral point, alternative ordering; see Krosnick, Judd, & Wittenbrink, 2005). Respondents also vary in their use of the scale: for example, young males tend to use extreme answer categories (Austin, Deary, & Egan, 2006), as do Hispanics in the USA (Marin, Gamba, & Marin, 1992), and there are large cultural effects in response style (Harzing, 2006).
30. Respondents can also fake their responses to appear more attractive for a variety of reasons (e.g., Griffith, Chmielowski, & Yoshita, 2007; Viswesvaran & Ones, 1999; Ziegler, MacCann, & Roberts, 2010), resulting in decreased validity (Pauls & Crost, 2005).
31. Several promising methods for collecting self-assessments, include using a multidimensional forced choice format (Stark, Chernyshenko, & Drasgow, 2005), using one's estimates of how others will respond to help control for faking (Prelac, 2004), and using vignettes to anchor the self-assessment (King, Murray, Salomon, & Tandon, 2004). These methods are discussed in the sections that follow.

Self-assessment: Bayesian Truth Serum

32. The Bayesian Truth Serum (BTS) approach is a method that has been proposed to reduce faking, and other possible confounding response style effects (Prelec, 2004). The technique provides incentives to respond honestly for respondents answering multiple-choice questions about their personal characteristics or opinions. The method requires the respondent to provide not only a personal answer to every question, but also to estimate in percentage terms how other respondents will answer that same question.
33. The method rests on an implication of Bayesian reasoning about population frequencies: Individuals with a particular characteristic should give higher estimates of the proportion of individuals in the population sharing the same characteristic, because possession of the characteristic is a positive signal about its overall

frequency. For example, if the respondent claims that they have high mathematics self-efficacy, then they should estimate a higher percentage of people saying that they also have high mathematics self-efficacy, than if they claim they do not have high mathematics self-efficacy. The BTS scoring method assigns high scores to answers whose actual frequency is greater than their predicted frequency.

34. There is some experimental evidence supporting these claims. For example, in one study, simulating lying about gender by changing gender in the data file for some respondents (but keeping other responses intact), showed reduced BTS scores for those respondents. In another study, BTS-based information scoring was more accurate than "majority rule" scoring in correctly identifying state capitals (Prelec, 2006).
35. Given the high validity of mathematics self-efficacy, it might be expedient to trial this approach for this construct, in the following way (adding this to the typical instruction related to self):

How confident do you think other students your same age in <your country> would feel about having to do the following Mathematics tasks?

	Not at all confident	**Not very confident**	**Confident**	**Very confident**
a. Using a <train timetable>, how long it would take to get from one place to another.	☐	☐	☐	☐
b. Calculating how much cheaper a TV would be after a 30 percent discount.	☐	☐	☐	☐
c. Calculating how many square metres of tiles you need to cover a floor.	☐	☐	☐	☐
d. Understanding graphs presented in newspapers.	☐	☐	☐	☐
e. Solving an equation like $3x + 5 = 17$.	☐	☐	☐	☐
f. Finding the actual distance between two places on a map with a 1:10,000 scale.	☐	☐	☐	☐
g. Solving an equation like $2(x+3) = (x + 3)(x - 3)$.	☐	☐	☐	☐
h. Calculating the petrol consumption rate of a car.	☐	☐	☐	☐

Self-assessment: Forced Choice

36. Another approach to reducing faking and other response styles is to change the format in which self-assessment statements are presented. Instead of presenting statements individually and asking examinees to indicate their levels of agreement, groups of statements can be organized into blocks. These are then administered to examinees with instructions telling them to choose the statement(s) that are most and/or least descriptive of them. Variations of this theme currently dominate the literature, as researchers explore formats ranging from pairwise preferences to tetrads (e.g., Christiansen, Burns, & Montgomery, 2005; Heggestad, Morrison, Reeve, & McCloy, 2006; Stark et al., 2005).
37. A common feature of forced choice tests is that the statements composing each block are matched in terms of social desirability so that respondents have a difficult time discerning which answers are better, making faking less effective. A variety of IRT models have been developed for scoring these items (see Stark, Chernyshenko, & Drasgow, 2010).
38. An example item that seems particularly relevant to PISA follows. It is likely that one of the problems previously encountered with an attempt to distinguish between co-operative and competitive learning rested in the fact that both were socially desirable. A forced-choice approach may ameliorate this problem.

Which of the following statements is most like you?

I like to work with other students. ☐
I like to try to be better than other students. ☐

39. The reduction of a usual four or more point Likert scale to two options (disagree vs agree) is also sometimes considered a forced-choice format in that it requires respondents to make a decision. Walker (2007) administered in the field trial of PISA 2006 the same items measuring enjoyment of science to half of the students using a 4 point Likert scale (i.e., 'strongly disagree', 'disagree', 'agree', 'strongly agree') and a dichotomous format (i.e., 'disagree' versus 'agree') to the other half of the student. He compared the means on items between countries using the different item formats (including a transformation of the 4-point Likert scale to a 2-point scale) and concluded

that there were only slight differences in the means and the correlations with achievement remained consistent.

Self-Assessment: Anchoring Vignettes

40. Anchoring vignettes ask respondents for self-assessments of the concept being measured along with assessments, on the same scale, of each of several hypothetical individuals described in anchoring vignettes (King et al., 2004). Since the actual levels for the people in the vignettes are invariant over respondents, the only reason answers to the vignettes will differ over respondents is interpersonal incomparability. The technique provides sufficient information for statistical models that are designed to correct the self-assessments.
41. The technique has gained momentum in policy research, and appears to address problems concerning cultural comparability. Buckley (2008) warns, however, of some important factors that need to be considered in item design (e.g., randomization of the vignettes, not always having the self-assessment come first). Its use, thus far, in education has been limited (see however, Buckley, 2008).
42. Many of the attitudinal constructs could be constructed so as to include anchoring vignettes. An example assessing teacher support in the mathematics classroom follows.

	Not very interested	**Not interested**	**Neutral**	**Inter-ested**	**Very Interested**
How interested is your mathematics teacher in getting students to work hard?	☐	☐	☐	☐	☐
a. Mr. <name> regularly sets mathematics homework but does not get the marks back on time for review before examinations. He encourages his students to pursue a career in mathematics, but does not always know the answers to questions. He is often late to class. How interested is Ms. <name> in getting his students to work hard?	☐	☐	☐	☐	☐
b. Ms. <name> sets mathematics homework only once a week but always gets the answers back on time for review before examinations. She encourages her students to pursue a career in mathematics and always know the answers to questions. She always arrives to class five minutes early. How interested is Ms. <name> in getting her students to work hard?	☐	☐	☐	☐	☐
c. Mr. <name> sets mathematics homework only once a week and yet does not get the marks back on time for review before examinations. He shows no interest in getting his students to pursue a career in mathematics and seldom knows the answers to questions. He is often late to class. How interested is Mr. <name> in getting his students to work hard?	☐	☐	☐	☐	☐

43. One issue with the use of vignettes as outlined here is the amount of development time required, particularly for cross-cultural use. As the three-year assessment cycle in PISA limits the time available for the development of new items to about six months, it is questionable whether vignettes that are comparable across the many countries and cultures participating in PISA can be developed. Previous studies that have used vignettes have done so in only two or three countries (e.g. King et al. 2004 in China and Mexico). Therefore, if the development of vignettes is considered useful, consideration needs to be given to developing these across two cycles whereby the development would be done during one cycle and the field-trial and use in the main study in the following cycle.

Self-Assessment: Biodata

44. Biographical data (biodata) are typically obtained by asking standardized questions about individuals' past behaviours, activities, or experiences. Respondents are given multiple-choice answer options or are requested to answer in an open format (e.g., frequency). Measures of biodata have been found to be incrementally valid beyond SAT and the Big Five in predicting student performance (e.g., Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004).

45. Obviously, biodata can be faked but there are several ways to minimize faking (e.g., Schmitt, Oswald, Kim, Gillespie, & Ramsay, 2003). Asking students to verify with details, for example, can minimize faking (Schmitt & Kunce, 2002).
46. PISA has traditionally implemented a variety of biodata-like questions. It is through this component – verifying responses – that we propose here a novel introduction. For example, consider these set of questions asked about computer use in PISA cycles in the past:

In your home, do you have:

A computer you can use for school work

Yes ☐
No ☐

If yes, what is the brand of the computer? _____

Educational software

Yes ☐
No ☐

If yes, what is the name(s) of the software? _____

A link to the Internet

Yes ☐
No ☐

If yes, what is the name of the internet provider? _____

47. Our proposal is not to do this manipulation with all such biodata questions, but with those deemed especially important for policy decisions or that have produced problematic findings in the past, possibly in concert with the rotation of student context questionnaires (see QEG[1002]2).
48. What has to be kept in mind in this context that frequency reports for autobiographical data for observable behaviours (e.g. absenteeism, being late for class, going to the library) have been shown to vary between cultures (e.g. between Chinese and Americans in Ji, Schwartz & Nisbett, 2000). Thus, people from more collectivist societies probably pay more attention to their social surroundings as well as to their own as well as others' needs and actions than people in more individualistic societies.

Situational Judgment Tests (SJTs)

49. A situational judgment test (SJTs) is one in which participants are asked how best to, or how they might typically deal with some situation. Situations can be described in words, audiotaped, or videotaped, and responses can be multiple choice, constructed response, ratings (how good would this response be?), and so forth (McDaniel, Morgesen, Finnegan, Campion, & Braverman, 2001).
50. SJTs may be developed to reflect more subtle and complex judgment processes than are possible with conventional tests. The methodology enables the measurement of many relevant attributes of individuals, including leadership, teamwork, achievement orientation, self-reliance, dependability, sociability, emotion management, and conscientiousness (e.g., Kyllonen & Lee, 2005; MacCann & Roberts, 2008; Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004; Wang, MacCann, Zhuang, Liu, & Roberts, 2009).
51. SJTs have been shown to predict many different criteria such as academic success (Lievens & Coestsier, 2002; Oswald et al., 2004), leadership (Legree, 1995), and managerial performance (Howard & Choi, 2000). Though applications in education have been more limited, there is emerging evidence they are effective predictors in educational domains (Lievens, Buyse, & Sackett, 2005; Oswald et al., 2004; Sternberg et al., 2000).
52. SJTs could be applied to assess many (but not all) of the constructs listed in Table 1. For example, consider the following as an indicator of co-operative versus competitive learning strategies:

You are taking part in a study group with your classmates in preparation for a particularly difficult exam in mathematics. As the first review session gets underway, it becomes clear that the other members of the group have not taken good notes and are not as familiar with the material as you are.

What are you likely to do in this situation?	**Very unlikely**	**Likely**	**Neither likely nor unlikely**	**Likely**	**Very likely**
a. Suggest that everyone read over the textbook in preparation for the next review session.	☐	☐	☐	☐	☐
b. Leave the group because you will be better off studying on your own.	☐	☐	☐	☐	☐
c. Offer to use your notes as the basis for the remaining review sessions.	☐	☐	☐	☐	☐
d. Ask the teacher to postpone the exam because many of the students are clearly not ready.	☐	☐	☐	☐	☐

Conditional Reasoning Tests (CRTs)

53. Conditional Reasoning Tests (CRTs) are multiple-choice tests consisting of items that look like logical reasoning items, but they really measure world-view, personality, biases, and motives (James, 1998; LeBreton, Barksdale, & Robin, 2007). Following a passage and a question, the CRT presents two or three logically incorrect alternatives, and two logically correct alternatives which reflect different world views. Participants are asked to state which of the alternatives seems to be most reasonable based on the information given in the text. Thus, respondents believe that they can solve a problem by reasoning about it, not realizing that there are two correct answers, and that their selection is guided by implicit assumptions underlying answer alternatives. To illustrate this idea consider the example from James (1998) below.

Studies of the stress-related causes of heart attacks led to the identification of the Type A personality. Type A persons are motivated to achieve, involved in their jobs, competitive to the point of being aggressive, and eager, wanting things completed quickly. Interestingly, these same characteristics are often used to describe the successful person. It would appear that people who wish to strive to be a success should consider that they will be increasing their risk for a heart attack.

Which of the following would most weaken the prediction that striving for success increases the likelihood of having a heart attack?

(A) Recent research has shown that it is aggressiveness and impatience, rather than achievement motivation and job involvement, that are that primary causes of high stress and heart attacks.
(B) Studies of the Type A personality are usually based on information obtained from interviews and questionnaires.
(C) Studies have shown that some people fear being successful.
(D) A number of non-ambitious people have heart attacks.

54. Alternatives (B) and (C) can be ruled out on logical grounds. Both (A) and (D) could be considered logically correct (or at least not incorrect), but reflecting different perspectives. Of the two responses, selecting (A) is taken as an indicator of achievement motivation, because to do so reflects a justification that achievement striving is a positive thing. A score reflecting an individual's level of achievement motivation is obtained by aggregating the answers to several of these kinds of items.
55. The CRT for achievement motivation has been shown to be unrelated to cognitive ability, reliable, and valid for predicting different behavioural manifestations of achievement (average r over 10 studies = .44) (James, 1998). A variant such as that given above could be implemented in PISA2012.

Other-Ratings

56. Other-ratings are assessments in which others (e.g., parents, teachers, colleagues, friends) rate individuals on various non-cognitive qualities. This method has a long history and countless studies have been conducted

that employed this methodology to gather information (e.g., Tupes & Christal, 1961/1992).

57. Other-ratings have an advantage over self-ratings in that they preclude socially desirable responding, although they do permit rating biases. Self- and others-ratings do not always agree (Oltmanns & Terkheimer, 2006), but others-ratings are often more predictive of outcomes than are self-ratings (MacCann et al., 2010; Wagerman & Funder, 2007).
58. Parents could be asked to rate their children on any of the constructs listed in Table 1, previously assessed as self-reports. For example, consider the following assessment of self-concept in mathematics:

How much do you disagree or agree with the following statements about your child and mathematics?

	Strongly disagree	**Disagree**	**Neither agree nor disagree**	**Agree**	**Strongly agree**
a. My child is just not good at mathematics.	☐	☐	☐	☐	☐
b. My child gets good <marks> in mathematics.	☐	☐	☐	☐	☐
c. My child learns mathematics quickly.	☐	☐	☐	☐	☐
d. I have always believed that mathematics is one of my child's best subjects.	☐	☐	☐	☐	☐
e. My child appears to understand even the most difficult mathematics concepts.	☐	☐	☐	☐	☐

59. It is also possible to use the approach to obtain more valid indicators of biodata, such as might be the case when assessing absenteeism in the Educational Career Questionnaire (QEG [1002]3[b]):

Did your child ever miss two or more consecutive months of <ISCED 2>

No, never ☐
Yes, once ☐
Yes, twice or more ☐

60. And recently, the approach has also been used with situational judgement test items, where a parent rates the likely course of action of their child (MacCann, Wang, Matthews, & Roberts, 2010). Thus, using the example above where an SJT was created to assess co-operative versus competitive learning, we would change this to the following form:

Your child is taking part in a study group with their classmates in preparation for a particularly difficult exam in mathematics. As the first review session gets underway, it becomes clear that the other members of the group have not taken good notes and are not as familiar with the material as your child is.

What is your child likely do in this situation?	**Very unlikely**	**Likely**	**Neither likely nor unlikely**	**Likely**	**Very likely**
a. Suggest that everyone read over the textbook in preparation for the next review session.	☐	☐	☐	☐	☐

Transcripts

61. Transcripts contain information on the courses students have taken, earned credits, grades, and grade-point average. As official records, transcript information can be taken as more accurate than self-reports. Transcript data can be standardized and used in validity studies. For example, the U.S. National Center for Educational Statistics supports an ongoing collection of transcripts (the NAEP High School Transcript Study, http://nces.ed.gov/nationsreportcard/hsts/), which classifies courses, computes grade-point average, and links resulting data to NAEP achievement scores (Shettle et al., 2007). The feasibility of collecting such data as part of the PISA data collection in many different countries with varying privacy and data protection laws would have to be explored.

THE PHENOMENON REQUIRES DIFFERENT WAYS OF ANALYSIS

62. Another consequence of differences in response styles are proposals regarding different ways of analysing data (Buckley, 2009; Harzing, 2006; Maij-de Meij, Kelderman and van der Flier; 2008, Rost, Carstensen and von Davier 1997; Smith 2003) including:
    a. Ipsative scoring (e.g. Schwartz, 2006);
    b. Standardization (e.g. Fischer, 2004);
    c. Mixture Item Response Theory models (e.g. Rost et al., 1997; van Rosmalen, et al., 2010);
    d. Multiple group latent class analyses (e.g. Moors, 2004);
    e. Structural equation modeling (e.g. Ziegler & Buehner, 2008);
    f. Factor analysis (e.g, Cheung & Rensvold, 2000; Lie & Turmo, 2005).

63. Ipsative scoring is defined as within person standardization, such that all respondents have the same mean response level. Such scoring of attitudinal scales has been successfully applied in the Schwartz (2006) values component of the European Social Survey. Likewise, at has been used in cross-national research to categorize teachers in terms of their approach as being either constructivist or explanatory.
64. An example of adjustments to correct for the phenomenon is the use of mixture IRT models (Rost et al., 1997, van Rosmalen, van Herk, Groenen, 2010) to group countries into categories that are similar with respect to response style. The goal of such analyses is to adjust responses country by country to enable the use of a common scale across countries.
65. Bolt and Johnson (2009) using data from a self-report measure of tobacco dependence demonstrate the applicability of a multidimensional item response theory model for the examination and control of response style effects in ordered rating scale data.
66. Other strategies include multiple group latent class analyses (e.g., Moors, 2004), structural equation modeling (e.g., Ziegler & Buehner, 2008), and factor analysis (e.g, Cheung & Rensvold, 2000; Lie & Turmo, 2005). What these techniques have in common is that they attempt to estimate and distinguish a factor or latent trait that measures a substantial content (i.e., the trait or construct to be measured) and a factor or latent trait that measures the response style. Both factors influence the scores on the items and the aim is to filter out the effect of the 'response-style factor'. For instance, Lie and Turmo (2005) undertook a factor analysis on a range of constructs such as learning strategies, motivation, self-concept, and school climate at the country level and extracted two components. The first factor showed high and positive loadings of all constructs and they interpreted this factor as the acquiescent response style whereas the second factor showed different loading for the different constructs. They found a strong negative association between countries' reading proficiencies in PISA 2000 and agreement response style suggesting that the lower performing countries showed the highest agreement response style. After correcting the country means for this agreement style, the correlations between the constructs and achievement changed dramatically. Many of the previously negative correlations of the constructs with performance switched to positive correlations (memorization strategies, cooperative learning, instrumental motivation, teacher support, student-teacher relations, self-concept) whereas other negative correlations were reduced in size (interest, attitudes to learning).
    Unlike Lie and Turmo (2005), Van de gaer (see Appendix A) using PISA 2003 data conducted analyses at the item level and estimated student-level attitudinal constructs while taking into account response bias. Results showed that the originally negative correlations at the between-country level between interest in mathematics, mathematics self-concept and mathematics performance became positive. These results suggest that a the extraction of an overarching response factor in this way might go some way to addressing the phenomenon.

RECOMMENDATIONS

67. Some further analyses are suggested to examine possible reasons for those attitudinal construct that have been shown to be less invariant across countries (e.g. attitudes towards school and memorisation strategies) and for which greater country effects have been reported (e.g. Vieluf et al. 2009a&b; Schulz, 2005).

68. To test new item types, particularly forced choice, biodata with verification, situational judgment, and variations of response scales in terms of number, labelling and direction of options in the field trial to examine the extent to which they might reduce some of the cross-cultural response style effects.

69. To suggest a development of vignettes to anchor ratings of self concept across countries that will extend over two cycles (i.e. 6 years) and to start this process in the current questionnaire development phase for PISA2012.

70. Explore the applicability of more recent techniques of measuring social desirability (e.g. overclaiming or idiosyncractic item types) during the field trial with the intention of examining the cross-cultural applicability of these measures and their ability to provide reliable and valid measures to account for SD in subsequent analyses.

71. To apply approaches such as those proposed by Rost et al. (1997), Maij-deMeij et al. (2008) or Bolt and Johnson (2009) to PISA 2003 data to examine the extent to which these models are able to adjust for response bias across countries and improve prediction of outcomes.

REFERENCES

Ainley, M. & Ainley, J. (under review). Interest in science: Part of the complex structure of student motivation in science. *Journal of Science Education*.

Atkinson, J.W. (1957). Motivational determinants of risk-taking behavior. *Psychological. Review, 64*, 359-373.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. Personality and Individual Differences, 40, 1235-1245.

Beretvas, S. N., Meyers, J. L., & Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement, 62*, 570-589.

Block, J. (1971). *Lives through time*. Berkeley, CA: Bancroft Books.

Bolt, D. M. & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement, 33*(5), 335-352.

Bratko, D., Chamorro-Pemuzic, T., & Saks, Z. (2006). Personality and school performance: Incremental validity of self- and peer-ratings over intelligence. Personality and Individual Differences, 41, 131-142.

Buckley, J. (2008). Survey context effects in anchoring vignettes. Retrieved February 10, 2010 from http://polmeth.wustl.edu/workingpapers.php.

Buckley, J. (2009). *Cross-national response styles in international educational assessments: Evidence from PISA 2006*. Last accessed 11/02/2010 https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31*, 187-212.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. Human Performance, 18, 267-307.

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation*. (pp. 75-146). San Francisco, CA: Freeman.

Fischer, R. (2004). Standardization to account for cross-cultural response bias: A classification of score adjustment procedures and review of research in JCCP. *Journal of Cross-Cultural Psychology, 35*(3), 263-282.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36*, 341-357.

Harzing, A.-W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management, 6*(2), 243-266.

Harzing, A-W. (2006). Response styles in cross-national survey research. *International Journal of Cross-Cultural Management, 6*, 243-266.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*, 9-24.

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*(4), 599-615.

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.

Heynman S.P. & Loxley, W.A. (1982). Influences on academic performance across high- and low-income countries: A re-analysis of IEA Data. *Sociology of Education, 55*, 13-21

Hofstede, G. (2001). *Cultures consequences: Comparing values, behaviours, institutions and organizations across nations* (2nd edn.). Thousand Oaks, CA: Sage Publications.

Holtgraves, T. (2004) Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, 30, 2, pp. 161-172.

Howard, A., & Choi, M. (2000). How do you assess a manager's decision-making abilities? The use of situational inventories. *International Journal of Selection and Assessment, 8*, 85-88.

Inglehart, R., Basanez, M., Dıez-Medrano, J., Halman, L. & Luijkx, R., (2004). *Human beliefs and values*. Ann Arbor: University of Michigan Press.

James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods, 1*, 131–163.

Ji, L.-J., Schwartz, N. & Nisbett, R.E. (2000). Culture, autobiographical memory, and behavioural frequency reports: Measurement issues in cross-cultural studies. *Personality and Social Psychology Bulletin, 26*, 585-593.

Johnson, T., Kulesa, P., Cho I.Y., Shavitt. S. (2008) *The Relation Between Culture and Response Styles: Evidence from 19 Countries.* Paper presented at the International Conference on Survey Methods in Multinational, Multiregional, and Multicultural Contexts (3MC), June, 2008, Berlin.

King, G., Murray, C. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. American Political Science Review ,98, 191-207.

Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2005). Attitude measurement. In D. Albarracin, B. T. Johnson, & M. P. Zanna (Eds.), Handbook of attitudes and attitude change. Mahwah, NJ: Erlbaum.

Kyllonen, P. C., & Lee, S. (2005). Assessing problem solving in context. In O. Wilhelm & R. W. Engle (Eds.), Handbook of understanding and measuring intelligence. (pp. 11-25). Thousand Oaks, CA: Sage.

LeBreton, J. M., Barksdale, C. D., & Robin, J. (2007). Measurement issues associated with Conditional Reasoning Tests: Indirect measurement and test faking. Journal of Applied Psychology, 92, 1-16.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor. *Intelligence, 21*, 247–266.

Leite, W. & Beretvas, N. (2005) Validation of scores on the Marlowe-Crowne Social Desirability Scale and the Balanced Inventory of Desirable Responding. *Educational and Psychological Measurement, 65*, 1, 140-154.

Lie, S., & Turmo, A. (2005). *Cross-country comparability of students' self-reports*. Report to the PISA Technical Advisory Group.

Lietz, P. (in press). Research into questionnaire design. A summary of the literature. *International Journal of Market Research*. To be published in Volume 52 Number 3.

Lievens, F., & Coestsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. International Journal of Selection and Assessment, 10, 245-257.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. Journal of Applied Psychology, 90, 442-452.

Mabe, P. A., & West, S. G. (1982). Validity of self evaluation of ability: A review and meta-analysis. Journal of Applied Psychology, 67, 280–296.

MacCann, C., & Roberts, R. D. (2008). Assessing emotional intelligence with situational judgment test paradigms: Theory and data. Emotion, 8, 540-551.

MacCann, C., Minsky, J., Ventura, M., & Roberts, R. D. (under review, 2010). Mother knows best: Comparing self- and parent-reported personality in predicting academic achievement. Personality and Social Psychology Bulletin.

MacCann, C., Wang, L., Matthews, G., & Roberts, R. D. (under review, 2010). Examining self-report versus other reports in a situational judgment test of emotional abilities. Emotion.

Maij-de Meij, A.M., Kelderman, H. & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement, 32*(8), 611-631.

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among Hispanics. Journal of Cross-Cultural Psychology, 23, 498-509.

Marsh, H. & Hau, K (2003). Big-fish-little-pond effect on academic self-concept: A cross-cultural (26 countries) test of the negative effect of academically selective schools. *American Psychologist, 58*(5), 364-376.

McDaniel, M. A., Morgesen, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. Journal of Applied Psychology, 86, 730-740.

Moorman, R.H. & Podsakoff, P.M. (1992) A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behavior research. *Journal of Occupational and Organizational Psychology*, 65, 131-149.

Moors, G. (2004). Facts and artefacts in the comparison of attitudes among ethnic minorities: A multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review, 20*(4), 303-320.

Oltmanns, T. F., & Turkheimer, E. (2006). Perceptions of self and others regarding pathological personality traits. In R.Krueger & J. Tackett (Eds.), Personality and psychopathology: Building bridges. New York: Guilford.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. Journal of Applied Psychology, 89, 187-207.

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). New York: Academic Press.

Paulhus, D.L. (1998). *Paulhus Deception Scales (PDS)*. Multi-Health Systems Inc., NY.

Paulhus, D.L. & Van Selst, M. (1990) The spheres of control scale: 10 years of research. *Personality and Individual Differences*, 10, 1029-1036.

Pauls, C. A., & Crost, N. W. (2005). Effects of different instructional sets on the construct validity of the NEO-PI-R. Personality and Individual Differences, 39, 297-308.

Prelec, D. (2004). A Bayesian truth serum for subjective data. Science, 306, 462-466.

Prelec, D. (2006). A Bayesian method for inducing truthful self-assessments. In P.C. Kyllonen (organizer), Solving the Faking Problem on Noncognitive Assessments (Invited symposium: Psychological Assessment and Evaluation). Athens, Greece: 26th International Congress of Applied Psychology.

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review 15*(3), 351–357.

Rost, J., Carstensen, D. & 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*. (Chapter 31, pp. 324-332). New York: Waxmann. Last accessed 11/02/2010 at: http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/c31.pdf

Schmitt, N., & Kunce, C. (2002). The effects of required elaboration of answers to biodata questions. Personnel Psychology, 55, 569-587.

Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., & Ramsay, L. J. (2003). Impact of elaboration on socially desirable responding and the validity of biodata measures. Journal of Applied Psychology, 88, 979-988.

Schwartz, S. H. (2006). Value orientations: Measurement, antecedents and consequences across nations. In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), Measuring attitudes cross-nationally - lessons from the European Social Survey. London: Sage.

Schwartz, S. H. and Sagiv, L. (1995). Identifying culture-specifics in the content and structure of values. *Journal of Cross-cultural Psychology*, *26(1)*, 92-116.

Schulz, W. (2005). *Testing parameter invariance for questionnaire indices using confirmatory factor analysis and item response theory*. Paper presented at the Annual Meeting of the American Educational Research Association in San Francisco, 7-11 April, 2005.

Shettle, C., Roey, S., Mordica, J., Perkins, R., Nord, C., Teodorovic, J., Brown, J., Lyons, M., Averett, C., & Kastberg, D. (2007). America's High School Graduates: Results from the 2005 NAEP High School Transcript Study (NCES 2007-467). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Smith, T.W. (2003). Developing comparable questions in cross-national surveys. In J. Harkenss, F. Van de Vijver & P. Ph. Mohler. (Eds.), *Cross-cultural survey methods* (Chapter 5, pp. 69-92). Hoboken, N.J.: John Wiley & Sons.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: An application to the problem of faking in personality assessment. Applied Psychological Measurement, 29, 184 – 201.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (in press, 2010). Constructing Fake-Resistant Personality tests using Item Response Theory: High stakes personality testing with Multidimensional Pairwise Preferences. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *Faking in personality assessment: Knowns and unknowns.* New York: Oxford University Press.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., (2000). Practical intelligence in everyday life. New York: Cambridge University Press.

Triandis, H. C., Bontempo, R., Villareal, M.J., Asai, M. and Lucca, N. (1988). Individualism and collectivistm: Cross-cultural perspectives on self-in-group relationships. *Journal of Personality and Social Psychology, 54*(2), 323-338

Tupes, E. C., & Christal, R. E. (1961/1992). Recurrent personality factors based on trait ratings. *Journal of Personality, 60*, 225-251.

Van de gaer, E., Grisay, A., Schulz, W. & Gebhardt, E. (2009). The reference group effect: An explanation of the paradoxical relationship between achievement and self-concept across countries. Paper presented at the PISA research conference, Kiel.

Van de gaer, E.,& Han (2009). The relationship between achievement, self-concept and attitudes: A cross-country analysis. ACER internal unpublished paper.

van Hemert, D.A., Poortinga, Y.H. & van de Vijver, F.J.R. (2007). Emotion and culture: A meta-analysis. *Cognition and emotion, 21*(5), 913-943.

van de Vijver, F. J. R. & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*(1), 29-37.

Vieluf, S., Lee, J. & Kyllonen, P. (2009). *The cross-cultural validity of variables from the PISA2003 student questionnaire.* QEG(0910)5b.doc, QEG meeting Offenbach, Germany, 19-21 October 2009.

Vieluf, S., Lee, J. & Kyllonen, P. (2009a). *The predictive power of variables from the PISA2003 student questionnaire.* QEG(0910)5a.doc, QEG meeting Offenbach, Germany, 19-21 October 2009.

Vieluf, S., Lee, J. & Kyllonen, P. (2009b). *The cross-cultural validity of variables from the PISA2003 student questionnaire.* QEG(0910)5b.doc, QEG meeting Offenbach, Germany, 19-21 October 2009.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197-210.

Wagerman, S. A., & Funder, D. C. (2006). Acquaintance reports of personality and academic achievement: A case for conscientiousness. *Journal of Research in Personality, 41*, 221-229.

Walker, M. (2007). Ameliorating culturally based extreme response tendencies to attitude items. *Journal of Applied Measurement, 8*(3), 267-278.

Wang, L., MacCann, C., Zhuang, X., Liu, L., & Roberts, R. D. (2009). Assessing teamwork and collaboration in high school students: A multimethod approach. Canadian Journal of School Psychology, 24, 108-124.

Ziegler, M., & Buehner, M. (2008). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement OnlineFirst*, published on October 15, 2008 as doi:10.1177/0013164408324469.

Ziegler, M., MacCann, C., & Roberts, R. D. (Eds.) (in press, 2010). Faking in personality assessment: Knowns and unknowns. New York: Oxford University Press.

Notes:

- Most of the material in the section entitled “Previous analyses of the phenomenon” has been taken from communication, analyses and papers internal to ACER that were produced by Ray Adams, Eveline Gebhart and Eva Van de gaer. The extensive contributions to these efforts by Aletta Grisay are herewith gratefully acknowledged.
- Explanations of the response styles taken from the draft note for the TAG and adjudication meeting to be held in Melbourne 15-18 March, 2010.

APPENDIX A RESPONSE STYLE BIAS: AN ANSWER TO THE ATTITUDE-ACHIEVEMENT PARADOX?

By Eva Van de gaer, Australian Council *for* Educational Research

1. In these analyses, we hypothesize that a response style bias is at least partially responsible for the attitude-achievement paradox. In order to test this hypothesis, we assume that attitudinal indices that show the phenomenon described in this paper are not only determined by the 'real' attitudes of students but also by an overarching "superfactor" (Lie and Turmo, 2005) of response style bias.

MODEL AND METHOD

In a first step, we selected 11 countries which participated in PISA 2003. These are AUS, BRA, FIN, FRA, DEU, HKG, KOR, IDN, IRL, JPN, and TUN. We based the selection of these countries on the position they hold in Figure 1 in this paper.
Three groups of countries can be identified:

1) countries that show a high mean interest in mathematics but low performance. Examples are TUN, IDN, and BRA.
2) countries that show a low mean interest in mathematics but a high performance. Examples are KOR, JPN, HKG, and FIN. Note: HKG shows a higher mean interest in mathematics than the other countries in this group.
3) countries that show a mean interest in mathematics and a mean performance. Examples are AUS, FRA, DEU, and IRL.

From these 3 groups of countries, we selected a number of countries for our analyses.

2. In the second step, we selected a number of constructs to be included in our conceptual model. In the first analyses, we included the constructs INMAT, SCMAT, and MATHEFF in our model. Figure 2 presents the statistical model. We used Mplus to analyse the data. We performed a confirmatory factor analysis that involved the estimation of:

   a) the loadings of the items ST30Q01, ST30Q03, ST30Q04, and ST30Q06 on the latent variable 'INTMAT',
   b) the loadings of the items ST32Q02, ST32Q04, ST32Q06, ST32Q07, and ST32Q09 on the latent variable 'SCMATH',
   c) the loadings of the items ST31Q01, ST31Q02, ST31Q03, ST31Q04, ST31Q05, ST31Q06, ST31Q07, and ST31Q08 on the latent variable 'MATHEFF".
   d) the loadings of the items ST30Q01, ST30Q03, ST30Q04, ST30Q06, and, ST32Q02, ST32Q04, ST32Q06, ST32Q07, ST32Q09 on the latent variable 'RESPONSE BIAS' (we restricted the loadings of the items ST31Q01, ST31Q02, ST31Q03, ST31Q04, ST31Q05, ST31Q06, ST31Q07, and ST31Q08 – items loading on MATHEFF - to zero on the latent variable 'RESPONSE BIAS').
   e) variances, and covariances between the latent constructs with the restriction that the covariances between the latent variable 'RESPONSE BIAS' and the other latent variables are set to zero
   f) residual variances
   g) factor scores.

3. We hypothesize that the attitude items are indicators of not only content constructs but also a response bias construct. A response bias factor may be responsible for the negative cross-country correlation of math attitudes such as interest in math and math self-concept with math performance. We expect that after taken into account the response bias factor, the negative between country correlations will become non-significant or – at least – will reduce in size.

4. Because math efficacy did not show the paradoxical pattern (i.e., switch in sign in correlations within and between countries with math performance), we restricted the loading of the items, which are indicators of the latent construct 'math efficacy', on the latent variable 'response bias' to zero. In Appendix B, we include an example of the Mplus code that was used to estimate the model. All the items scores (except ST24Q01, ST24Q02, and ST32Q02) were inverted so that a higher item score corresponds to a higher attitude.

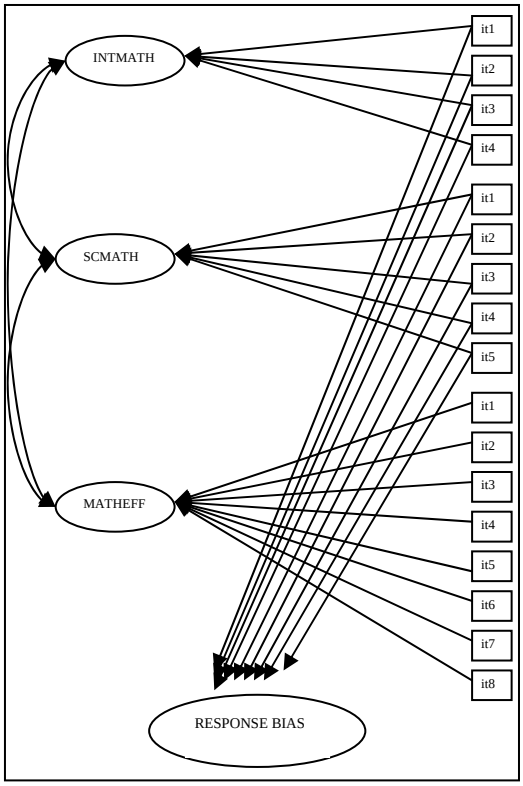


Figure 2. Conceptual Model

RESULTS

WITHIN COUNTRY CORRELATIONS

5. Tables 1 to 4 show the within country correlations between the constructs interest in math, math self-concept, math efficacy, and response bias with math performance. We also included the within country correlations between the original PISA 2003 constructs for comparison.
6. Overall, the results show that when we take into account response bias, the within country correlations between interest in math, math self-concept, math efficacy math performance become somewhat larger after taking into account response bias. For BRA, the within country correlation between intmat, scmath and pvmath even becomes positive. Only for IDN the within country correlation between intmat, scmath and pvmath remain negative (see Table 1 and 2). The results seem to suggest that we are taken some bias into account.
7. Table 3 shows that also for MATHEFF the correlations become a little bit larger after taken response bias into account. Although we restricted the loadings of the items belonging to MATHEFF and the response bias factor to zero, there still might be some indirect controlling for response bias through the correlations between MATHEFF and SCMATH and INTMAT.
8. Table 4 presents the within-country correlations between the response bias factor and pvmath. We found the strongest negative correlations for BRA and TUN indicating that in these two countries the lower performing students show higher response bias. In the other countries the correlations were close to zero meaning that there was no clear relationship between students' math performance and response bias. All the correlations were estimated using final students weights and BRR

.

Table 1. Correlations between interest in math and math performance in PISA 2003 for 11 countries

CORR F_INTMATH with PVMATH (estimated with Model in Figure 2)					CORR INTMAT with PVMATH (original correlations)			
CNT	CORR	SE	NU_cases*	N_cases	CORR	SE	NU_cases*	N_cases
AUS	0.319	0.013202	12460	234233	0.187	0.013096	12421	233576
BRA	0.110	0.026404	4344	1899581	-0.117	0.026543	4231	1849670
DEU	0.171	0.01872	4440	831613	0.118	0.019256	4424	828450
FIN	0.546	0.013159	5792	57831	0.334	0.016516	5697	56789
FRA	0.279	0.01759	4259	727263	0.222	0.018766	4237	722862
HKG	0.337	0.014028	4466	72259	0.303	0.014027	4465	72237
IDN	-0.068	0.029568	10723	1966952	-0.073	0.029435	10475	1927137
IRL	0.323	0.018282	3843	54207	0.196	0.020303	3826	53946
JPN	0.425	0.021557	4691	1235779	0.281	0.021742	4688	1234840
KOR	0.465	0.012654	5441	533182	0.393	0.013471	5438	532850
TUN	0.156	0.019166	4704	150315	0.102	0.019815	4670	149221

* NU_cases (Number of unweighted cases) is higher in the analyses taken into account response bias because Mplus imputes factor scores

Table 2. Correlations between math self-concept and math performance in PISA 2003 for 11 countries

CORR F_SCMATH with PVMATH (estimated with Model in Figure 2)					CORR SCMATH with PVMATH (original correlations)			
CNT	CORR	SE	NU_cases*	N_cases	CORR	SE	NU_cases*	N_cases
AUS	0.454	0.011498	12460	234233	0.409	0.012143	12403	233438
BRA	0.210	0.025748	4344	1899581	0.207	0.023423	4248	1856797
DEU	0.280	0.016829	4440	831613	0.267	0.016753	4416	826828
FIN	0.598	0.011416	5792	57831	0.575	0.012194	5766	57599
FRA	0.340	0.01779	4259	727263	0.320	0.018776	4212	718772
HKG	0.388	0.01657	4466	72259	0.348	0.017176	4463	72204
IDN	-0.053	0.028706	10723	1966952	-0.054	0.026678	10531	1933486
IRL	0.442	0.017015	3843	54207	0.376	0.019189	3823	53919
JPN	0.250	0.017709	4691	1235779	0.202	0.017823	4684	1233773
KOR	0.515	0.01226	5441	533182	0.462	0.013471	5435	532582
TUN	0.266	0.018569	4704	150315	0.275	0.017731	4653	148696

* NU_cases (Number of unweighted cases) is higher in the analyses taken into account response bias because Mplus imputes factor scores

Table 3. Correlations between math efficacy and math performance in PISA 2003 for 11 countries

CORR F_MATHEFF with PVMATH (estimated with Model in Figure 2)					CORR MATHEFF with PVMATH (original correlations)			
CNT	CORR	SE	NU_cases*	N_cases	CORR	SE	NU_cases*	N_cases
AUS	0.548	0.010492	12460	234233	0.522	0.011263	12418	233600
BRA	0.323	0.032489	4344	1899581	0.307	0.035265	4292	1875654
DEU	0.524	0.014342	4440	831613	0.508	0.015552	4419	827415
FIN	0.579	0.012507	5792	57831	0.524	0.01432	5697	56752
FRA	0.521	0.012723	4259	727263	0.504	0.014147	4210	718444
HKG	0.584	0.013311	4466	72259	0.556	0.015359	4464	72198
IDN	0.102	0.028024	10723	1966952	0.105	0.027113	10524	1935425

IRL	0.538	0.013696	3843	54207	0.529	0.013503	3828	53968
JPN	0.597	0.017945	4691	1235779	0.586	0.018845	4687	1234469
KOR	0.615	0.011547	5441	533182	0.576	0.012904	5436	532692
TUN	0.402	0.022736	4704	150315	0.370	0.024374	4516	144259

* NU_cases (Number of unweighted cases) is higher in the analyses taken into account response bias because Mplus imputes factor scores

Table 4. Correlations between response bias and math performance in PISA 2003 for 11 countries

CORR F_RESPBIAS with PVMATH				
CNT	CORR	SE	NU_cases	N_cases
AUS	-0.012	0.011715	12460	234233
BRA	-0.347	0.021035	4344	1899581
DEU	-0.086	0.016189	4440	831613
FIN	-0.011	0.016173	5792	57831
FRA	-0.027	0.019674	4259	727263
HKG	0.050	0.017177	4466	72259
IDN	0.060	0.017084	10723	1966952
IRL	0.001	0.021251	3843	54207
JPN	0.021	0.017566	4691	1235779
KOR	0.019	0.01469	5441	533182
TUN	-0.182	0.015666	4704	150315

BETWEEN COUNTRY CORRELATIONS

9. We aggregated the factor scores that were estimated using Mplus by country and correlated them with each of the five aggregated PV's for math. Then we calculated the mean of the five correlations.

10. When we compare Table 5 with Table 6, we conclude that the original negative correlations between interest in math, math self-concept and math performance become positive. The highest positive correlation was found between F_INTMATH and PVMATH. Surprisingly the correlation between MATHEFF and PVMATH became smaller across countries when taken into account response bias.

Table 5. Cross-country correlations between interest in math, math self-concept, math efficacy, and response bias and math performance for 11 countries.

Between country correlations (estimated with Model in Figure 2)					
	CNT_ PVMATH*	CNT_ F_INTMATH	CNT_ F_SCMATH	CNT_ F_MATHEFF	CNT_ F_RESPBIAS
CNT_PVMATH	1				
CNT_F_INTMATH	0.246	1			
CNT_F_SCMATH	0.196	0.918	1		
CNT_F_MATHEFF	0.015	0.766	0.814	1	
CNT_F_RESPBIAS	-0.316	0.654	0.439	0.4	1

*The correlations between each of the 5 PV's and the constructs were averaged

Table 6. Cross-country correlations between interest in math, math self-concept, math efficacy, and math performance for 11 countries.

Between country correlations (original correlations)				
	CNT_ PVMATH*	CNT_ F_INTMATH	CNT_ F_SCMATH	CNT_ F_MATHEFF
CNT_PVMATH	1			
CNT_INTMAT	-0.896	1		
CNT_SCMAT	-0.544	0.574	1	
CNT_MATHEFF	0.370	-0.111	0.408	1

*The correlations between each of the 5 PV's and the constructs were averaged

11. Table 7 and 8 present the mean factor scores for each country with and without controlling for response bias, respectively. FIN and IDN show the most negative country mean response bias scores whereas BRA, AUS, and IRL show the highest positive country mean response bias scores. We put the correlations in red font in Table 7 when they become more negative after taken into account response bias, in green font when they become more positive, and in black font when they remain the same compared to the original country mean scores. For example, the original country mean INTMAT scores for BRA, FIN, HKG, IDN, IRL, and TUN become much more negative (see red font in Table 7) after taken into account response bias whereas the original country mean scores for AUS, JPN, and KOR become much more positive. The country means for DEU, and FRA do not really change when taken response bias into account.

Table 7. Countries mean factor scores (estimated with Model in Figure 2) and pv.

CNT	F_INTMATH	F_SCMATH	F_MATHEFF	F_RESPBIAS	F_PVMATH1
AUS	71.61	34.90	68.24	23.97	524.08
BRA	-21.44	-37.67	-37.74	66.22	355.52
DEU	5.99	-3.75	-36.17	5.42	503.08
FIN	-171.31	-99.42	-109.60	-90.95	544.17
FRA	2.11	20.56	-68.43	0.11	511.47
HKG	-54.17	-70.59	-136.48	-0.71	549.43
IDN	-157.69	-86.98	-60.75	-25.86	360.09
IRL	-27.55	-53.25	-43.97	9.51	503.48
JPN	22.82	-1.60	23.70	3.24	533.64
KOR	77.25	74.91	71.48	-0.26	541.63
TUN	-1.06	6.35	-4.00	0.36	358.92

Note. All the factor scores were multiplied by 10000

Table 8. Countries mean original indices scores and pv.

CNT	CNT_INTMAT	CNT_SCMAT	CNT_MATHEFF
AUS	0.91	13.15	10.21
BRA	56.55	3.49	-37.86
DEU	4.49	14.89	15.21
FIN	-24.24	1.1	-15.42
FRA	4.48	-16.91	-0.74
HKG	22.44	-26.14	11
IDN	73.51	10.87	-30.63

IRL	-4.91	-2.91	-2.92
JPN	-38.65	-52.83	-52.69
KOR	-12.04	-35.12	-41.94
TUN	94.26	14.91	-29.47

Note. All the aggregated indices were multiplied by 100

12. Table 9 shows the correlations between the original attitude indices, the response bias factor score, and the attitude factor scores after taken into account response bias. The original and the 'new' attitude factor scores are negatively correlated. The response bias factor shows the highest correlation with both the original INTMAT as the 'new' FINTMAT. The correlation between the response bias factor and the original SCMAT and MATHEFF are very small and close to zero indicating that these constructs are less affected by response bias.

Table 9. Correlations between country mean original indices and country mean factor scores for 11 countries in PISA 2003 (estimated with Model in Figure 2).

	CNT_INTMAT	CNT_SCMAT	CNT_MATHEFF	CNT_FINTMATH	CNT_FSCMATH	CNT_FMATHEFF	CNT_FRESPBIAS
CNT_INTMAT	1						
CNT_SCMAT	.574	1					
CNT_MATHEFF	-.111	.408	1				
CNT_FINTMATH	-.223	-.282	.007	1			
CNT_FSCMATH	-.180	-.246	-.108	.918**	1		
CNT_FMATHEFF	-.144	-.120	-.357	.766**	.814**	1	
CNT_FRESPBIAS	.261	.002	-.037	.654*	.439	.400	1

RESPONSE STYLE 'SIMPLE' INDICES

Next to the Mplus analyses, we also calculated a number of simple indices that measure response bias on the PISA 2003 data. Table 10 gives an overview of the indices and clarifies how they were measured.

Table 10. An overview of response bias indices

Response Style		*Definition*	*Measurement*
ARS0_all_PRP	Acquiescence or agreement response style	The tendency to agree with items regardless of content.	Proportion of 'strongly agree' responses on all the 75 Likert type items in PISA 2003. Coding: strongly agree: score 1; agree/disagree/strongly disagree: score 0 (following definition of Buckley, 2009)
ARS1_all_PRP	Acquiescence or agreement response style	The tendency to agree with items regardless of content.	Proportion of 'strongly agree' responses on all the 75 Likert type items in PISA 2003. Coding: strongly agree: score 2, agree: score 1, disagree/strongly disagree: score 0

			(following definition of Cheung & Rensvold, 2001)
ARS0_5it_PRP	Acquiescence or agreement response style	The tendency to agree with items regardless of content.	Proportion of 'strongly agree' responses on 5 most heterogeneous Likert type items in PISA 2003. Coding: strongly agree: score 1; agree/disagree/strongly disagree: score 0 (following definition of Buckley, 2009)
ARS1_5it_PRP	Acquiescence or agreement response style	The tendency to agree with items regardless of content.	Proportion of 'strongly agree' responses on 5 most heterogeneous Likert type items in PISA 2003. Coding: strongly agree: score 2, agree: score 1, disagree/strongly disagree: score 0 (following definition of Cheung & Rensvold, 2001)
ARS_BD_PRP	Acquiescence or agreement response style	The tendency to agree with items regardless of content.	The acquiescence or agreement response style was measured using balanced data. Balanced data are items belonging to the same scale that have the same meaning but are negative and positively worded. We calculated the proportion of agreement on 5 pairs of balanced items. Coding for each pair of items: Strongly agree – strongly agree: score 3 Strongly agree – agree: score 2 agree – Strongly agree: score 2 agree – agree: score 1 other combinations: score 0.
DARS0_all_PRP	Disacquiescence of disagreement response style	The tendency to disagree with items regardless of content.	Proportion of 'strongly disagree' responses on all the 75 Likert type items in PISA 2003. Coding:

			strongly disagree: score 1; /disagree/agree/strongly disagree: score 0 (following definition of Buckley, 2009)
DARS1_all_PRP	Disacquiescence of disagreement response style	The tendency to disagree with items regardless of content.	Proportion of 'strongly disagree' responses on all the 75 Likert type items in PISA 2003. Coding: strongly disagree: score 2, disagree: score 1, agree/strongly agree: score 0 (following definition of Cheung & Rensvold, 2001)
DARS0_5it_PRP	Disacquiescence of disagreement response style	The tendency to disagree with items regardless of content.	Proportion of 'strongly agree' responses on 5 most heterogeneous Likert type items in PISA 2003. Coding: strongly disagree: score 1; /disagree/agree/strongly disagree: score 0 (following definition of Buckley, 2009)
DARS1_5it_PRP	Disacquiescence of disagreement response style	The tendency to disagree with items regardless of content.	Proportion of 'strongly agree' responses on 5 most heterogeneous Likert type items in PISA 2003. Coding: strongly disagree: score 2, disagree: score 1, agree/strongly agree: score 0 (following definition of Cheung & Rensvold, 2001) (following definition of Cheung & Rensvold, 2001)
ERS0_all_PRP	Extreme response style	The tendency to endorse the most extreme responses regardless of content.	Proportion of extreme responses 'strongly agree' or 'strongly disagree' on all the 75 Likert type items in PISA 2003. Coding: strongly agree/strongly disagree: score 1; disagree/agree: score 0

ERS0_5it_PRP	Extreme response style	The tendency to endorse the most extreme responses regardless of content.	Proportion of extreme responses 'strongly agree' or 'strongly disagree' on 5 most heterogeneous Likert type items in PISA 2003. Coding: strongly agree/strongly disagree: score 1; disagree/agree: score 0
NCR_PRP	Non contingent response style.	The tendency to respond to items carelessly, randomly, or non-purposefully	Average absolute difference between responses of pairs of items, where the items in each pair are maximally correlated.

13. In order to reduce the number of missing values on the response style indices, we calculated a proportion whenever less than 50% of the responses items were missing. When more than 50% of the responses of items were missing, a response style index was not calculated.

14. For the measurement of ARS0_all_PRP, ARS1_all_PRP, DARS0_all_PRP, DARS1_all_PRP, and ERS0_all_PRP, we used 75 Likert items. These are all the items belonging to the attitude indices in PISA 2003 (ATSCHL, STUREL, BELONG, INTMAT, INSTMOT, MATHEFF, ANXMAT, SCMAT, CSTRAT, ELAB, MEMOR, COMPLRN, COOPLRN, TEACHSUP, DISCLIM) except for the ICT-indices as the ICT questionnaire was a national option (and not all countries filled out this questionnaire). All items are Likert type bipolar items with four categories 'Strongly Agree', 'Agree', 'Disagree', and 'Strongly Disagree'. There are, however, some exceptions. The items belonging to the scale MATHEFF and DISCLIMA are also bipolar with four categories but they used different wordings 'Very confident', 'Confident', 'Not very confident', 'Not at all confident' and 'Every lesson', 'Most lessons', 'Some lessons', 'Never or hardly any lesson', respectively.

15. For the measurement of ARS0_5it_PRP, ARS1_5it_PRP, DARS0_5it_PRP, DARS1_5it_PRP, and ERS0_5it_PRP, we used a subset of five Likert type items (out of the 75 Likert items) that are heterogeneous[1]. These items are ST24Q01, ST26Q02, ST32Q02, ST37Q06, and ST37Q10. In a first step, we identified the attitude constructs that showed the lowest intercorrelations (ATSCHL, SCMATH, COOPLRN, STUREL, and COMPLRN) and, in a second step, we selected items that showed the lowest intercorrelations among each other. The average correlation between the 5 items was 0.32 with a minimum of 0.25 and a maximum of 0.62.

16. For the measurement of ARS_BD_PRP, we selected five pairs of items belonging to the same scale that were positively and negatively worded. These pairs are (ST24Q01, ST24Q03), (ST24Q02, ST24Q04), (ST27Q04, ST27Q03), (ST27Q02, ST27Q06) and (ST32Q02, ST32Q06).

17. For the measurement of NCR_PRP, we selected five pairs of items that were highly correlated. These are (ST30Q03, ST30Q04), (ST30Q02, ST30Q05), (ST31Q05, ST31Q07), (ST37Q05, ST37Q07), and (ST38Q06, ST38Q08). The average correlation was 0.66 ranging from 0.603 to 0.744.

18. In the literature, we found two other response styles that may be interesting. MPR or the tendency to respond towards the midpoint of the scale (this is actually equivalent to 1-ERS in PISA 2003) and NARS or the tendency to show greater acquiescence than disacquiescence (=ARS-DARS). We didn't include them because they are linear combinations from ERS, ARS, and DARS.

[1] In the literature it is recommended to calculate response styles using a heterogeneous set of items.

(Note: another response style that was mentioned by Baumgartner and Steenkamp (2001) is RR or the tendency to use a narrow or wide range of response categories around the mean response and is usually measured by the standard deviation of a person's responses across many heterogeneous items. We did not calculate RR.)

Countries mean math performance and mean response styles

19. Figures 3 to 5 present some examples of correlations between different response styles and math performance for *all* 41 countries in PISA 2003. Table 11 shows the actual cross-country correlations between math performance and all of the response styles for all 41 countries in PISA 2003.
20. Figures 3 and 4 show a positive and negative relationship between ARS and DARS with PV1MATH across countries, respectively. This means that the higher performing countries show fewer tendencies to agree with statements (ARS) but a higher tendency to disagree with statements (DARS) than the lower performing countries.
21. The between-country relationship between ERS and NCR with PV1MATH shown in Figures 5 and 6 are both negative but the correlation is much smaller in size compared to Figures 3 and 4. The results indicate the poor performing countries show a higher tendency to show extreme and non-contingent responses than the higher performing countries.

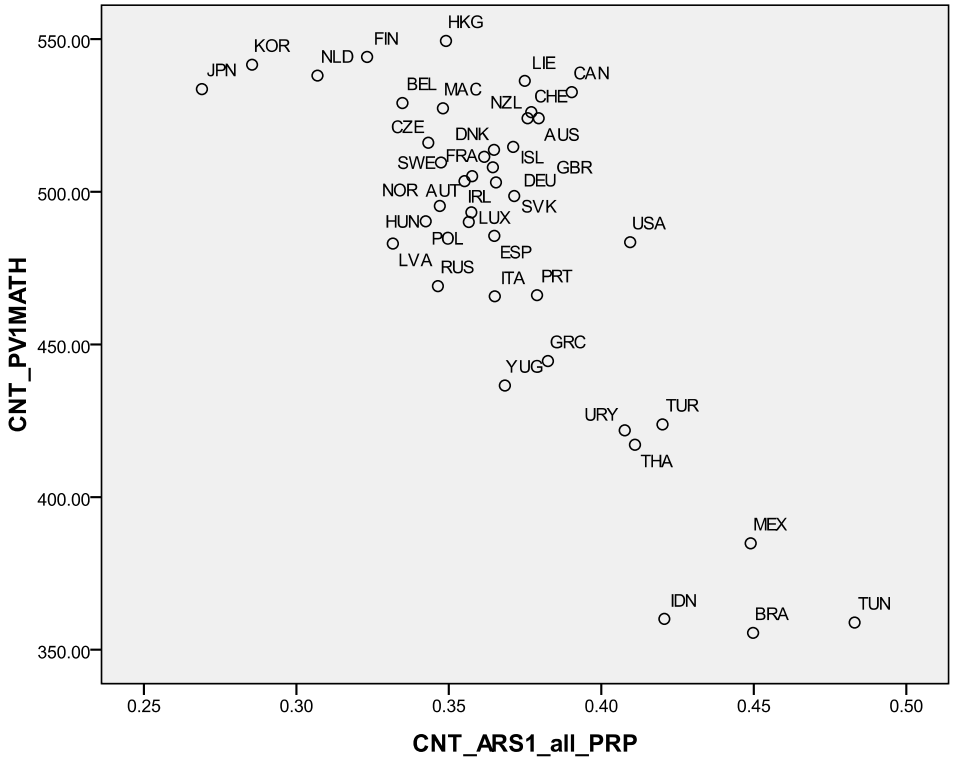


Figure 3. Country mean acquiescence response styles by country mean PV1MATH for all 41 countries in PISA 2003

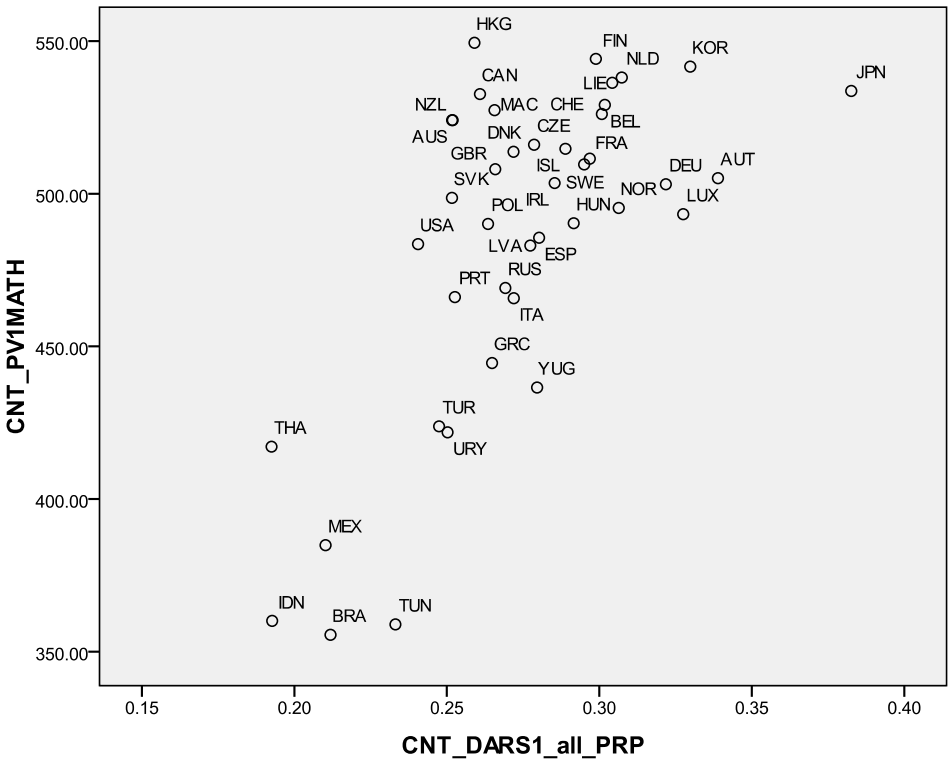


Figure 4. Country mean disacquiescence response style by country mean PV1MATH for all 41 countries in PISA 2003

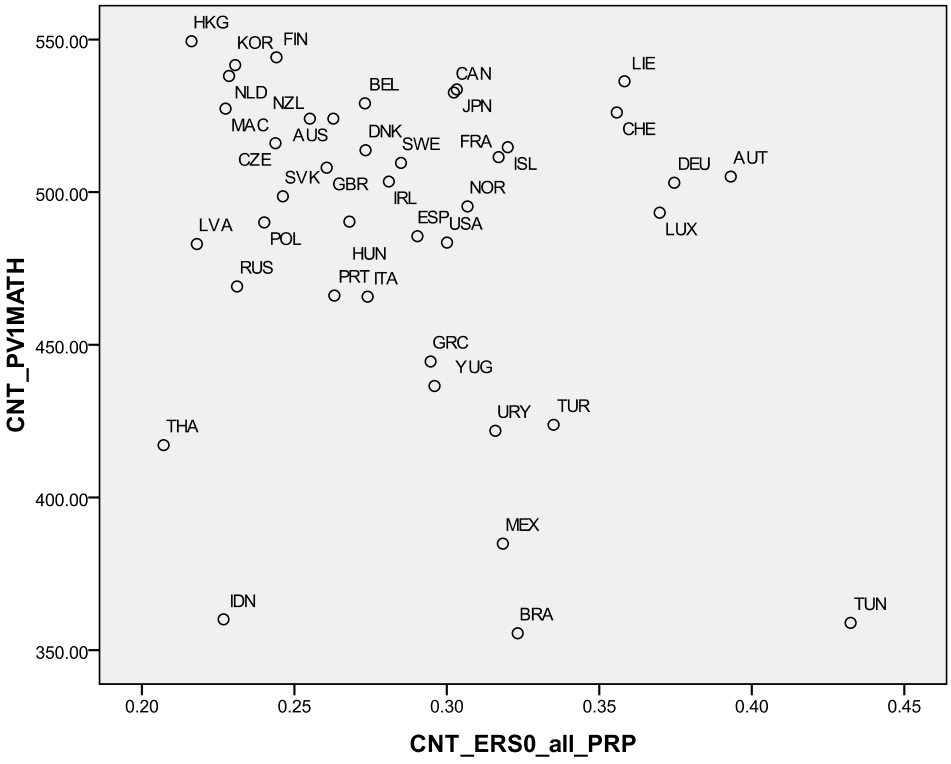


Figure 5. Country mean extreme response style by country mean PV1MATH for all 41 countries in PISA 2003

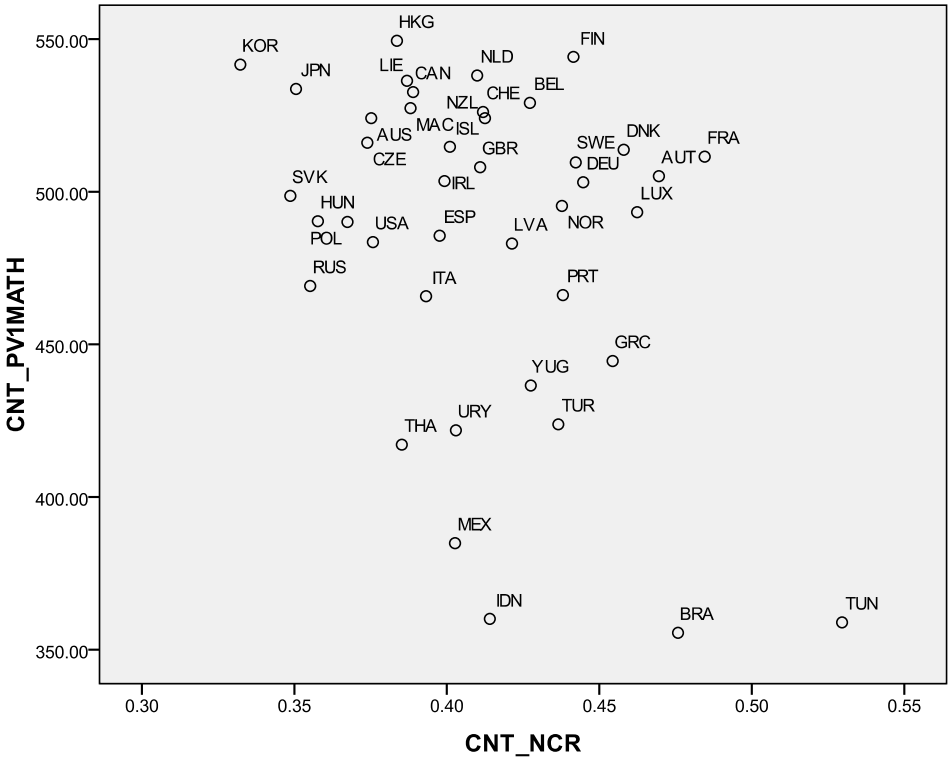


Figure 6. Country mean non-contingent response style by country mean PV1MATH for all 41 countries in PISA 2003

22. In Appendix C, Table 14 shows the country mean math performance in descending order and the country mean response style scores.

23. Next to the cross-country correlations between math performance and the response styles for all 41 countries in PISA 2003, Table 11 also shows the correlations between the different response styles. The results show that the different measurements of ARS are highly correlated (correlations ranging from .605 to .837) as well as the different measurements of DARS (correlations ranging from .594 to .875). As can be expected ERS is positively correlated both with ARS and DARS measures. NCR shows the largest correlation with ERS. The results suggest that using all 75 Likert items or only a subset of 5 heterogeneous items (or balanced data for ARS) to measure ARS, DARS, and ERS doesn't seem to matter too much.

Table 11. Cross-country correlations between math performance and response styles for all 41 countries in PISA 2003

	CNT_PVMATH	CNT_ARS0_all_PRP	CNT_ARS1_all_PRP	CNT_ARS0_5it_PRP	CNT_ARS1_5it_PRP	CNT_DARS0_all_PRP	CNT_DARS1_all_PRP	CNT_DARS0_5it_PRP	CNT_DARS1_5it_PRP	CNT_ERS0_all_PRP	CNT_ERS0_5it_PRP	CNT_ARS_BD_PRP	CNT_NCR
CNT_PVMATH	1.000												
CNT_ARS0_all_PRP	-.630	1.000											
CNT_ARS1_all_PRP	-.785	.861	1.000										
CNT_ARS0_5it_PRP	-.686	.871	.772	1.000									
CNT_ARS1_5it_PRP	-.724	.619	.791	.837	1.000								
CNT_DARS0_all_PRP	.424	.076	-.402	.023	-.425	1.000							
CNT_DARS1_all_PRP	.683	-.369	-.784	-.363	-.695	.875	1.000						
CNT_DARS0_5it_PRP	.232	.376	-.055	.170	-.330	.841	.594	1.000					
CNT_DARS1_5it_PRP	.573	-.165	-.542	-.419	-.837	.765	.804	.783	1.000				
CNT_ERS0_all_PRP	-.225	.803	.412	.674	.215	.655	.243	.787	.332	1.000			
CNT_ERS0_5it_PRP	-.305	.820	.477	.773	.342	.557	.142	.757	.227	.954	1.000		
CNT_ARS_BD_PRP	-.642	.527	.605	.708	.788	-.284	-.488	-.276	-.648	.230	.291	1.000	
CNT_NCR	-.363	.568	.425	.521	.300	.253	-.056	.471	.066	.582	.649	.049	1.000

Country mean attitude and mean response styles

24. Table 12 shows the correlations between country mean attitude and country mean response styles. We highlighted the attitude indices in yellow when they show a positive within-country but a negative between country correlations with achievement (cf. paradox). Positive high correlations (> .6) are shown in green font and negative high correlations (< -.6) are shown in red font. There seems to a pattern among the attitude indices that show the paradox. They all show positive correlations with ARS and negative with DARS, except for CNT_SCMATH (We suspect that this index is also affected by other type of bias such as the BFLPE). The indices that do not show the paradox (CNT_MATHEFF, CNT_BELONG, CNT_ANXMAT, and CNT_DISCLIM) do not seem to show a clear pattern in their correlations with response styles. This suggests that they are indeed less susceptible to response bias and that this may explain why they do not show the paradox.

Table 12. Cross-country correlations between math performance and attitude indices for all 41 countries in PISA 2003

	CNT_ INTMAT	CNT_ SCMAT	CNT_ MATHEFF	CNT_ ATSCHL	CNT_ STUREL	CNT_ BELONG	CNT_ INSTMOT
CNT_ARS0_all_PRP	.573	.525	.069	.595	.483	.389	.483
CNT_ARS1_all_PRP	.808	.574	-.053	.774	.716	.230	.784
CNT_ARS0_5it_PRP	.622	.225	-.275	.460	.466	.158	.450
CNT_ARS1_5it_PRP	.786	.151	-.387	.473	.621	-.069	.644
CNT_DARS0_all_PRP	-.598	-.115	.228	-.392	-.498	.434	-.635
CNT_DARS1_all_PRP	-.808	-.387	.185	-.676	-.715	.135	-.844
CNT_DARS0_5it_PRP	-.329	.296	.302	-.014	-.183	.585	-.276
CNT_DARS1_5it_PRP	-.698	.041	.408	-.318	-.532	.352	-.597
CNT_ERS0_all_PRP	.077	.329	.188	.217	.069	.551	-.014
CNT_ERS0_5it_PRP	.201	.340	.012	.296	.191	.482	.121
CNT_ARS_BD_PRP	.644	.078	-.292	.265	.410	-.223	.373
CNT_NCR	.294	.332	-.176	.453	.269	.382	.235

	CNT_ ANXMAT	CNT_ CSTRAT	CNT_ ELAB	CNT_ MEMOR	CNT_ COMPLRN	CNT_ COOPLRN	CNT_ TEACHSUP	CNT_ DISCLIM
CNT_ARS0_all_PRP	.243	.812	.594	.556	.603	.679	.378	-.271
CNT_ARS1_all_PRP	.354	.768	.865	.821	.801	.838	.705	-.361
CNT_ARS0_5it_PRP	.495	.652	.569	.498	.632	.563	.410	-.278
CNT_ARS1_5it_PRP	.588	.525	.728	.645	.722	.647	.660	-.312
CNT_DARS0_all_PRP	-.376	-.078	-.693	-.577	-.515	-.420	-.685	.162
CNT_DARS1_all_PRP	-.405	-.430	-.896	-.822	-.751	-.713	-.834	.311
CNT_DARS0_5it_PRP	-.446	.116	-.382	-.316	-.206	-.149	-.381	.032
CNT_DARS1_5it_PRP	-.586	-.250	-.678	-.595	-.582	-.510	-.668	.229
CNT_ERS0_all_PRP	-.040	.569	.037	.077	.150	.263	-.122	-.109
CNT_ERS0_5it_PRP	.040	.507	.131	.126	.286	.277	.027	-.163
CNT_ARS_BD_PRP	.542	.452	.585	.499	.568	.405	.480	-.028
CNT_NCR	-.008	.449	.252	.180	.214	.425	.065	-.466

COMPARING THE RESPONSE STYLE FACTOR ESTIMATED WITH MODEL IN FIGURE 2 WITH THE 'SIMPLE' MEASURES OF RESPONSE STYLES ACROSS 11 COUNTRIES IN PISA 2003

25. In order to give an answer to the question what the response style factor that was extracted using the Model shown in Figure 2 using Mplus actually measures, we calculated the correlations between this factor and the simple response style measures for 11 countries in PISA 2003. Table 13 presents the results.

26. First, we found positive correlations between the response style factor and the ‘simple’ acquiescence (CNT_ARS1_5it showed the highest correlation) and extreme response style measures and negative correlation between the ‘simple’ disacquiescence (CNT_DARS1_5it showed the highest negative correlation) responses style. These results indicate that the response style factor that was extracted using the Model shown in Figure 2 represents an acquiescence response style.
27. A second interesting finding also seems to support the conclusion that the response style factor represents an acquiescence response style. When we look at the correlations between the simple ARS and DARS response style measures and CNT_FINTMATH, CNT_FSCMATH and compare these with the correlation between the simple ARS and DARS with the original country mean attitudes CNT_INTMAT and CNT_SCMATH, we notice that the correlations between CNT_FINTMATH, CNT_FSCMATH and ARS are near zero whereas they are high and positively correlated with the original attitude country mean constructs. This indicates that ARS has been controlled for when estimating FINTMATH, FSCMATH. However, the correlations between DARS with CNT_FINTMATH, CNT_FSCMATH are positive compared to the correlations between DARS with the original INTMAT, SCMATH indicating that there still remains some disacquiescence response bias.

Table 13. Cross-country correlations between attitude indices, 'simple' response style measures, and response style factor for 11 countries in PISA 2003

	CNT_PVMATH	CNT_INTMAT	CNT_SCMAT	CNT_MATHEFF	CNT_F_INTMATH	CNT_F_SCMATH	CNT_F_MATHEFF	CNT_F_RESPBIAS
CNT_PVMATH	1.000	-.896	-.544	.370	.246	.196	.015	-.316
CNT_INTMAT	-.896	1.000	.574	-.111	-.223	-.180	-.144	.261
CNT_SCMAT	-.544	.574	1.000	.408	-.282	-.246	-.120	.002
CNT_MATHEFF	.370	-.111	.408	1.000	.007	-.108	-.357	-.037
CNT_F_INTMATH	.246	-.223	-.282	.007	1.000	.918	.766	.654
CNT_F_SCMATH	.196	-.180	-.246	-.108	.918	1.000	.814	.439
CNT_F_MATHEFF	.015	-.144	-.120	-.357	.766	.814	1.000	.400
CNT_F_RESPBIAS	-.316	.261	.002	-.037	.654	.439	.400	1.000
CNT_ARS0_all_PRP	-.780	.798	.649	.005	.056	.050	.008	.381
CNT_ARS1_all_PRP	-.869	.922	.766	.070	-.175	-.179	-.135	.330
CNT_ARS0_5it_PRP	-.756	.736	.331	-.250	.143	.080	.071	.489
CNT_ARS1_5it_PRP	-.810	.874	.366	-.154	-.054	-.168	-.126	.519
CNT_DARS0_all_PRP	.464	-.603	-.404	-.058	.430	.427	.280	-.052
CNT_DARS1_all_PRP	.746	-.846	-.678	-.099	.383	.383	.258	-.194
CNT_DARS0_5it_PRP	.036	-.124	.169	.057	.375	.474	.296	-.024
CNT_DARS1_5it_PRP	.564	-.684	-.188	.072	.286	.424	.300	-.360
CNT_ERS0_all_PRP	-.437	.373	.353	-.029	.298	.290	.168	.314
CNT_ERS0_5it_PRP	-.509	.449	.319	-.145	.295	.302	.203	.328
CNT_ARS_BD_PRP	-.648	.695	.098	-.324	-.121	-.185	-.091	.247
CNT_NCR	-.614	.622	.600	.108	-.259	-.179	-.364	.028

APPENDIX B MPLUS INPUT

```
TITLE: CFA with continuous factor indicators
DATA:
!replace AUS with cntcode
FILE IS
P:\PISA\AttitudeAchParadox\IRT_ResponseStyle\Mplus\AUS\AUS_Mplus.txt;

VARIABLE:
NAMES ARE country schoolid stdidst atschl1-atschl4 sturel1-sturel5
        intmat1-intmat4 instmot1-instmot4 matheff1-matheff8
        scmath1-scmath5 cooplrn1-cooplrn5 math1-math83 read1-read28
        scie1-scie34 prob1-prob19 pvmath1-pvmath5 pvread1-pvread5
        pvscie1-pvscie5 pvprob1-pvprob5 w_fstuwt;

USEVARIABLES ARE intmat1-intmat4 scmath1-scmath5
 matheff1-matheff8;

MISSING ARE intmat1-intmat4 scmath1-scmath5
 matheff1-matheff8 (9);

AUXILIARY=country schoolid stdidst;

!WEIGHT IS w_fstuwt;

ANALYSIS: ITERATIONS = 3000;

MODEL:

      f1 BY intmat1-intmat4;
      f2 BY scmath1-scmath5 ;
      f3 BY matheff1-matheff8;
      f4 BY intmat1-intmat4 scmath1-scmath5 matheff1-matheff8@0;
      f4 with f1@0;
      f4 with f2@0;
      f4 with f3@0;


OUTPUT: TECH1 TECH3 TECH4 MODINDICES;
SAVEDATA:
FILE IS AUS_Mplus_Fscores_INTMATHSCMATH.txt;
SAVE=FSCORES;
```

APPENDIX C TABLE 14. COUNTRY MEAN MATH PERFORMANCE AND MEAN RESPONSE STYLES IN PISA 2003

CNT	CNT_PVMATH*	CNT_ARS 0_all_PRP	CNT_ARS 1_all_PRP	CNT_ARS 0_5it_PRP	CNT_ARS 1_5it_PRP	CNT_DARS 0_all_PRP	CNT_DARS 1_all_PRP	CNT_DARS 0_5it_PRP	CNT_DARS 1_5it_PRP	CNT_ERS 0_all_PRP	CNT_ERS 0_5it_PRP	CNT_ARS_BD_PRP	CNT_NCR
HKG	550.38	.13	.35	.10	.34	.08	.26	.06	.24	.22	.16	.05	.38
FIN	544.29	.12	.32	.06	.24	.12	.30	.12	.35	.24	.18	.03	.44
KOR	542.23	.10	.29	.07	.25	.13	.33	.11	.34	.23	.18	.03	.33
NLD	537.82	.11	.31	.06	.24	.12	.31	.10	.34	.23	.16	.03	.41
LIE	535.80	.20	.37	.11	.27	.16	.30	.15	.35	.36	.25	.05	.39
JPN	534.14	.11	.27	.12	.28	.19	.38	.13	.34	.30	.25	.05	.35
CAN	532.49	.19	.39	.11	.29	.11	.26	.13	.32	.30	.24	.03	.39
BEL	529.29	.14	.33	.09	.27	.13	.30	.11	.33	.27	.20	.03	.43
MAC	527.27	.15	.35	.11	.32	.08	.27	.07	.27	.23	.17	.05	.39
CHE	526.55	.20	.38	.13	.29	.15	.30	.16	.35	.36	.29	.04	.41
AUS	524.27	.16	.38	.10	.29	.10	.25	.11	.31	.26	.21	.03	.38
NZL	523.49	.16	.38	.09	.30	.10	.25	.10	.30	.26	.20	.03	.41
CZE	516.46	.14	.34	.07	.25	.10	.28	.09	.33	.24	.16	.03	.37
ISL	515.11	.18	.37	.11	.27	.14	.29	.16	.36	.32	.27	.03	.40
DNK	514.29	.16	.36	.12	.33	.12	.27	.12	.29	.27	.24	.03	.46
FRA	510.80	.17	.36	.11	.28	.14	.30	.14	.35	.32	.25	.03	.48
SWE	509.05	.15	.35	.09	.26	.13	.30	.15	.36	.28	.24	.03	.44
GBR	508.26	.15	.36	.10	.30	.11	.27	.11	.30	.26	.20	.03	.41
AUT	505.61	.20	.36	.11	.26	.19	.34	.19	.39	.39	.31	.03	.47
DEU	502.99	.20	.37	.13	.28	.17	.32	.16	.36	.37	.29	.04	.44
IRL	502.84	.16	.36	.09	.28	.13	.29	.11	.32	.28	.20	.03	.40
SVK	498.18	.16	.37	.08	.27	.09	.25	.07	.31	.25	.15	.03	.35
NOR	495.19	.16	.35	.12	.30	.15	.31	.13	.32	.31	.25	.03	.44

LUX	493.21	.19	.36	.13	.29	.18	.33	.17	.36	.37	.30	.05	.46
POL	490.24	.14	.36	.09	.28	.10	.26	.10	.31	.24	.19	.05	.37
HUN	490.01	.15	.34	.09	.28	.12	.29	.09	.31	.27	.17	.04	.36
ESP	485.11	.16	.36	.13	.32	.13	.28	.11	.29	.29	.24	.04	.40
LVA	483.37	.12	.33	.07	.26	.10	.28	.10	.33	.22	.18	.03	.42
USA	482.88	.20	.41	.13	.33	.10	.24	.12	.30	.30	.25	.06	.38
RUS	468.41	.14	.35	.07	.24	.09	.27	.10	.35	.23	.17	.04	.36
PRT	466.02	.16	.38	.12	.34	.11	.25	.09	.26	.26	.21	.04	.44
ITA	465.66	.16	.37	.11	.32	.12	.27	.08	.28	.27	.20	.04	.39
GRC	444.91	.18	.38	.13	.33	.11	.26	.10	.28	.29	.23	.05	.45
YUG	436.87	.18	.37	.11	.28	.12	.28	.12	.33	.30	.22	.05	.43
TUR	423.42	.22	.42	.17	.39	.11	.25	.10	.25	.34	.28	.09	.44
URY	422.20	.21	.41	.13	.32	.11	.25	.11	.30	.32	.24	.05	.40
THA	416.98	.15	.41	.12	.40	.05	.19	.03	.18	.21	.15	.08	.39
MEX	385.22	.23	.45	.19	.42	.08	.21	.09	.22	.32	.28	.08	.40
IDN	360.16	.17	.42	.11	.33	.06	.19	.06	.25	.23	.17	.05	.41
TUN	358.73	.31	.48	.22	.39	.12	.23	.16	.30	.43	.39	.06	.53
BRA	356.02	.24	.45	.18	.39	.08	.21	.09	.25	.32	.27	.05	.48

*The countries are ordered by math performance