Cross-National Response Styles in International Educational Assessments: Evidence from PISA 2006

Jack Buckley*

Department of Humanities and Social Sciences in the Professions

Steinhardt School of Culture, Education, and Human Development

New York University

June 2, 2009

*Send all correspondence to the author at 246 Greene St, New York, NY 10003, USA, voice: (212) 992-7676, e-mail: *spb5@nyu.edu*.

Abstract

Scholars and researchers of international education frequently use data from large-scale cross-national studies such as the Programme for International Student Assessment (PISA), a periodic international assessment consisting of cognitive, attitudinal, and sociodemographic measures conducted by the Organisation for Economic Co-operation and Development (OECD). Several components of PISA contain conventional Likert-type scales of psychological attitudes which are administered to students, parents, and school personnel in many nations. Secondary analysts of these attitudinal data, however, generally ignore the issue of cultural differences in response style or scale usage heterogeneity, leading to descriptive statistics and inferences that may be biased and misleading. In this paper I explore this issue several ways and illustrate the extent, consequences, and some possible solutions using data from PISA 2006.

Keywords: PISA, response style, international education, scale usage heterogeneity

Introduction

Large-scale international educational assessments—such as the Organisation for Economic Co-operation and Development's (OECD's) Programme for International Student Assessment (PISA) and the International Association for the Evaluation of Educational Achievement's (IEA's) Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy (PIRLS)—form one of the cornerstones of empirical research in international education. These studies, repeated at regular intervals, provide researchers with high-quality assessment data for students from many nations in mathematics, science, and reading at several age or grade levels. In addition, a typical administration of these large-scale international assessments also produces a wealth of noncognitive data by means of attitudinal survey items presented in various formats to students, parents, and school personnel. The primary function of these assessments is the production of a series of official reports issued by the organizing agencies, their member governments, and affiliated nongovernmental organizations. The most prominent of these official publications are the compilations of the relative rankings of nations in the various academic subject areas that are frequently cited in the news media and that form the basis of a recurring cycle of debate in education policy wherein analysts seek pedagogical and organizational best practices from the top-performing nations.

In addition to these "league tables," the data are also widely used by secondary researchers interested in exploring the relationship between pedagogy, sociodemographic factors, or education policy and academic outcomes both within and between nations (e.g. Loveless 2007; Haahr, Nielsen, Hansen, and Nielsen 2005; Woessmann 2001). These analysts typically move beyond simple analysis of the aggregate assessment data and empirically examine the relationships between assessment outcomes and other factors at various levels of analysis—including, occasionally, the individual student level. Often these analyses both center on attitudinal measures and are cross-national in nature. For example, Loveless (2006, pp. 13-21) examines the relationship between enjoyment of mathematics and

academic performance as measured by TIMSS 2003 at the country level.

Secondary analysts who use attitudinal data from international education assessments are at risk of reaching erroneous conclusions if they do not consider the issue of cultural differences in survey response style. Although findings of bias due to cultural response styles or scale usage heterogeneity are well known in cross-cultural psychology, marketing, and public opinion research, there appears to be little attention paid to these issues in international education research.

A central goal of cross-cultural research in fields like public opinion, marketing, and education is to identify important differences in attitudes and perceptions and to link these differences to other outcomes of interest like voting behavior, purchasing decisions, or academic achievement. Unfortunately, there is an increasingly large body of evidence that suggests that many observed cross-national or cross-cultural differences are, in fact, contaminated by artifacts of measurement (Javaras and Ripley 2007; King, Murray, Salomon, and Tandon 2004; Johnson 2003; Rossi, Gilula, and Allenby 2001; Baumgartner and Steenkamp 2001; Heine, Takata, and Lehman 2000; de Vijver and Leung 1997; Chen, Lee, and Stevenson 1995; Mullen 1995; Greenleaf 1992; Poortinga 1989).

Much of this research focuses particularly on cross-cultural differences in the usage of Likert (1932) scales or individual categorical items drawn from such scales. Baumgartner and Steenkamp (2001) provide a useful summary of the various *response styles* or differences in response scale usage that can lead to bias in cross-cultural attitude research. In the present study, I am particularly interested in four styles. Acquiescence response style (ARS), or positivity bias, is a tendency to agree with items regardless of actual attitude. Its opposite, disacquiescence response style (DARS) is a tendency to disagree with items regardless of their content. Extreme response style (ERS) is a tendency to choose the endpoints of an item's scale (e.g. "very satisfied" or "very dissatisfied"), again regardless of the actual, underlying attitude. Finally, noncontingent responding (NCR) is a term used to describe the random

or careless response to items.[1]

Although individual respondents' idiosyncratic usage of different response styles adds noise to attitude survey data, systematic differences in response style across nations or cultures can introduce far more serious biases in both descriptive statistics and inferential results from more complex models. Unfortunately, there is much empirical evidence of such systematic biases between cultures. For example, Chen, Lee, and Stevenson (1995) report that Chinese and Japanese secondary students are more likely to use the midpoint of a seven-point Likert-type item, while U.S. students exhibit a greater tendency toward ERS than the Asian students or their Canadian counterparts (although Chen and colleagues find little effect on cross-national comparisons of item means). Watkins and Cheung (1995) examine response styles of high school students from five countries and report substantial variation in the tendency to exhibit several response styles, including ERS and NCR, on academic self-esteem items. Marin, Gamba, and Marin (1992) compare Hispanics to non-Hispanic Whites and find a greater incidence of both ERS and ARS among the Hispanic population, particularly the less educated and less acculturated. Bachman and O'Malley (1984) find similar results comparing Black with White respondents. Using international marketing data, Clarke (2001) reports cross-national differences in ARS and ERS that lead to biased inference, if uncorrected.

These findings, particularly the cross-national research on secondary school populations, suggest that heterogeneity in response style could be a potential source of bias in the secondary analysis of PISA and other international assessment data. In this paper I investigate the extent, form, and consequences of cross-cultural differences in response style or scale usage using data from the PISA 2006 student questionnaire and science assessment (OECD 2007). In the next section, I briefly introduce the PISA data and then turn to an exploratory analysis of response style heterogeneity across PISA nations using some simple methods sug-

[1]There are a variety of other response styles discussed in the literature that are less relevant to the PISA data. The tendency to use the middle of the response scale (midpoint responding), for example, is not observable in the PISA attitude data considered here as each item has a four-point response scale.

gested in the cross-cultural psychology literature. This analysis is followed by the estimation of a more sophisticated model suggested by Rossi, Gilula, and Allenby (2001). I conclude with some suggestions for minimizing bias due to response style in future PISA administrations.

Exploring Response Styles in PISA 2006

OECD's PISA is an international assessment of 15-year-olds in science, reading, and mathematics that is conducted triennially. In each administration, one of the three subject areas, on a rotating basis, is chosen as the focus. In 2006, the focus area was science, and the assessment was given to approximately 400,000 students in 57 countries. In addition to the assessment, which primarily consisted of cognitive items but also contained some attitude items for the science focus area, PISA 2006 also included student, school, parent, and information communication technology questionnaires (Organisation for Economic Co-operation and Development 2007).[2]

In addition to numerous socioeconomic and demographic background items, the student questionnaire also included seven Likert-type scales measuring various attitudes toward science. Each of the conceptual scales is composed of items on a four-point response scale (1 = strongly agree, 2 = agree, 3 = disagree, 4 = strongly disagree). Surveys were administered via pencil and paper with no reverse-scored items and all of the items in each scale clearly grouped together on the form. The student science attitude scales are summarized in Table 1 (scale names used here are not official).

As is evident from Table 1, each science attitude scale appears internally consistent, with the minimum observed value of Cronbach's (1951) $\alpha = .762$. Over all 41 items, the average interitem correlation is .293, with a range of .079 to .792, all positive.

[2]Public use data from the assessment and questionnaires, along with survey instruments and codebooks, is available online at http://pisa2006.acer.edu.au/downloads.php.

Table 1: Attitude scales in the PISA 2006 student questionnaire

	Number of items	α	n
Science Enjoyment	5	.904	389,721
Science Value	10	.851	380,106
Environmental Responsibility	7	.762	378,918
Usefulness for Science Career	4	.808	388,028
Science in Future (a)	4	.916	386,728
Science in Future (b)	5	.916	362,823
Science Learning	6	.909	360,570

Note: α denotes Cronbach's (1951) α; n is number of observations (listwise deletion of missing values). The two Science in Future scales combine to a single scale with $\alpha = .927$. All scale items are on a four-point response scale. All statistics reported are unweighted. Source: PISA 2006 student questionnaire data file.

Measuring Response Style

The literature on response style suggests several potential ways to measure ARS, DARS, ERS, and NCR. Here, in the case of ARS and DARS, I follow Bachman and O'Malley (1984) and Baumgartner and Steenkamp (2001) and compute a simple acquiescence index and a disacquiescence index based on responses to a heterogeneous subset of the 41 attitude items. The choice of items is based on an ad hoc examination of both the factor structure of the superset of items and also the item-test correlations estimated after considering all 41 items to be on a single scale ($\alpha = .945$). The resulting five-item index appears to meet the criterion of heterogeneity of content ($\alpha = .482$, average item intercorrelation = .158, range of correlations = .101–.261).[3] Once the five items are chosen, I construct the ARS measure simply by computing, for each respondent, the proportion of "strongly agree" responses to the five items. The DARS measure is constructed by computing the proportion of "strongly disagree" responses.[4]

[3]The five items included in the index are `st18q04, st26q01, st26q07, st27q01`, and `st37q04`, which are drawn from the scales measuring Science Value, Environmental Responsibility (two items), Usefulness for a Science Career, and Science Learning, respectively.

[4]As a sensitivity test, I also compute an alternative (D)ARS measure using the proportion of "strongly (dis)agree" responses over the entire set of 41 items. While this set of items is clearly too homogeneous, these second measures nevertheless correlate with the measures constructed from the five-item subset at .741 (ARS) and .668 (DARS).

More formally,

$$\widehat{ARS_i} = \frac{1}{n_q} \sum_{j=1}^{n_q} 1[x_{ij} = 1], \tag{1}$$

and,

$$\widehat{DARS_i} = \frac{1}{n_q} \sum_{j=1}^{n_q} 1[x_{ij} = 4], \tag{2}$$

for each student respondent i over the $n_q = 5$ heterogeneous items, x_j, and $1[\ ]$ denotes the indicator function.

An analogous method of estimating ERS is to compute the proportion of extreme responses to a set of highly heterogeneous items (Baumgartner and Steenkamp 2001; Greenleaf 1992). Since each response scale has only four points in this case, my estimate of ERS is simply equal to the sum of the estimated ARS and DARS. Thus,

$$\widehat{ERS_i} = \widehat{ARS_i} + \widehat{DARS_i}. \tag{3}$$

For the fourth measure of response style, NCR, the methodology is somewhat different. Watkins and Cheung (1995; see also Baumgartner and Steenkamp 2001) suggest constructing an index using a series of item pairs chosen from items that are as highly correlated as possible (given the data in hand), have similar means, and are scored in the same direction. For the PISA 2006 data, I choose the five pairs of items with the largest interitem correlations (range = .7165–.7925). All five pairs are drawn from the two Science in Future scales.[5] Once these items are chosen, NCR is estimated simply as the sum of absolute differences between the item pairs:

$$\widehat{NCR_i} = \sum_{j=1}^{n_p} |x_{ij} - y_{ij}|, \tag{4}$$

where $n_p = 5$ is the number of pairs of highly similar items x_j, y_j observed for student i. The

[5]The five pairs of items are {st29q01,st29q02}, {st29q03,st29q04}, {st35q01,st35q02}, {st35q04,st35q05}, and {st35q02,st35q05}, in order of greatest to least correlation.

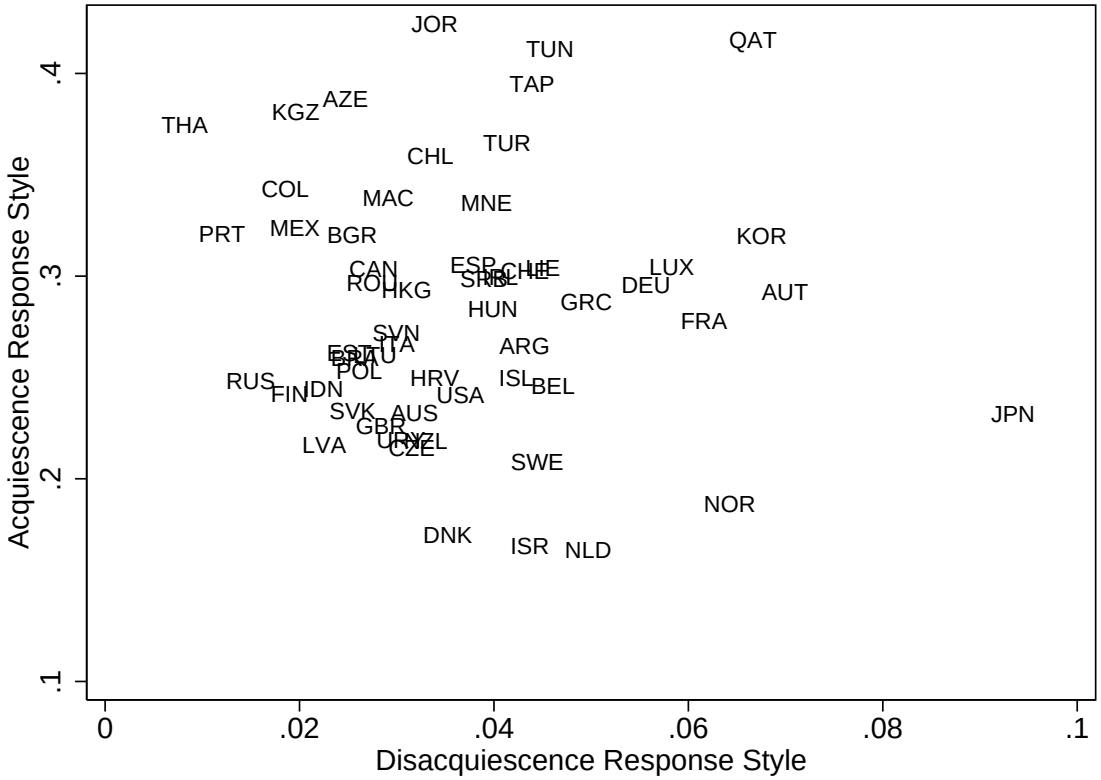
logic underlying this measure of NCR is that responses to highly similar items should differ very little; if respondents exhibit large absolute differences in response to similar items, this suggests an erratic or random response style.

With these simple, ad hoc, measures of the four response styles defined, I now turn to estimating each measure for each country in the PISA 2006 student data. I compute each measure for each student in the dataset using equations (1)–(4) and then compute the sample mean using the final student weight provided on the file to account for differential probability of selection, unit nonresponse, and coverage bias. I also compute standard errors for each measure using the survey design variables supplied on the file (stratum and primary sampling unit identifiers) via Taylor series linearization. The complete results for each country are in Appendix Table A1, but the results are summarized graphically in Figures 1 and 2, below.

Figure 1 presents ARS versus DARS for all 57 PISA 2006 countries. As the figure shows, the estimated values of ARS are substantially larger on average than the estimates of DARS, suggesting a positive net acquiescence response style (or directional bias) internationally (Greenleaf 1992). Several nations appear to exhibit relatively high average ARS, including Jordan, Tunisia, and Qatar, implying a tendency to agree with scale items regardless of content. Qatar also has a relatively large estimated DARS, the propensity to disagree with items regardless of content, although Japan appears to be an outlier on this measure ($\widehat{DARS} = .093$).

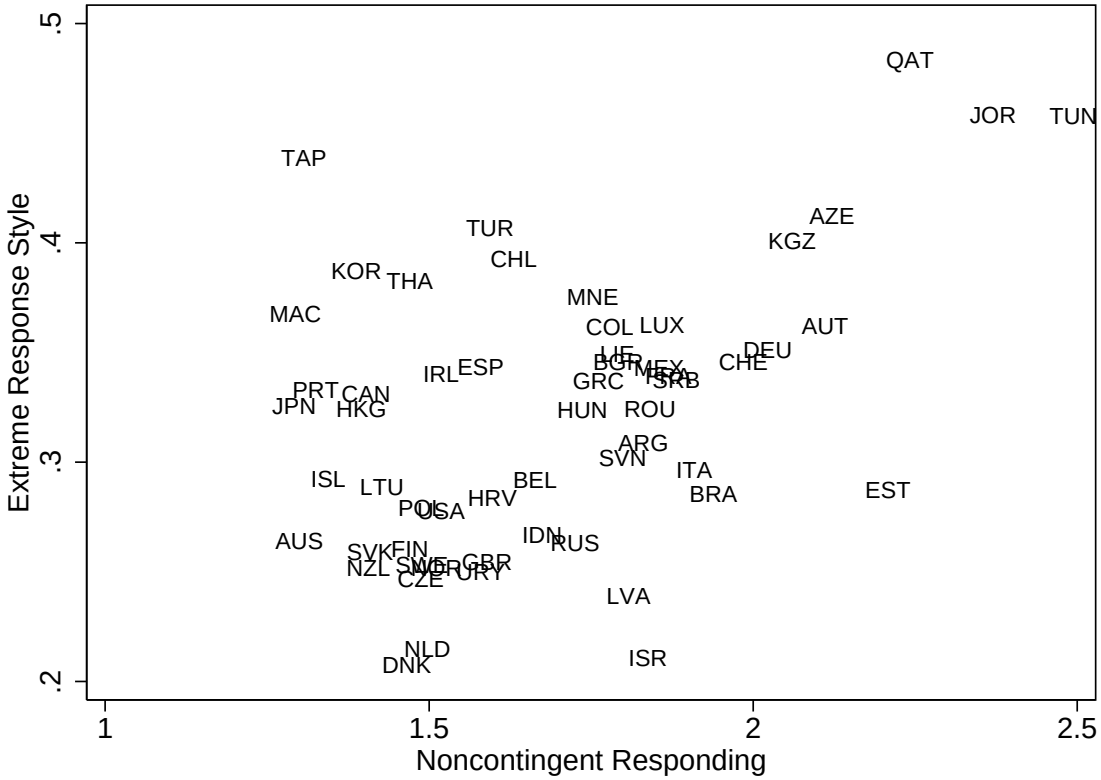
Figure 2 shows ERS versus NCR for the same 57 nations. As one would expect given its relatively large estimated values of ARS and DARS, Qatar has one of the largest estimated values of ERS, along with Jordan, Tunisia, and Taipei, suggesting that students from these nations have a tendency to use both anchors of the response scale regardless of the item content. Qatar, Jordan, and Tunisia also have the largest estimated values of NCR, suggesting a tendency to report attitudes randomly or carelessly. Taipei, in contrast, has a relatively low estimated NCR (implying consistency in responses to similar items), as do

Figure 1: Acquiescence response style (ARS) versus disacquiescence response style (DARS) for students from all 57 countries in PISA 2006



Note: Results are survey weighted means of individual student responses computed via equations (1)–(4). Complete results, including standard errors, are presented in Table A1.

Figure 2: Extreme response style (ERS) versus noncontingent responding (NCR) for students from all 57 countries in PISA 2006



Note: Results are survey weighted means of individual student responses computed via equations (1)–(4). Complete results, including standard errors, are presented in Table A1.

Japan, Australia, and Macao.[6]

By construction of the measures, one would expect a negative correlation between ARS and DARS and a positive correlation between ERS and both ARS and DARS at the individual student level; there is no expected correlation between NCR and the other measures. For a nation like Qatar, for example, which appears to have relatively high measures of all response styles, this correlation structure implies separate subgroups of respondents exhibiting high levels of either ARS or DARS and another subgroup with lower but still relatively large levels of ARS *and* DARS (the ERS respondents). The students exhibiting high levels of NCR, however, could cut across these subgroups.

A Simple Model of Response Style Effect

What are the consequences of this observed cross-cultural heterogeneity in response style? One approach to examining this important question at the scale score level is to assume a very simple linear classical measurement model. Let y_i denote an observed scale score (simply the sum of the responses to all items in the scale divided by the number of items) for student i, and let y_i^* denote the unobserved true score. If we assume linear and additive measurement effects due to response style in the population:

$$y_i = y_i^* + \beta_1 ARS_i + \beta_2 DARS_i + \beta_3 NCR_i, \tag{5}$$

then we can estimate y_i^* by fitting the linear regression model,

$$y_i = \beta_0 + \beta_1 \widehat{ARS_i} + \beta_2 \widehat{DARS_i} + \beta_3 \widehat{NCR_i} + \epsilon_i, \tag{6}$$

and estimating:

$$\hat{y_i^*} = \hat{\beta}_0 + \hat{\epsilon}_i. \tag{7}$$

[6]It is important to note that all nations are, on average, far from the theoretical maxima (1.0 for ARS, DARS, and ERS, and 15.0 for NCR) on all four response style measures. However, within nations, there are individuals who attain the maxima or large values on the various measures.

Note that equation (5) does not include ERS, as it is a linear combination of ARS and DARS as measured in the present study.

In Table 2, I present the results of the least squares estimation of equation (6) for two arbitrarily selected scales, Science Value and Science Enjoyment. Both models are adjusted for survey weights, and the reported standard errors are computed via Taylor series linearization using the design variables. The scales are constructed using the original coding, so smaller values of the scale indicate greater agreement.

Table 2: Estimating the effects of response style on two scales

	Science Value		Science Enjoyment	
	Coefficient	(s.e.)	Coefficient	(s.e.)
ARS	-1.449	(.006)	-1.676	(0.008)
DARS	1.467	(.015)	2.875	(0.023)
NCR	.012	(.001)	-0.009	(0.001)
Intercept	2.182	(.003)	2.480	(0.004)
n	358446		347768	
R^2	.575		.617	

Note: Results presented are from independent linear regressions of the Science Value and Science Enjoyment scales on acquiescence response style (ARS), disacquiescence response style (DARS), and noncontingent responding (NCR) for students from all 57 countries in PISA 2006. n is number of observations (listwise deletion of missing values) and R^2 is the proportion of variation accounted for in each linear model. All statistics reported are survey weighted and standard errors (s.e.) are adjusted for design via Taylor linearization. Source: PISA 2006 student questionnaire data file.

As Table 2 illustrates, the effect of ARS is, on average, to reduce the observed value on both scales. Moving from 0 (no acquiescence response style) to 1 (a response of 1 on all five of the ARS index items) predicts, on average, a decrease of 1.449 points for Science Value and 1.676 points for Science Enjoyment on a four-point scale. The estimated effect of DARS is, as expected, in the opposite direction and similar in magnitude in the case of Science Value, although almost double in magnitude for Science Enjoyment. The estimated effect of NCR is relatively small and unsystematic, in keeping with the hypothesis that NCR, in general, serves to add noise to measurement (Baumgartner and Steenkamp 2001).

With these estimates in hand, I use equation (7) to produce scale scores adjusted for

these response styles. I then compute the survey weighted means by country and standard errors using Taylor linearization for both the unadjusted and adjusted scales. The full results for each country are reported in Appendix Table A2. In the case of the Science Value scale, the unadjusted and adjusted values are correlated at .651; the correlation for the Science Enjoyment scale is .693. Over the 57 nations, the average difference (unadjusted minus adjusted) for the Science Value scale is -.227, and the average difference for the Science Enjoyment scale is -.215.

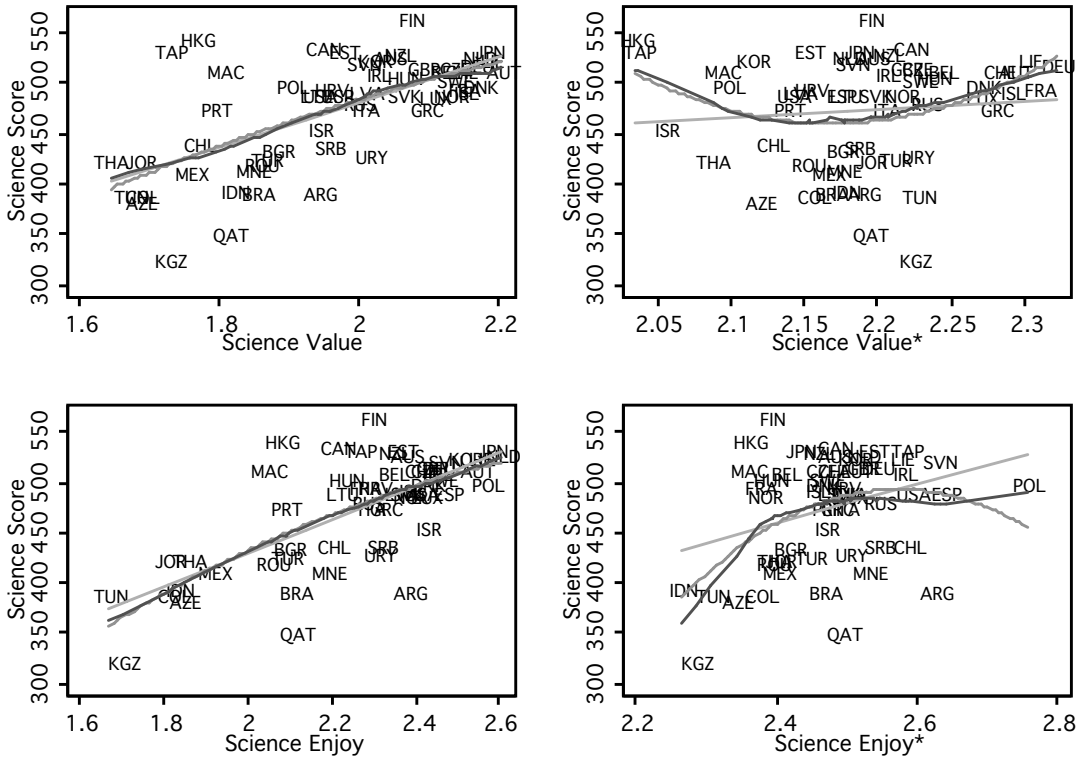
A key question is whether or not cross-cultural differences in response style lead to any biased inference. Using these measures adjusted for response style it is possible to investigate this issue. It is well documented in both PISA and TIMSS that, at the country level, assessment scores and attitudinal measures are often negatively correlated in a seemingly illogical way. For example, an OECD report assessing the validity of the embedded attitude scales[7] in the PISA 2006 (Organisation for Economic Co-operation and Development 2008) finds negative correlations between science achievement and the embedded scales for Interest in Science, Enjoyment of Science, General Value of Science, and Personal Value of Science at the country level and discusses similar results in reading in PISA 2000 and mathematics in PISA 2003. Loveless (2006), using data from TIMSS 2003, reports a similar paradoxical relationship at the country level between mathematics achievement and enjoyment of the subject. While it is possible that this negative relationship is a result of Simpson's Paradox or aggregation bias, it is also possible that it is at least partially a result of response style heterogeneity at the country level.

I investigate this latter possibility in Figure 3, which shows scatterplots of the unadjusted (left panels) and adjusted (right panels) scale scores versus the PISA science achievement scores, by country (Table A2). In each plot, I present three simple models: the lightest line shows a linear fit, the middle line shows a quadratic fit, and a local regression (loess) fit is represented by the darkest line. The left two panels replicate the familiar finding discussed

[7]Unlike the student survey items discussed above, these embedded items are included along with the cognitive items as part of the test booklets.

above: at the country aggregate level, more positive attitudes toward science appear to be linked to lower average levels of science achievement (recall that the attitude scales use the original coding–larger values indicate more negative attitude). The plots based on the the adjusted attitude scales, on the other hand, suggest a more complex nonlinear, relationship between achievement and attitude at the aggregate level.

Figure 3: Relationship between science achievement and attitudes towards science



Note: The left plots show the average PISA 2006 science achievement score versus the average score on the Science Value (top) and Science Enjoyment (bottom) scales, by country. Larger values on the attitude scales indicate more negative attitude. The right plots replace the two scale scores with the response style adjusted scores. Three fit lines are shown in each graph. The lightest line is a linear fit, the middle line is a quadratic fit, and the darkest line is a loess fit.

For both scales, the adjustment for response style appears to alter the linear relationship between achievement and attitude. According to the nonlinear model fits, in the case of the Science Value scale, countries with both high and low average attitudes appear to have, on

average, similar high levels of achievement. In the case of the Science Enjoyment scale, the original direction of the relationship is preserved, but with a great deal more nonlinearity.

What about at the individual student level? To explore the possible bias due to response style in estimating bivariate relationships, I consider data from three countries based on the results of the ad hoc estimation of ARS, DARS, ERS, and NCR above. Specifically, to cover an interesting range of estimated response styles, I choose:

1. Japan: high DARS, low ARS, moderate ERS, low NCR;
2. Tunisia: moderate DARS, high ARS, high ERS, high NCR, and;
3. United States: moderate ARS, DARS, ERS, and NCR.

I then consider the following pairs of simple models for each country independently:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \epsilon_{ij}, \tag{8}$$

$$y_{ij} = \beta_0 + \beta_1 x_{ij}^* + \beta_2 x_{ij}^{*2} + \epsilon_{ij}, \tag{9}$$

where y_i is the average over the five plausible values [8] of the estimated science achievement score for student i in country j, x_{ij} is the measure of attitude toward science (either Science Value or Science Enjoyment), and x_{ij}^* is the attitudinal measure adjusted for response style per equations (5)–(7). The results for each of the models, estimated via ordinary least squares, are presented in Table 3.

Since the estimated coefficients of the quadratic models are somewhat difficult to interpret by direct inspection, Figure 4 provides a graphical interpretation by plotting the predicted values for each pair of measures (unadjusted and adjusted) for each country. For all but one of the plots, the relationship between science achievement and the attitudinal measure is

[8] Due to the sampling design of the assessment, students are not assigned individual scores on the academic assessment components of PISA. Instead, researchers are given five draws from the estimated posterior distribution of student achievement on each subject test. Here I use the simple average of these "plausible values" and ignore the contribution of this uncertainty toward the estimation of variance. Standard errors presented in Table 3 are thus biased toward zero.

Table 3: Relationship between science achievement and attitudes at the individual level in three nations

	Japan		Tunisia		United States	
Value	38.69	(16.29)	-72.92	(13.63)	-64.99	(11.51)
Value^2	-21.97	(3.59)	6.25	(3.58)	0.87	(2.59)
Intercept	558.381	(18.38)	492.66	(12.88)	614.77	(13.48)
n	5581		4296		5400	
R^2	.113		.073		.097	
Value*	289.46	(38.24)	125.89	(33.64)	255.79	(31.23)
Value*^{2}	-67.09	(8.91)	-32.69	(7.81)	-64.26	(7.01)
Intercept	228.99	(40.69)	277.67	(39.97)	246.30	(34.66)
n	5883		3839		5227	
R^2	.018		.010		.018	
Enjoy	-15.42	(8.99)	-96.70	(8.98)	-57.70	(9.03)
Enjoy^2	-6.01	(1.82)	18.95	(2.21)	3.08	(1.66)
Intercept	615.72	(55.13)	491.64	(9.05)	611.04	(13.60)
n	5880		4286		5367	
R^2	.138		.038		.091	
Enjoy*	109.76	(15.46)	82.55	(15.30)	138.63	(24.12)
Enjoy*^{2}	-25.92	(3.31)	-15.60	(3.07)	-30.26	(4.54)
Intercept	426.40	(19.30)	288.59	(18.43)	343.64	(31.66)
n	5803		3832		5196	
R^2	.018		.007		.021	

Note: Results presented are from independent linear regressions of the mean of five science assessment plausible values on the unadjusted and adjusted Science Value and Science Enjoyment scales for students from three countries in PISA 2006. n is number of observations (listwise deletion of missing values). All statistics reported are survey weighted and standard errors are adjusted for design via Taylor linearization. Source: PISA 2006 student questionnaire data file.

approximately linear and in the opposite direction from the aggregate, country-level results. That is, more positive attitudes toward science are associated with higher scores on the science assessment. This reversal of the linear relationship observed at the aggregate level suggests that the reported paradoxical findings are, indeed, likely the result of aggregation bias or ecological fallacy.

After adjusting the attitudinal measures for response style, however, the relationship between achievement and attitude becomes, in each case, nonlinear, with an achievement peak somewhere in the middle of the attitude scale and lower levels of achievement associated, on average, at the extremes of both attitude scales.[9] This added level of complexity when response styles are considered suggests that ecological fallacy in the linear model results may not be a sufficient explanation for the observed differences between individual and aggregate relationships. That is, it is no longer accurate to say simply that the observed positive relationship between achievement and attitudes at the individual level is reversed at the country level; there appears to be a more complicated nonlinear relationship at both levels.

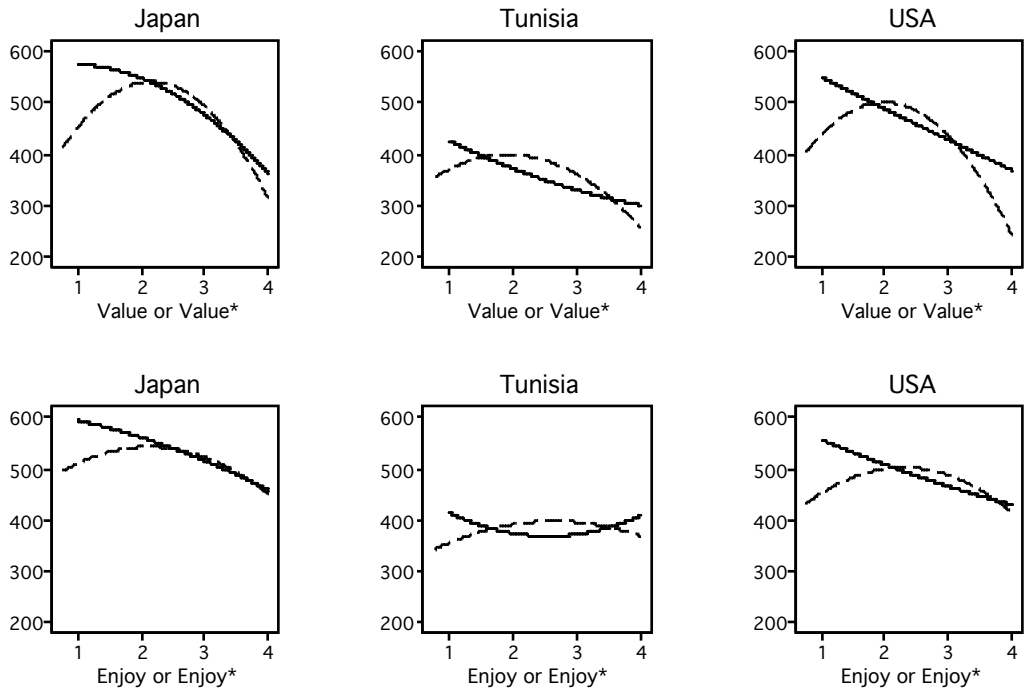
A Less Ad Hoc Approach?

While the methods used above for exploring cross-national response style in the PISA data yield some interesting results at both the aggregate and individual levels, it may be preferable to estimate response style with a method that relies less on the judgment of the researcher (in terms of the measurement of ARS, DARS, ERS, and NCR via ad hoc index construction) and more on the data. Moreover, the simple linear measurement model presented above is inappropriate for adjusting a single categorical item for response style, as it assumes continuous measures (scales) with elliptically symmetric distributions.

Recently, several more statistically sophisticated approaches have been suggested in the

[9]The exception to this pattern is Tunisia, in which the relationship between achievement and the unadjusted Science Enjoyment scale appears to be quadratic and mirroring the results for the adjusted scale. Given the average estimated levels of the various response styles for Tunisian students, particularly the relatively high ERS, this is not entirely unexpected.

Figure 4: Predicted values from the regression models of science achievement at the individual level



Note: Solid lines are the unadjusted (Value, Enjoy) measures and dashed lines are the response style adjusted (Value*, Enjoy*) measures. Complete results are presented in Table 3.

applied statistics literature with applications in public health, political science, education, and marketing. King et al. (2004) introduce a parametric model, compound hierarchical ordered probit, and a nonparametric variant, for estimating response scale usage heterogeneity (see also King and Wand 2007). While their approach appears to be promising and has even seen some application in the education context (Buckley and Schneider 2007, pp. 170-204), it has a major limitation with regard to the secondary analysis of PISA: to identify the differences in response style, respondents must be presented with additional *vignettes*, or items designed by the researcher to elicit a transitive ordering of attitudinal response. For example, in addition to being asked about their attitude toward their own school, students must also be asked about their attitudes toward several hypothetical schools of varying characteristics chosen by the survey researcher. Thus, the only way these methods could be used in PISA would require a sizable redesign of the survey instrument.

Javaras and Ripley (2007) present a method that avoids this limitation by using data from the entire survey to estimate response style at the individual level. Moreover, their approach—the multidimensional unfolding model—also models survey response as an "unfolding" process, which may be more desirable theoretically for certain attitude items (Roberts, Donoghue, and Laughlin 2000). In 2008, the National Center for Education Statistics (NCES), part of the U.S. Department of Education, conducted a small pilot study that attempted to estimate Javaras and Ripley's (2007) multidimensional unfolding model using student responses from two nations in the 2006 PISA data. While NCES judged the approach to be promising, they found the estimates to be very sensitive to initial conditions set during the estimation process and, thus, problematic as official statistics (National Center for Education Statistics 2008).

Here I choose to apply a different model, the Bayesian hierarchical approach suggested by Rossi, Gilula, and Allenby (2001). Their model is motivated by the problem of adjusting customer satisfaction data for response style or scale usage heterogeneity at the individual level. Unlike the approach of King and colleagues, Rossi, Gilula, and Allenby's model does

not require additional vignettes to identify response styles. And unlike Javaras and Ripley's model, this alternative model is fully Bayesian and thus estimable via Markov chain Monte Carlo methods. Although this model does not allow for the unfolding process as in the Javaras and Ripley approach, it still appears to allow sufficient flexibility to permit the modeling of several of the types of response style discussed above while using a functional form appropriate for categorical, Likert-type scale items.

Applying Rossi, Gilula, and Allenby's (2001) Model

Let the vector $y_i' = [y_{i1}, ..., y_{iM}]$ denote i's latent response to M questions in a scale, each with K response options, and x_{ij} the observed responses. The model resembles an ordinal probit (Aitchison and Silvey 1957), but with a joint multivariate normal distribution for the latent variables:

$$\begin{aligned} x_{ij} &= k \text{ if } c_{k-1} \leq y_{ij} \leq c_k, && (10) \\ y_i &\sim N(\mu_i^*, \Sigma_i^*), && (11) \\ \mu_i^* &= \mu + \tau_i \iota, && (12) \\ \Sigma_i^* &= \sigma_i^2 \Sigma, && (13) \end{aligned}$$

where τ_i is a respondent-specific location shift and σ_i is a scale shift. As Rossi, Gilula, and Allenby note, the (τ_i, σ_i) parameters flexibly model some response styles. For example, large positive values of τ and small values of σ correspond to use of the top end of the scale (i.e. DARS for the PISA data). Conversely, large negative values of τ imply overuse (relative to true attitudes) of the bottom end of the scale (ARS in the PISA case). Large σ and $\tau = 0$ model ERS. As the authors note, when $\sigma = 0$, this model could be considered a generalization of the polytomous Rasch model allowing correlated normal errors across items where τ is a latent attitude.

Since (τ_i, σ_i) are likely to be correlated, the model further departs from the usual approach

to Bayesian modeling of categorial data by assuming a joint bivariate normal prior:

$$\begin{bmatrix} \tau_i \\ \ln \sigma_i \end{bmatrix} \sim N(\phi, \Lambda), \tag{14}$$

where $E(\tau) = 0$, $E(\sigma^2) = 1$ are assumed for identification, which implies priors $\phi_1 = 0$ and $\phi_2 = -\lambda_{22}$. Rossi, Gilula, and Allenby further increase the flexibility of the model by allowing nonlinear spread of the c_k cutpoints:

$$\begin{aligned} c_k &= a + bk + ek^2, \quad k = 1, ..., K-1, && (15) \\ \sum_k c_k &= m_1, && (16) \\ \sum_k c_k^2 &= m_2, && (17) \end{aligned}$$

with m_1 and m_2 chosen so that for a given value of K, $e = 0$ implies even spacing of the cutpoints. Thus e is the only free parameter to be estimated. A negative value of e implies large spaces between cutpoints at the low end of the scale and tighter spaces at the high end, massing probability at the lower end.

The remaining prior distributions are specified as conjugate but diffuse (noninformative):

$$\begin{aligned} \pi(\mu) &\propto \text{constant}, && (18) \\ \pi(e) &\propto \text{unif}[-.2, .2], && (19) \\ \Sigma^{-1} &\sim W(v_\Sigma, V_\Sigma), && (20) \\ \Lambda^{-1} &\sim W(v_\Lambda, V_\Lambda). && (21) \end{aligned}$$

The model is estimated via Markov chain Monte Carlo using the modified and accelerated Gibbs sampler described in Rossi, Gilula, and Allenby (2001) (*R* function `rscaleUsage` in package `bayesm`).

To estimate this model using the 2006 PISA data, I use the student responses from the

same three countries from the 2006 PISA dataset as in the individual-level exploratory results above. I estimate the model on this subset of data using the 10 items in the Science Value scale. There are approximately 5,000 students in each country in sample (total $n = 15,577$). The model is estimated using the pooled data from all three countries. Table 4 provides estimated means and standard deviations of the posterior distributions of e, the item means over all respondents, μ, and the elements of the covariance matrix, Λ.[10]

Table 4: Some posterior quantities for the science value scale data

	Posterior mean	(Standard deviation)
e	-.060	(.004)
Λ_{11}	.218	(<.0005)
$\Lambda_{12}, \Lambda_{21}$	-.080	(<.0005)
Λ_{22}	.243	(.004)
μ_1	1.56	(.007)
μ_2	1.61	(.007)
μ_3	2.11	(.007)
μ_4	1.75	(.008)
μ_5	2.18	(.008)
μ_6	1.69	(.007)
μ_7	2.00	(.008)
μ_8	1.89	(.007)
μ_9	1.86	(.007)
μ_{10}	2.12	(.008)

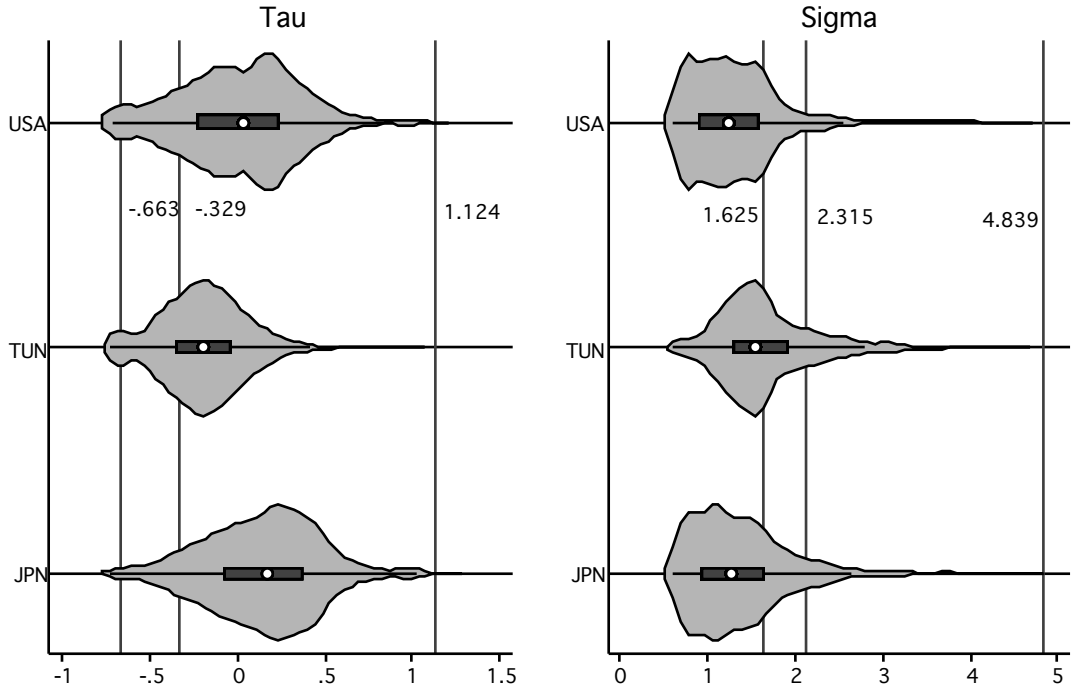
Note: Results based on 1,000 draws from Gibbs sampler (after 29,000 discarded as burn-in) as specified in Rossi, Gilula, and Allenby (2001). Data are 15,577 student responses from Japan, Tunisia, and the United States to the Science Value scale in PISA 2006.

As the table shows, e is slightly negative, suggesting a compression of the cutpoints at the high end of the scale (more negative) and more probability mass at the lower end. The negative estimated posterior mean for Λ_{12} means that there is a negative correlation between τ and σ overall across students from all three countries. The estimated item means μ are all very close to the empirical means, as expected.

[10]Results reported are based on 1,000 draws from the Gibbs sampler after 29,000 are discarded as burn-in. Visual inspection of the posterior draws and computation of the Geweke (1992) and Heidelberger and Welch (1983) diagnostics do not suggest nonconvergence. Additional posterior means for the matrix Σ are available upon request.

The real quantities of interest are the posterior values of τ and σ. Recall, however, that these are estimated at the individual level–15,577 estimated posterior distributions for each quantity. Thus instead of a table, I present these results graphically in Figure 5, which shows horizontal violin plots (Hintze and Nelson 1998) of the means of the posterior values for each student's estimated density of τ and σ. These plots combine the information in a boxplot (the dark rectangular region denoting the interquartile range and white circle denoting the median) with a density trace showing the distribution of the estimated quantities.

Figure 5: Horizontal violin plots of the estimated posterior means of τ and σ for students in three nations



Note: results are plotted from the estimated posterior means of 15,577 student responses using the PISA 2006 student Science Value scale data. The three vertical lines on each plot show the results of the "worst case" response styles.

Because direct interpretation of the posterior means for τ and σ is not obvious, I also include vertical lines corresponding to the estimates for three hypothetical students. The first student's response pattern on the Science Value scale item is simply a vector of 1's,

corresponding to all "strongly agree." Their estimated posterior means for τ and σ are -.663 and 1.625, respectively. The next student's response pattern is a vector of 4's, corresponding to all "strongly disagree," and their posterior means are 1.124 and 2.315. Finally, the third hypothetical response pattern is an alternating series of 1's and 4's (five of each) and their posterior means are -.329 and 4.839.

As Figure 5 shows, none of the three nations has median estimated τ and σ values corresponding to any of these three extreme hypothetical cases. The U.S. median student τ is close to 0 (.033) and the median $\sigma = 1.229$, which does not suggest any particular response style in the aggregate, although individual students vary substantially around these medians. In particular, the density plot for the posterior means of the τ parameter for U.S. students appears bimodal, with a cluster of students in the left tail exhibiting estimated τ's consistent with the acquiescence response "worst case" pattern.

Japan's median student τ of .169 and σ of 1.265 suggest perhaps a slight disacquiescence response style, but not as extreme as the "worst case" DARS response pattern. The medians of Tunisia's estimated τ and σ are -.197 and 1.560 respectively, suggesting perhaps slight acquiescence response style. Additionally, the distribution of τ for Tunisia exhibits a secondary mode in the left tail. Further examination reveals that this is made up of approximately 300 students with a median estimated posterior mean $\tau = -.673$ and $\sigma = 1.574$, almost identical to the hypothetical "worst case" extreme ARS response pattern.

It is not immediately obvious which estimated quantity from the Rossi, Gilula, and Allenby model is the appropriate one to compare to the simple Science Value scale as an adjusted overall attitude. In Rossi and colleagues' marketing example, their scale includes an "overall" or summative satisfaction item and they demonstrate that the posterior estimates of z_i for this item are the most predictive of purchasing behavior. They note, however, that the τ_i might also be a good indicator of overall satisfaction, analogous to the row mean of the scale items. Since the PISA student Science Value scale does not include an overall summative item, I use the estimated posterior means of the τ_i for comparison.

As a first comparison, Table 5 shows, by country, unweighted linear correlations between the student-level estimates of τ, and the Science Value scale and the value scale adjusted via the simple linear measurement model discussed above. As the table shows, τ correlates quite highly with the simple scale across all three countries. The correlations between τ and the adjusted scale, value*, are moderate in comparison (and similar to the correlations observed between the unadjusted and adjusted value scales).

Table 5: Comparing the methods of correcting for response style

Japan	τ	value*
Value*	0.594	
Value	0.836	0.636
Tunisia	τ	value*
Value*	0.417	
Value	0.864	0.529
United States	τ	value*
Value*	0.594	
Value	0.894	0.647

Note: Correlations shown are unweighted product-moment correlations between the student-level Science Value scale, the adjusted scale using the estimates of ARS, DARS, and NCR from the simple linear measurement model and the estimated latent τ's from the Rossi, Gilula, and Allenby (2001) model.

The relatively close relationship between τ and the simple scale suggests that there is likely to be more similarity between any regression of an additional variable like science achievement on these measures than there was in the comparison between the regressions of achievement on the unadjusted and adjusted Science Value scales above. To test this, I estimate the same quadratic regression models as in equations (8) and (9):

$$y_{ij} = \beta_0 + \beta_1 s(x_{ij}) + \beta_2 s(x_{ij})^2 + \epsilon_{ij}, \tag{22}$$

$$y_{ij} = \beta_0 + \beta_1 s(\tau_{ij})^* + \beta_2 s(\tau_{ij})^2 + \epsilon_{ij}, \tag{23}$$

where y_i is again the average over the five plausible values of the estimated science achievement score for student i in country j, $s(x_{ij})$ is the standardized measure of Science Value, and $s(\tau_{ij})$ is the standardized student estimated posterior mean of τ. The results for each

of the models, estimated via ordinary least squares using survey weights and Taylor series linearization, are presented in Table 6.

Table 6: Comparing the simple Science Value scale and the estimates from the Bayesian hierarchical model using their relationship to science achievement.

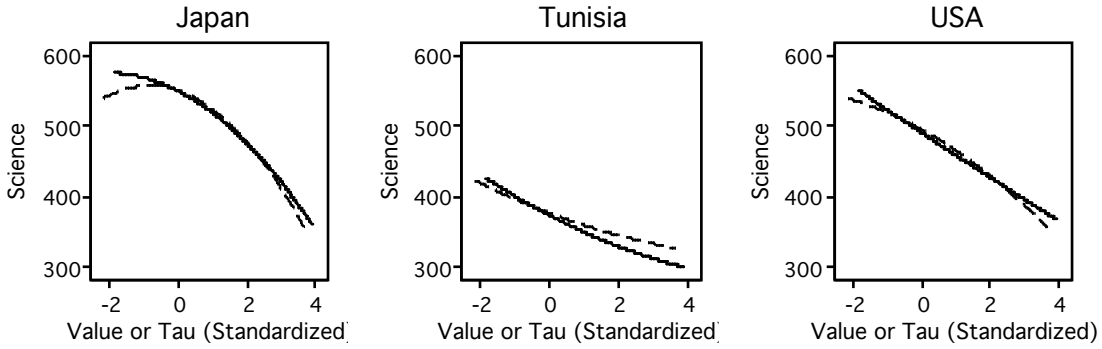
	Japan		Tunisia		United States	
Standardized Value	-24.95	(1.93)	-24.25	(1.61)	-32.58	(1.57)
Standardized Value^2	-6.06	(1.00)	1.72	(1.00)	0.24	(0.72)
Intercept	549.75	(2.85)	373.69	(2.77)	490.63	(3.12)
n	5581		4296		5400	
R^2	.113		.073		.097	
Standardized τ	-17.39	(1.69)	-20.11	(-10.09)	-28.12	(1.41)
Standardized τ^2	-10.17	(1.16)	0.50	(0.43)	-3.52	(0.95)
Intercept	550.89	(3.01)	377.29	(3.08)	495.29	(3.55)
n	5581		4296		5400	
R^2	.083		.040		.077	

Note: Results presented are from independent linear regressions of the mean of five science assessment plausible values on the unadjusted and standardized Science Value scale and on the standardized estimated posterior means of τ for students from all three countries in PISA 2006. n is number of observations (listwise deletion of missing values) and R^2 is the proportion of variation accounted for by the model. All statistics reported are survey weighted, and standard errors are adjusted for design via Taylor linearization. Source: PISA 2006 student questionnaire data file.

Because it is somewhat difficult to compare the results in Table 6 directly, Figure 6 shows the predicted values from the two models. The solid lines–the results for the regression of achievement on the simple scale and its square–are identical to the solid lines in Figure 4, above. The dashed lines show the results for the regression of science achievement on the standardized estimates of τ and its square. Whereas the regression results using the adjusted scale (the dashed lines in Figure 4) show a pronounced nonlinearity when compared to the unadjusted scale values, the results using τ shown in Figure 6 are more similar to those obtained using the unadjusted scale. In the case of Japan, there is an increase in curvature at the tails (and also, to a lesser extent, for the United States) indicating lower predicted values of achievement for students with attitudes at either extreme as compared with the unadjusted scale. In the case of Tunisia, however, this difference is reversed at the low end

of the attitude scale. As expected, the Bayesian hierarchical adjustment for response style does not produce as extreme changes as does the ad hoc approach discussed above; this can be seen in a visual comparison between Figure 6 and Figure 4.

Figure 6: Predicted values from the regression models of science achievement at the individual level



Note: Solid lines are the unadjusted Science Value measures and dashed lines are the response style adjusted (τ) estimated posterior means. Complete results are presented in Table 6.

Discussion

Despite the large and growing body of literature illustrating the measurement pitfalls of cross-national and cross-cultural surveys, researchers all too often ignore these issues both in instrument design and analysis (Smith 2003, p. 69). Large-scale assessments like TIMSS and PISA set the standard for international psychometric assessment and careful instrument design, but even these data collection efforts may still be overlooking issues of cross-national

validity.

The present study is an exploration of methods that can be used to diagnose and, in some cases, correct such issues. Focusing on the student questionnaire from PISA 2006, I find evidence of cross-national response style variation using both a set of ad hoc methods and a more systematic Bayesian hierarchical approach. While promising, the Bayesian approach employed here is computationally cumbersome (estimation for the entire set of countries would be extremely computationally intensive) and rather opaque to the casual secondary user of international assessment data. The ad hoc methods, on the other hand, are simple to compute and relatively simple to explain. However, their validity relies on the selection of a sufficiently heterogeneous set of items (or homogeneous set of pairs of items, in the case of NCR). Given that, at present, the PISA attitude surveys do not contain any scales eliciting attitudes about any targets other than the main academic focus of the test, it would appear that the items selected here for measurement of ARS, DARS, and ERS are as heterogeneous as possible without some significant alteration of the survey content.

While further research is required to determine the most appropriate method of detecting response style differences in datasets like PISA, it seems clear that the problem should not be ignored. Part of this paper focuses on statistical adjustments that may be used to correct, at least in part, variation in measurement due to response style or scale use heterogeneity. However, an alternative and complementary avenue of research would be to investigate potential changes in item design that may mitigate the problem.

One example is related to the response scale used on the PISA student survey instrument. All seven of the scales (41 items in total) are coded using the identical four point scale discussed above (where 1 = strongly disagree,..., 4 = strongly agree). However, balancing scales by reverse-coding some items is a well-known method of reducing the impact of ARS and DARS on measurement (Paulhus 1991) and Baumgartner and Steenkamp (2001) present empirical evidence that this simple and virtually cost-free method is effective in reducing bias due to response style in the cross-national context. Beyond simple scale reversals, other

methods of reducing contamination due to ARS and DARS include reversals in the actual question wording and alternative item formats, such as dichotomous forced-choice items (Smith 2003, p. 81).

Research suggests that ERS can be mitigated by varying the number of response options on items within the scale (Hui and Triandis 1989) or by changing format to a ranking rather than a rating of items (Smith 2003, p. 82). Baumgartner and Steenkamp (2001) note that the severity of ERS appears related to distance between the scale midpoint and the scale mean, which suggests that bias due to ERS may be mitigated through more careful question design or the replacement of scale items that elicit average responses too far from the scale midpoint.

With respect to PISA, one strategy for the OECD would be to use a series of survey experiments as part of the piloting phase of the survey to assess the extent to which one or more of these simple modifications (or more complex interventions, such as the use of anchoring vignettes) reduces response style artifacts. Given the large number of nations involved in PISA and other international educational assessments, such survey experiments are much more feasible than in-depth ethnographic studies. In parallel to these experiments, OECD or member nations could also sponsor additional research on more advanced statistical methods for measuring and correcting scale usage heterogeneity.

References

Aitchison, J. and S. D. Silvey (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika 44*, 131–140.

Bachman, J. G. and P. M. O'Malley (1984). Yea-saying, nay-saying, and going to extremes: Black-White differences in response styles. *Public Opinion Quarterly 48*(2), 491–509.

Baumgartner, H. and J. E. M. Steenkamp (2001). Response styles in marketing research: A Cross-National investigation. *Journal of Marketing Research 38*(2), 143–156.

Buckley, J. and M. Schneider (2007). *Charter Schools: Hype or Hope?* Princeton, N.J.: Princeton University Press.

Chen, C., S. Lee, and H. W. Stevenson (1995). Response style and cross-cultural comparisons of rating scales among east asian and north american students. *Psychological Science 6*(3), 170–175.

Clarke, I. (2001). Extreme response style in cross-cultural research. *International Marketing Review 18*(3), 301–324.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika 16*(3), 297–334.

de Vijver, F. J. R. V. and K. Leung (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, and J. Pandey (Eds.), *Handbook of Cross-Cultural Psychology, Volume 1: Theory and Method*, pp. 257–300. Boston, Mass.: Allyn and Bacon.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*. Oxford, UK.: Clarendon Press.

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

Haahr, J. H., T. K. Nielsen, M. E. Hansen, and S. T. Nielsen (2005). *Explaining Student Performance: Evidence from the International PISA, TIMSS and PIRLS Surveys*. Danish Technological Institute.

Heidelberger, P. and P. D. Welch (1983). Simulation run length control in the presence of an initial transient. *Operations Research 31*, 1109–44.

Heine, S. J., T. Takata, and D. R. Lehman (2000). Beyond Self-Presentation: evidence for Self-Criticism among japanese. *Personality and Social Psychology Bulletin 26*(1), 71–78.

Hintze, J. L. and R. D. Nelson (1998). Violin plots: A box plot-density trace synergism. *The American Statistician 52*(2), 181–184.

Hui, C. H. and H. C. Triandis (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology 20*(September), 253–260.

Javaras, K. N. and B. D. Ripley (2007). An 'Unfolding' latent variable model for likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association 102*(478), 454–463.

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika 68*, 563–583.

King, G., C. J. L. Murray, J. A. Salomon, and A. Tandon (2004). Enhancing the validity and Cross-Cultural comparability of measurement in survey research. *American Political Science Review 44*(April), 341–355.

King, G. and J. Wand (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis 15*, 46–66.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology 140*.

Loveless, T. (2006). *The 2006 Brown Center Report on American Education: How Well Are American Students Learning?* Washington, D.C.: Brookings Institution.

Loveless, T. (Ed.) (2007). *Lessons Learned: What International Assessments Tell Us about Math Achievement*. Washington, D.C.: Brookings Institution Press.

Marin, G., R. J. Gamba, and B. V. Marin (1992, December). Extreme response style and acquiescence among hispanics: The role of acculturation and education. *Journal of Cross-Cultural Psychology 23*(4), 498–509.

Mullen, M. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies 26*, 573–596.

National Center for Education Statistics (2008). PISA Attitude Analysis Study. Internal Report.

Organisation for Economic Co-operation and Development (2007). *PISA 2006: Science Competencies for Tomorrow's World, Volume 1: Analysis*. Paris: OECD.

Organisation for Economic Co-operation and Development (2008). Validation of the embedded attitudinal scales: 25th meeting of the PISA governing board. Technical Report EDU/PISA/GB(2008)21, OECD, Warsaw, Poland.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, and L. S. Wright (Eds.), *Measures of Personality and Social Psychological Attitudes*, pp. 17–59. San Diego, Calif.: Academic Press.

Poortinga, Y. P. (1989). Equivalence of Cross-Cultural data: An overview of basic issues. *International Journal of Psychology 24*, 737–756.

Roberts, J. S., J. R. Donoghue, and J. E. Laughlin (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement 24*(1), 3–32.

Rossi, P. E., Z. Gilula, and G. M. Allenby (2001). Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association 96*(453), 20–31.

Smith, T. W. (2003). Developing comparable questions in cross-national surveys. In J. A. Harkness, F. J. R. V. de Vijver, and P. P. Mohler (Eds.), *Cross-Cultural Survey Methods*, pp. 69–92. Hoboken, N.J.: Wiley.

Watkins, D. and S. Cheung (1995). Culture, gender, and response bias: An analysis of responses to the Self-Description questionnaire. *Journal of Cross-Cultural Psychology 26*(5), 490.

Woessmann, L. (2001). Why students in some countries do better: International evidence on the importance of education policy. *Education Matters 1*(2), 67–74.

A Supplemental Tables

Table A1: Estimates of response styles, by country.

Country	ARS	s.e.	DARS	s.e.	ERS	s.e.	NCR	s.e.
ARG	0.266	0.006	0.043	0.038	0.309	0.006	1.830	0.038
AUS	0.232	0.004	0.032	0.029	0.264	0.004	1.300	0.016
AUT	0.292	0.005	0.070	0.064	0.362	0.005	2.111	0.035
AZE	0.388	0.007	0.025	0.021	0.412	0.007	2.122	0.042
BEL	0.246	0.004	0.046	0.042	0.292	0.004	1.663	0.019
BGR	0.320	0.005	0.025	0.022	0.346	0.005	1.792	0.033
BRA	0.260	0.005	0.026	0.023	0.286	0.004	1.939	0.037
CAN	0.304	0.004	0.028	0.026	0.331	0.004	1.403	0.020
CHE	0.303	0.005	0.043	0.040	0.346	0.004	1.985	0.023
CHL	0.359	0.006	0.033	0.030	0.393	0.005	1.630	0.022
COL	0.343	0.005	0.019	0.015	0.362	0.005	1.779	0.035
CZE	0.215	0.004	0.032	0.028	0.247	0.005	1.487	0.028
DEU	0.296	0.004	0.056	0.050	0.351	0.004	2.022	0.026
DNK	0.172	0.004	0.035	0.031	0.208	0.004	1.465	0.022
ESP	0.306	0.004	0.038	0.035	0.344	0.004	1.579	0.020
EST	0.262	0.004	0.025	0.022	0.287	0.004	2.208	0.030
FIN	0.242	0.004	0.019	0.017	0.261	0.004	1.470	0.019
FRA	0.278	0.005	0.062	0.056	0.339	0.004	1.869	0.030
GBR	0.226	0.004	0.028	0.026	0.254	0.004	1.589	0.019
GRC	0.287	0.004	0.049	0.046	0.337	0.004	1.761	0.038
HKG	0.293	0.005	0.031	0.028	0.324	0.005	1.396	0.024
HRV	0.250	0.004	0.034	0.031	0.284	0.004	1.597	0.023
HUN	0.284	0.005	0.040	0.036	0.324	0.005	1.737	0.030

IDN	0.244	0.004	0.022	0.020	0.267	0.004	1.674	0.031
IRL	0.300	0.005	0.041	0.037	0.340	0.005	1.519	0.025
ISL	0.250	0.005	0.042	0.039	0.292	0.005	1.345	0.023
ISR	0.167	0.004	0.044	0.039	0.211	0.004	1.838	0.043
ITA	0.267	0.004	0.030	0.028	0.297	0.004	1.909	0.023
JOR	0.425	0.005	0.034	0.030	0.459	0.005	2.370	0.037
JPN	0.232	0.004	0.093	0.089	0.326	0.004	1.292	0.022
KGZ	0.381	0.007	0.020	0.017	0.401	0.007	2.060	0.034
KOR	0.320	0.004	0.068	0.064	0.387	0.004	1.388	0.021
LIE	0.304	0.014	0.045	0.031	0.349	0.014	1.790	0.091
LTU	0.261	0.005	0.028	0.025	0.289	0.005	1.427	0.023
LUX	0.304	0.004	0.058	0.054	0.363	0.004	1.860	0.026
LVA	0.217	0.004	0.023	0.020	0.239	0.004	1.808	0.029
MAC	0.338	0.005	0.029	0.026	0.368	0.006	1.293	0.028
MEX	0.324	0.004	0.020	0.017	0.343	0.004	1.855	0.025
MNE	0.336	0.005	0.039	0.035	0.376	0.005	1.752	0.029
NLD	0.165	0.004	0.050	0.046	0.215	0.005	1.498	0.025
NOR	0.187	0.004	0.064	0.058	0.252	0.004	1.511	0.023
NZL	0.219	0.005	0.033	0.029	0.252	0.005	1.406	0.021
POL	0.253	0.004	0.026	0.024	0.279	0.004	1.488	0.021
PRT	0.321	0.005	0.012	0.010	0.333	0.005	1.325	0.027
QAT	0.417	0.003	0.067	0.063	0.483	0.003	2.242	0.028
ROU	0.297	0.006	0.027	0.023	0.324	0.006	1.841	0.040
RUS	0.248	0.005	0.015	0.013	0.263	0.005	1.725	0.061
SRB	0.299	0.005	0.039	0.036	0.338	0.005	1.882	0.028
SVK	0.233	0.004	0.025	0.022	0.259	0.004	1.409	0.028
SVN	0.272	0.004	0.030	0.027	0.302	0.005	1.799	0.026

SWE	0.208	0.006	0.044	0.040	0.253	0.006	1.489	0.024
TAP	0.395	0.003	0.044	0.041	0.439	0.004	1.307	0.016
THA	0.375	0.005	0.008	0.007	0.383	0.005	1.470	0.028
TUN	0.412	0.004	0.046	0.042	0.458	0.004	2.494	0.051
TUR	0.366	0.005	0.041	0.035	0.407	0.005	1.594	0.036
URY	0.219	0.005	0.030	0.027	0.250	0.005	1.580	0.029
USA	0.241	0.005	0.037	0.033	0.278	0.005	1.518	0.037

Note: Acquiescence response style (ARS), disacquiescence response style (DARS), extreme response style (ERS), and noncontingent responding (NCR) for students from all 57 countries in PISA 2006. Results are survey weighted means of individual student responses computed via equations (1)–(4). Standard errors (s.d.) computed via Taylor series linearization using survey design variables. Source: PISA 2006 student questionnaire data file.

Table A2: Results of simple linear scale adjustment.

Country	Value	s.e.	Value*	s.e.	Enjoy	s.e.	Enjoy*	s.e.
ARG	1.943	0.011	2.189	0.007	2.379	0.019	2.626	0.012
AUS	2.041	0.008	2.196	0.003	2.376	0.013	2.481	0.006
AUT	2.206	0.012	2.293	0.006	2.541	0.023	2.510	0.011
AZE	1.686	0.010	2.119	0.008	1.845	0.018	2.342	0.009
BEL	2.127	0.009	2.244	0.005	2.350	0.015	2.412	0.006
BGR	1.883	0.010	2.175	0.006	2.095	0.014	2.416	0.008
BRA	1.854	0.008	2.167	0.004	2.112	0.012	2.468	0.008
CAN	1.950	0.008	2.222	0.004	2.217	0.011	2.483	0.006
CHE	2.148	0.009	2.283	0.004	2.406	0.016	2.479	0.008
CHL	1.770	0.012	2.127	0.005	2.203	0.017	2.586	0.009
COL	1.686	0.010	2.155	0.006	1.818	0.015	2.378	0.009
CZE	2.139	0.010	2.226	0.006	2.423	0.015	2.465	0.008

DEU	2.166	0.011	2.323	0.006	2.451	0.019	2.540	0.009
DNK	2.176	0.008	2.273	0.004	2.430	0.017	2.466	0.010
ESP	1.973	0.007	2.176	0.005	2.479	0.012	2.642	0.007
EST	1.980	0.008	2.153	0.005	2.365	0.013	2.539	0.009
FIN	2.075	0.008	2.195	0.005	2.300	0.012	2.393	0.007
FRA	2.150	0.012	2.310	0.005	2.281	0.017	2.377	0.007
GBR	2.092	0.008	2.219	0.004	2.436	0.012	2.522	0.007
GRC	2.096	0.009	2.281	0.005	2.329	0.015	2.485	0.009
HKG	1.765	0.008	2.035	0.005	2.078	0.011	2.361	0.007
HRV	1.957	0.009	2.154	0.005	2.298	0.014	2.490	0.009
HUN	2.066	0.009	2.238	0.006	2.238	0.014	2.392	0.010
IDN	1.821	0.008	2.177	0.004	1.835	0.014	2.266	0.007
IRL	2.030	0.011	2.206	0.005	2.461	0.017	2.584	0.008
ISL	2.153	0.009	2.293	0.005	2.390	0.013	2.457	0.007
ISR	1.946	0.014	2.055	0.009	2.433	0.025	2.470	0.017
ITA	2.007	0.007	2.204	0.003	2.292	0.011	2.495	0.006
JOR	1.686	0.008	2.195	0.005	1.814	0.014	2.405	0.010
JPN	2.187	0.011	2.188	0.006	2.581	0.019	2.435	0.010
KGZ	1.730	0.010	2.224	0.006	1.702	0.014	2.286	0.009
KOR	2.021	0.009	2.113	0.006	2.510	0.017	2.510	0.008
LIE	2.194	0.029	2.304	0.018	2.555	0.040	2.580	0.028
LTU	1.939	0.008	2.178	0.005	2.224	0.012	2.480	0.008
LUX	2.110	0.010	2.271	0.006	2.421	0.014	2.505	0.008
LVA	2.012	0.009	2.147	0.005	2.369	0.012	2.505	0.007
MAC	1.809	0.007	2.094	0.006	2.047	0.011	2.360	0.009
MEX	1.759	0.006	2.165	0.004	1.920	0.010	2.400	0.007
MNE	1.849	0.008	2.177	0.006	2.197	0.011	2.535	0.008

NLD	2.170	0.010	2.180	0.006	2.605	0.018	2.520	0.009
NOR	2.136	0.013	2.217	0.006	2.391	0.015	2.383	0.008
NZL	2.059	0.010	2.206	0.004	2.348	0.015	2.457	0.006
POL	1.904	0.007	2.098	0.005	2.566	0.014	2.755	0.009
PRT	1.797	0.008	2.140	0.006	2.088	0.011	2.471	0.008
QAT	1.817	0.008	2.195	0.005	2.117	0.011	2.498	0.008
ROU	1.856	0.008	2.152	0.006	2.054	0.012	2.392	0.009
RUS	2.002	0.008	2.232	0.005	2.282	0.016	2.547	0.010
SRB	1.960	0.010	2.187	0.005	2.320	0.016	2.547	0.009
SVK	2.065	0.010	2.198	0.006	2.376	0.014	2.492	0.008
SVN	2.008	0.007	2.184	0.004	2.475	0.011	2.631	0.008
SWE	2.139	0.012	2.229	0.005	2.455	0.018	2.471	0.012
TAP	1.728	0.007	2.036	0.004	2.264	0.012	2.582	0.008
THA	1.643	0.007	2.087	0.003	1.861	0.011	2.396	0.007
TUN	1.670	0.010	2.226	0.007	1.670	0.013	2.307	0.009
TUR	1.867	0.015	2.210	0.007	2.087	0.018	2.445	0.009
URY	2.017	0.010	2.226	0.006	2.309	0.015	2.505	0.010
USA	1.942	0.010	2.142	0.006	2.407	0.014	2.592	0.008

Note: Survey weighted average scale scores for the student Science Value and Science Enjoyment scales and the scores adjusted via the simple linear measurement model, equations (5)–(7) for all 57 countries in PISA 2006. Standard errors computed via Taylor series linearization using survey design variables. Source: PISA 2006 student questionnaire data file.