JOURNAL OF APPLIED MEASUREMENT, *8*(3), 267-278

Ameliorating Culturally Based Extreme Response Tendencies To Attitude Items

Maurice Walker
Australian Council for Educational Research

Using data from the PISA 2006 field trial, Rasch item response models are used to demonstrate that extreme response tendency was exhibited differentially across culturally distinct countries when answering Likert type attitude items. A single attitude scale is examined across eight culturally distinct countries in this paper. Two avenues to ameliorate this tendency are explored: first using dichotomous variants of the items, and second incorporating the country specific response tendency into the Rasch item response model. Analysis of the item variants reveals similar scale outcomes and correlations with achievement but preference for the Likert variant when test information is considered. A hierarchical analysis using facet models reveals that the data fit significantly better in a model that incorporates an interaction effect between the country and the item delta parameters. The implications for reporting attitudes measured with Likert items across cultures are outlined.

Requests for reprints should be sent to Maurice Walker, Australian Council for Educational Research, 19 Prospect Hill Rd., Camberwell, VIC 3124, Australia, e-mail: walker@acer.edu.au.

Introduction

In international comparative surveys such as educational studies undertaken by the International Association for the Evaluation of Educational Achievement (IEA) or the Organization for Economic Co-operation and Development (OECD), questionnaires are commonly used to measure students' attitudes, beliefs, opinions and self-reported activities. One of the main reasons for collecting such information—hereafter referred to collectively as students' *background* information—is to provide data that may help explain patterns in achievement collected at the same time. Associations between some background scales and achievement differ markedly *across* countries (Kirsch et al., 2002; Martin, Mullis, Gonzalez and Chrostowski, 2004; Mullis, Martin, Gonzalez and Chrostowski, 2004; OECD, 2001, 2004) and there have been a number of suggestions for the cause of such variation. For example, Martin et al. (2004) suggest that differences in the association between science self-efficacy and science achievement across countries may be due to shared cultural conditions. Kirsch et al. (2002) note that correlations between reading engagement and reading literacy are lowest among countries with low mean reading literacy scores. Lie and Turmo (2005) observe that students in countries with low mean mathematics literacy scores tend to respond more positively to items on mathematics motivation than their counterparts in countries with high mean mathematics literacy.

There is some speculation that these observed differences may in part be attributed to response biases. Paulhus (1991, p. 17) defines response bias as "a systematic tendency to respond to a range of questionnaire items on some basis other than the questionnaire content". According to Paulhus the three main types of response bias are:

- **social desirability**, or the tendency to provide responses that the respondent believes are those which make him or her 'look good';
- **acquiescence**, or the tendency to agree rather than disagree with any statement; and
- **extreme response bias**, or the tendency to respond towards the extremes of a response scale rather than the centre of the scale.

This paper is concerned solely with extreme response bias. Whilst many studies note cross-cultural differences in response styles (Choi, Mericle and Harachi, 2006; Heine, Lehman, Peng and Greenholz, 2002) some studies more specifically conclude that distinct groups tend to differentially exhibit extreme response bias (Lee, Jones, Mineyama and Zhang, 2002; van Herk, Poortinga and Verhallen, 2004). Central to the investigation reported in this paper is the notion that cultural factors may influence response tendencies in international comparative educational surveys.

Typically, in international comparative studies a background trait is measured with a series of Likert type items forming a scale. However, because of response differences to Likert items between cultural groups Heine asserts that the use of such items "is most valid for identifying differences within rather than between groups" (Heine et al., 2002, p. 914). A common exploration in cross-cultural psychology is whether differences between cultures in answering Likert type questions can be explained by the location of the culture along an individualism-collectivism dimension. An individualist society is characterized as one that emphasizes the goals, values and rights of the individual; in contrast, a collective society is characterized as one that primarily promotes the goals of the society as a whole (see Heine et al., 2002 for a concise review of this literature). An additional factor posited as influencing extreme response bias is the literacy of the respondent (e.g. Flaskerud, 1988): it is argued that less literate individuals are less able to differentiate the subtleties between concepts such as *agree somewhat*, *agree* and *strongly agree*, and as such will usually opt for the least modified expression of their position (in this case *agree*). Another element in the debate is the linguistic equivalence of translations for the response options: in some languages the equivalent of *total* agreement is sought rather than *strong* agreement.

However, it is not the purpose of this paper to explore the potential reasons behind such differences. The fact that differences do exist is relatively uncontroversial, whereas their causes are equivocal in the literature. Rather, the study undertaken here arose from a simple pragmatic question: if there is extreme response bias exhibited differentially across cultures when answering Likert type items, what can be done to ameliorate this in cross-national survey research?

One potential remedy is to determine the optimal number of response options in Likert type scales that would reduce the effects of culturally related response bias. For example, Lee et al. (2002) reported that construct validity of a *sense of coherence* scale was stronger for Chinese and American respondents when there were four response categories; and stronger for Japanese respondents when there were seven response categories.

But one more obvious solution to extreme response bias is to remove the extremes of Likert type response scales altogether and use only dichotomous items. While this may not remove or ameliorate acquiesence, extreme response bias becomes a non-issue. This reasoning prompted the current investigation. To compare Likert and dichotomous variants, parallel versions of item batteries for measuring two constructs were administered in the field trial of the questionnaires for the PISA 2006 assessment of Scientific Literacy.

The following research questions were investigated:

- do cultures differentially exhibit extreme response tendencies when answering Likert type items?
- does item response format influence the key outcomes from questionnaire instruments in comparative studies?
- is it possible to ameliorate extreme response bias at the analysis stage of research?

Method

Two parallel versions of item batteries measuring the construct *enjoyment of science* (ENJSCI), were administered randomly in the PISA 2006 field trial: one variant was administered as Likert type items to half of the students; the other half received the same items in dichotomous format. In addition to comparing the results of the dichotomous items directly with those from the Likert type items, this paper also compares the Likert type items treated as though they were dichotomous items (by collapsing *strongly disagree* and *disagree* into a single category, and *strongly agree* and *agree* into another).

Translation of the items from two source languages, English and French, incorporated double blind translation, then reconciliation of the independent translations, followed by independent linguistic verification. This did not preclude the possibility, however, that countries' results were influenced by linguistic nuances of the items. Grisay (2002) provides a full description of the PISA translation procedures.

English source versions of the item battery for '*enjoyment of science*' appear in Table 1. As Likert type items, the English responses categories were in English *strongly agree*, *agree*, *disagree* and *strongly disagree*; as dichotomous items, the response categories were *agree* and *disagree*.

Table 1

Items used to measure enjoyment of science in PISA 2006 Field Trial

Item 1	I generally have fun when I am learning science topics
Item 2	I enjoy reading about science
Item 3	I am happy doing science problems
Item 4	I enjoy acquiring new knowledge in science
Item 5	I am interested in learning about science

To trial a large number of items in the PISA 2006 field trial, four questionnaires were randomly administered to students within each participating school. Two of the questionnaires contained the dichotomous variants of the ENJSCI items; the other two contained the Likert variants. The questionnaires contained a range of demographic and other items, including items about attitudes and motivation towards science. The questionnaires were administered after a two hour test on scientific literacy. Thirty minutes were scheduled for administration of the questionnaires although students were able to continue if they had not completed.

Participants

This paper compares the results from eight different countries that participated in the PISA 2006 field trial. The eight countries were selected for comparison to provide a range of test languages (each country administered the test in a different language), measures of mean science achievement, and geographic location.[1] This range of countries was deliberately chosen to emphasise the cross-cultural aspect of the investigation. While the cultural range of students within each country was not homogeneous, each country investigated had a different single dominant language of instruction within schools indicating the students were from distinct cultural societies.

The intent of the field trial was to examine the efficacy of test instruments, field operations and data processing procedures in preparation for the main PISA study. No reporting of findings was intended and the sample was not, therefore, designed to be rigorously representative of the population. For the purposes of the field trial, the sampling design was select a *convenient* sample of schools that was broadly representative of the different school types and study programmes within each country.

Table 2

Sample sizes for enjoyment of science variants

	Dichotomous	Likert
Country A	2745	2733
Country B	583	606
Country C	696	682
Country D	637	633
Country E	857	859
Country F	1104	1109
Country G	622	622
Country H	882	893
Total (pooled sample)	8126	8137

Usually 30-40 schools were selected and 30-40 15-year-old students within each school were randomly selected. Some countries opted to sample more than this so that comparisons could be made between sub-national entities (e.g., geographic regions). The data were collected in 2005. Table 2 provides the sample sizes.

Results

Differential exhibition of extreme response bias

Initial examination of the response frequencies indicated differences in response tendencies across countries. Figure 1 presents the proportions of responses for two countries—referred to hereafter as Country A and Country B—for item 4 (*I enjoy acquiring new knowledge in sci-*

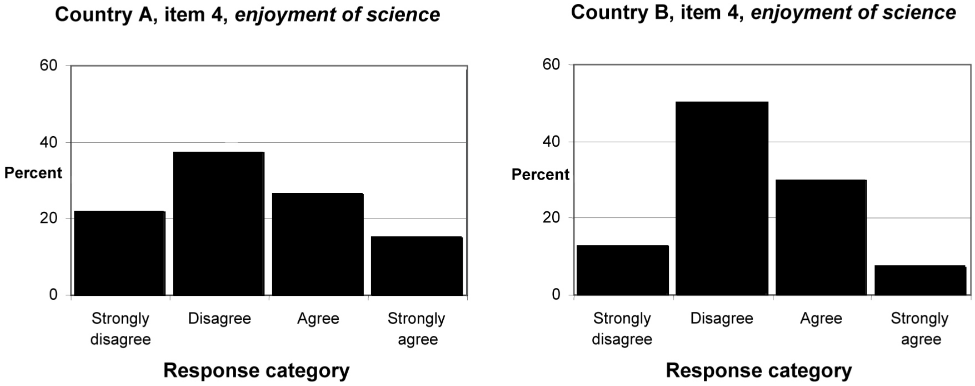

Figure 1. For Countries A and B, proportions in each response category of item 4, *enjoyment of science*

[1] As the data used here are from the PISA field trial and were not publicly released, the countries have not been identified.

ence). From these simple frequencies it is clear that students from Country A opted for more extreme response categories, both positive and negative, than those in Country B. This pattern was generally consistent across the five items measuring the construct although item 4 was the most extreme case. Therefore item 4 is used in this paper for illustrative purposes. Similarly, consistent response patterns were seen across the other six countries but Country A and Country B were the most extreme cases and are again used in this paper for illustrative purposes.

However, while these simple frequencies suggest differential response tendencies, they say nothing about the underlying latent trait of the respondents.

The two item batteries were therefore scaled with the Partial Credit Model (Masters, 1982) using ACER ConQuest software (Wu, Adams and Wilson, 1997).

Item fit was assessed using the ConQuest weighted mean-square statistic (infit), which is a residual-based fit statistic (Wu, 1997). Weighted infit statistics ranged between 0.94 and 1.02 for the item parameters and between 0.90 and 1.00 for the delta parameters (Masters, 1982), which all indicate good fit. All fit analyses were undertaken on the pooled sample data.

The goodness of fit for item 4 can be illustrated by examining the observed data for the pooled sample overlaid on the modelled cumulative probability curves, as in Figure 2. This shows the observed data (in dotted lines) closely matching the model (the solid lines), reflecting good fit. The figure plots the cumulative probability that a respondent will select a particular response category (or beyond) by the strength of their latent trait (in this case *enjoyment of science*). Given that the response categories are scored as 0 for *strongly disagree,* 1 for *disagree*, 2 for *agree,* and 3 for *strongly agree*, the left hand curves represent the probability that a respondent with a particular level of the underlying trait (shown on the horizontal axis) will score at least a one, the middle set at least two, and the right hand set, three.

However when the data was grouped by country and the observed values for each country were plotted alongside the same model, consider-

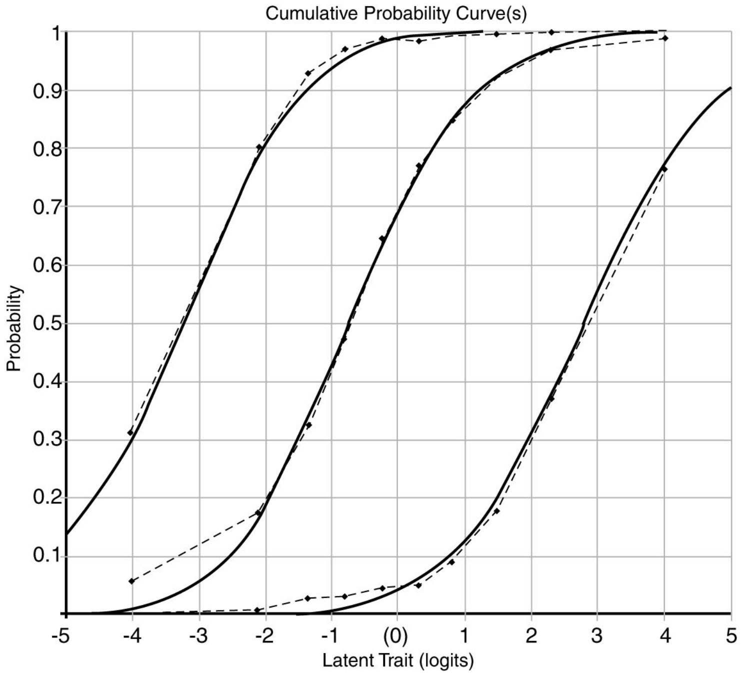

Figure 2. For the pooled sample, the cumulative probability curves for item 4 (Likert) in *enjoyment of science*, analyzed using the partial credit model

able country response differences were revealed. More specifically, the country data did not always fit the model. Figure 3 shows the observed data from countries A and B overlaid on the modelled cumulative probability curves for the item 4.

Of interest are the left and the right hand sets referring to the probability of selecting the extreme response categories. The left hand set of probability curves show that a person in Country B had a much higher probability of selecting *disagree* (or beyond) over *strongly disagree* than a person in Country A with the same level of *enjoyment of science*. For example, persons in Country B at –2 logits on the *enjoyment of science* scale had a probability of around 0.96 that they selected a category other than *strongly disagree*, whereas this was only about 0.77 for a person in Country A at –2 logits on the *enjoyment of science* scale. In other words, it was more likely that a respondent in Country A chose the extreme of *strongly disagree*.

The right hand set of probability curves show that a person in Country B had a much lower probability of selecting *strongly agree* than a person in Country A with the same level of *enjoyment of science*. A person in Country B at two logits on the *enjoyment of science* scale had a probability of around 0.11 to opt for *strongly agree*, whereas this was about 0.38 for a person in Country A with the same level of *enjoyment of science*. In other words, it was more likely that a respondent in Country A chose the extreme of *strongly agree*. In summary, the data shows that respondents in Country B tended to opt for less extreme responses than respondents in Country A with the same level of *enjoyment of science*.

Although Figure 3 represents the most severe case of misfit found across the eight countries and five items, misfit of this nature, if not magnitude, was a typical result.

Comparison with dichotomous variants

Having established a degree of misfit by country to the model, and with some evidence pointing towards differential extreme response tendencies, the dichotomous item variants were compared alongside 1) the Likert variants and 2) the Likert variants treated as dichotomous.

Table 3 provides the mean weighted likelihood estimate (WLE, Warm, 1989) for each country, using the partial credit model for the Likert variant and the Rasch model for the dichotomous variants. A ranking based on the country mean is

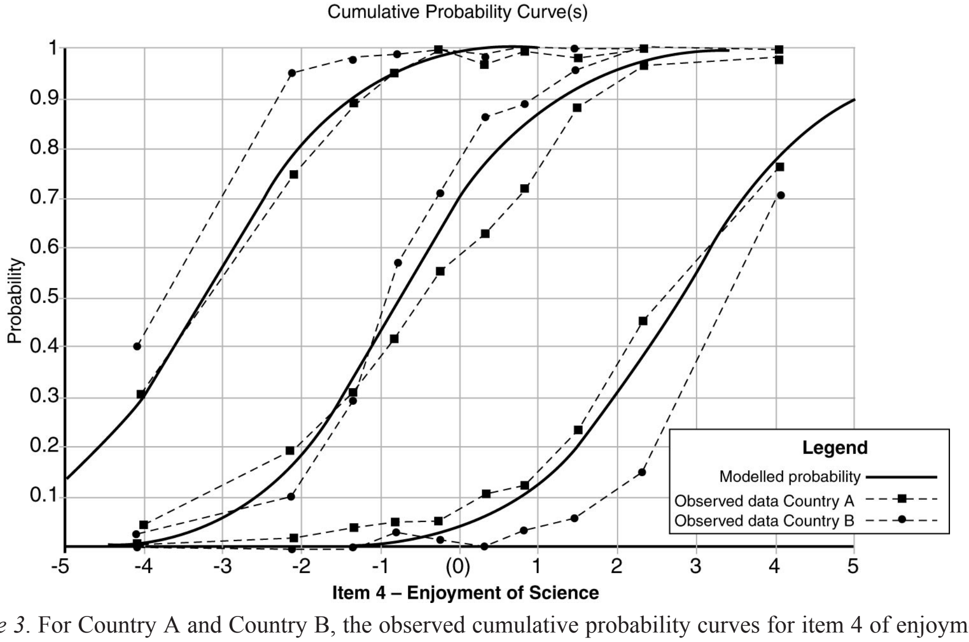

Figure 3. For Country A and Country B, the observed cumulative probability curves for item 4 of enjoyment of science, analyzed using the partial credit model

provided. The (attenuated) correlation between the attitude estimates and the score for science achievement (also a WLE) for each country is also given.

The data in Table 3 show that using simple Rasch models, the different methods of data collection and treatment effect only slight differences in the relative ordering of the mean *enjoyment of science* scores, and the correlations with science achievement remain fairly consistent. Thus, on these criteria, no preference clearly emerges for any one variant.

However, the Likert scale provides relatively more information about the respondents with high or low levels of the latent trait. The test information functions plotted in Figure 4 illustrate this. Note that while the absolutes value of the information scales can not be directly compared, the shape of the test information curves differ markedly. The dichotomous variant provided relatively less information about the respondents lying more than one standard deviation from the mean of the scale—those who lay beyond negative one and positive one logit. The test information curve for the Likert variant shows that this scale yielded more consistent information across the range of the population—the curve does not drop off until three standard deviations from the mean. One reason that the latter type of information curve is superior is that attitudes such as 'enjoyment of science' are considered educational outcomes and researchers are often interested in examining the characteristics of students with strong attitudes and the relationship between attitudinal outcomes and achievement.

Table 3

For each country, the mean enjoyment of science score (WLE), the correlation between enjoyment of science and science achievement, and rank order for mean enjoyment of science. Results for each method of data collection and treatment

	Dichotomous			Likert			Likert treated as dichotomous		
Country	Mean WLE	Correlation with Science Achievement	Rank	Mean WLE	Correlation with Science Achievement	Rank	Mean WLE	Correlation with Science Achievement	Rank
A	–0.092	0.26	6	–0.010	0.21	6	–0.029	0.20	6
B	0.457	0.23	3	0.250	0.26	4	0.410	0.24	5
C	0.432	0.26	4	0.388	0.18	3	0.528	0.19	3
D	0.307	0.30	5	0.222	0.31	5	0.447	0.30	4
E	–0.509	0.31	8	–0.513	0.31	8	–0.541	0.28	8
F	–0.234	0.32	7	–0.376	0.32	7	–0.398	0.32	7
G	0.474	0.03	2	0.536	0.11	2	0.551	0.11	2
H	0.594	0.21	1	0.565	0.17	1	0.697	0.19	1

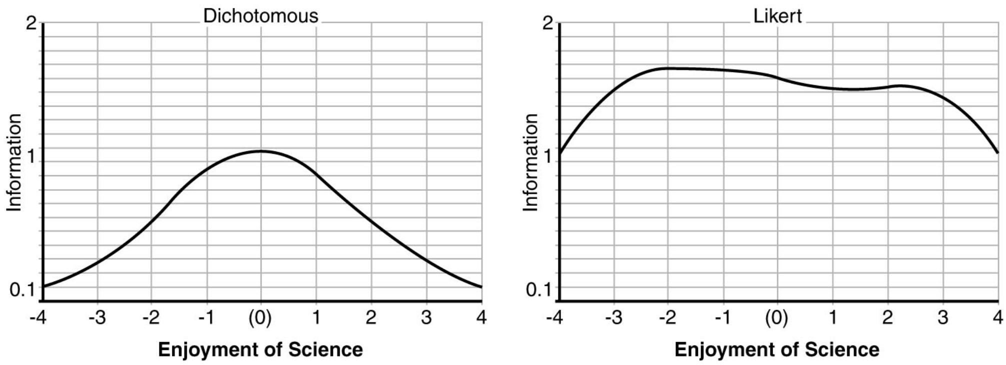

Figure 4. Test information curves for the dichotomous and Likert variants of the enjoyment of science construct

Hierarchical analysis of Rasch models

Having examined the constructs with the basic partial credit model, the following analyses use a multi-faceted Rasch model (Linacre, 1994) to examine variations across countries in the item difficulty and category parameters. Table 4 presents the facet and ConQuest model terms that are employed in the analyses and a description of the effects referred to by each term or parameter.

A series of models was tested hierarchically to find the most parsimonious one. In this procedure, the pooled sample is first calibrated with a complex model, then effects are removed from the model one at a time, creating increasingly simple models. The deviance statistics that result from each calibration are compared. If the difference in deviation is statistically significant, the data better fit the more complex of the two models being compared. Table 5 presents the seven different models that were compared. The results of the statistical comparisons appear in Table 6. The first column in Table 6 indicates the two models being compared. The final column indicates that all comparisons were statistically significant. As such, Model 1, the most complex model, is significantly better fitting than all other models.

Figure 5 illustrates the good fit of Model 1 by plotting the observed data from Countries A and B overlaid on the modelled cumulative probability

Table 4

Description of ConQuest model terms

ConQuest model term	Parameter	Description of the effect being modelled
ITEM	δ_i	A general item effect
STEP	τ_k	A general item response category effect
CNT	α_c	A general country effect
ITEM*STEP	τ_{ik}	An effect of the interaction between the item and the item response category
ITEM*CNT	β_{ic}	An effect of the interaction between the item and the country
ITEM*CNT*STEP	γ_{ick}	An effect of the interaction between the item, the country and the item response category

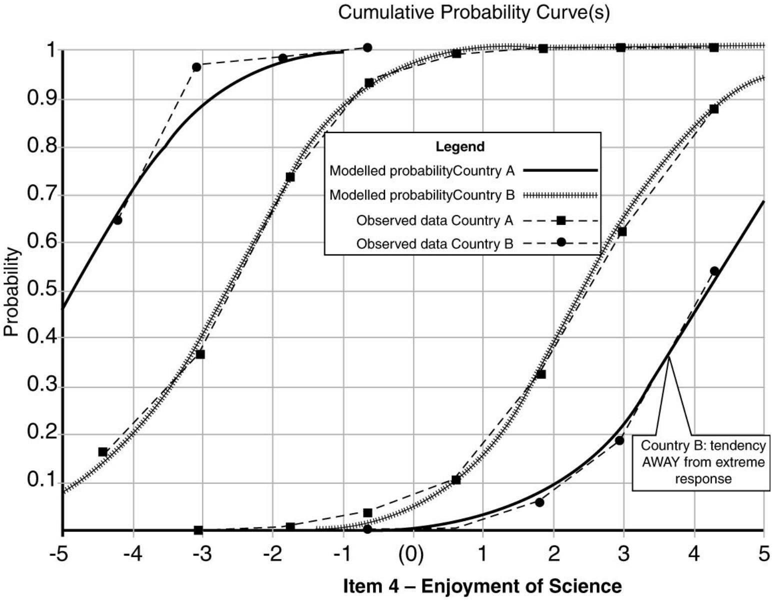

Figure 5. For Countries A and B, the observed cumulative probability curves for item 4 of *enjoyment of science*, from analysis that used Model 1 of the multi-faceted Rasch models.

curve for item 4. Note that for the purposes of this paper it is not necessary to examine the probability curves relating to the middle score category (i.e. the probability of *agree* or beyond) and they have been omitted from Figure 5 to reduce clutter (so only the probability curves for the extreme score categories remain in Figure 5). The data fit the complex model in Figure 5 much closer than the simple model in Figure 3.

In summary, the data better fits a model that includes an interaction effect between the country and the item and the item response category (delta). This effect term will be used below to describe the data in terms of extreme response bias.

Examination of delta parameter estimates

Having determined the best fitting item response model, the final analysis reported in this paper is an examination of the delta parameter estimates for the Likert style items. As reported above, the hierarchical analysis of models revealed that the most complex provided a significantly better fit than any of the less complex models examined. The best fitting model allows for interactions between the delta parameters and both the country effect and the main item effect. In other words, for each item, the distance between categories (from *strongly disagree* to *disagree*, from *disagree* to *agree*, and from *agree* to *strongly agree*) is influenced by a country effect.

It can be reasoned, then, that countries with a tendency towards extreme responses can be identified by examining these delta parameters. Table 7 presents the delta parameter estimates for Countries A and B, for item 4.

Table 7 also shows the difference between deltas 1 and 3. The larger the difference in these delta parameters, the less tendency there is for the

Table 5

Models hierarchically compared in ConQuest for Likert style variants of constructs

	Facet model formulation	ConQuest model term	Deviance	Number of parameters
Model 1	$\ln\left(\frac{p_{nic,k-1}}{p_{nic,k}}\right)=\theta_n-\alpha_c-\delta_i-\beta_{ic}-\gamma_{ick}$	CNT + ITEM+ ITEM*CNT + ITEM*CNT*STEP	79612.7	121
Model 2	$\ln\left(\frac{p_{nic,k-1}}{p_{nic,k}}\right)=\theta_n-\alpha_c-\delta_i-\beta_{ic}-\tau_{ik}$	CNT + ITEM + ITEM*CNT + ITEM*STEP	80927.1	51
Model 3	$\ln\left(\frac{p_{nic,k-1}}{p_{nic,k}}\right)=\theta_n-\alpha_c-\delta_i-\tau_{ik}$	CNT + ITEM + ITEM*STEP	82846.5	23
Model 4 [the Partial Credit Model]	$\ln\left(\frac{p_{nic,k-1}}{p_{nic,k}}\right)=\theta_n-\delta_i-\tau_{ik}$	ITEM + ITEM*STEP	83040.8	16
Model 5	$\ln\left(\frac{p_{nic,k-1}}{p_{nic,k}}\right)=\theta_n-\delta_i-\tau_k$	ITEM + STEP	83374.7	8
Model 6	$\ln\left(\frac{p_{nic,k-1}}{p_{nic,k}}\right)=\theta_n-\alpha_c-\delta_i-\tau_k$	CNT + ITEM + STEP	83183.5	15
Model 7	$\ln\left(\frac{p_{nic,k-1}}{p_{nic,k}}\right)=\theta_n-\alpha_c-\delta_i-\beta_{ic}-\tau_k$	CNT + ITEM + ITEM*CNT + STEP	81228.4	43

Table 6

Results from Chi-square analyses to test for differences between seven multi-faceted Rasch models

Models tested for difference	Chi Square	Degrees of freedom	Significance
2 – 1	1314.4	70	$p < 0.001$
3 – 2	1919.4	28	$p < 0.001$
4 – 3	194.3	7	$p < 0.001$
5 – 4	333.9	8	$p < 0.001$
5 – 6	191.2	7	$p < 0.001$
6 – 7	1955.1	28	$p < 0.001$
4 – 6	142.6	1	$p < 0.001$
2 – 7	301.3	8	$p < 0.001$

Table 7

For Countries A and B, delta parameter estimates for the fourth Likert item in enjoyment of science

	Country A	Country B
Delta parameter 1 (disagree and beyond)	–2.538	–4.263
Delta parameter 2 (agree and beyond)	–0.009	–0.341
Delta parameter 3 (strongly agree)	2.546	4.603
Difference between delta parameter 3 and delta parameter 1	5.084	8.866

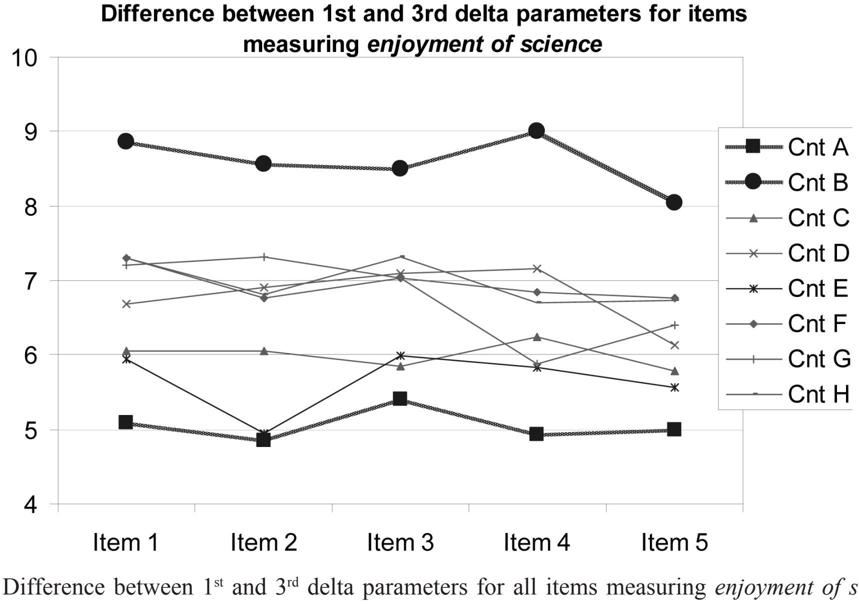

Figure 6. Difference between 1st and 3rd delta parameters for all items measuring *enjoyment of science*, by country

respondents within that country to opt for extreme categories. This is because, in this model, any general effects of the country and the item, and any interactive effect between the country and the item are already accounted for. Thus, only the country by delta interactions are represented by these parameter estimates.

With this in mind, the difference between deltas 3 and 1 can be plotted for each country for all items measuring the construct (Figure 6).

This figure reveals consistent patterns of extreme response tendencies across all items measuring *enjoyment of science*. For Country A, differences between the first and third delta parameters were consistently low for all items (with a magnitude of about five), indicating a tendency towards extreme responses. Country B on the other hand has, for all items, high differences between the first and third delta parameters (magnitudes between eight and nine) indicating a tendency away from extreme responses. This consistency demonstrates that tendency towards extreme response was not item specific, at least within the set of items measuring *enjoyment of science.*

Conclusion

The original question prompting this examination was: if there is extreme response bias exhibited differentially across cultures when answering Likert type items, what can be done to ameliorate this in cross-national survey research? The proposal was that using dichotomous items would *ipso facto* eliminate extreme response bias. When scaled with simple Rasch models, the patterns of correlation between science achievement and *enjoyment of science* were very similar across the Likert variant, the dichotomous variant, and the Likert variant treated as though it were dichotomous, for each country. However, when test information is considered, and other things being equal, a preference for the Likert variant emerges because of the increased information about those in the population distributed further from the mean latent trait score.

The subsequent hierarchical test of the different facets models, however, indicated that a more complex Rasch model, incorporating an interaction between the country and the delta parameter, better fit the data. Analysis of the Likert type items with this more complex model illustrated consistent patterns of what *might* be interpreted as extreme response bias, country by country.

An important point is the degree to which one interprets response tendencies as *extreme*. Rather, it could be argued that countries labelled in this paper as having *low tendency towards extreme response bias* could better be interpreted as having *high tendency towards central response bias*.

The examination presented here is limited, examining only one construct trialled in parallel Likert and dichotomous forms in the PISA 2006 field trial. Further analyses could be undertaken to measure the latent correlations between these constructs, measured and treated differently, and other constructs from both the background questionnaires and the achievement booklets. Having a wider range of constructs trialled in parallel form, particularly ones not specifically related to the topic of science, may reveal different findings. In fact, until a wider range of constructs is investigated, a cultural tendency towards extreme response bias is not established: the effect reported upon here might only be related to the *enjoyment of science* trait.

What can be concluded from the current examination is that tendencies towards extreme response bias may be more prevalent in some countries than others, at least for some attitudinal constructs. On balance it appears prudent to continue the use of Likert type items for international comparative background measures. Researchers would be wise to closely investigate all constructs, measured by Likert type items, on a country by country basis. If extreme response bias is believed to be present, then treating the items as dichotomous in analyses ameliorates this. The choice of analytical model should also be carefully considered. One may be tempted to build facets into the model which allow for (i.e. do not constrain) the interaction between country and the item delta parameter. Importantly however, while the latter option may appeal to the researcher as a rigorous analytical method as the data fits better, this approach masks cultural effects — sometimes the very subject of interest when examining student outcomes. In other words, incorporating country effects into the model results in scales which compensate for some of the differences between countries: people are measured differently depending on their country, but placed on the same reporting scale. So it would be difficult to communicate to those not familiar with such techniques why it is that a student in one country who ‘agrees’ with all statements aimed at measuring a latent construct receives a different final estimate of that trait than a student from another country with exactly the same response pattern. If extreme response bias is found to be significant and consistent it may be wise to report two sets of scores: one adjusted for the bias and one unadjusted.

However, if further research reveals that culturally specific response biases exist and are independent of the latent trait being measured, then there is a much greater argument for building this bias into the model. As noted earlier, bias has not been demonstrated by the analyses in this paper, only hinted at.

References

Choi, Y., Mericle, A., and Harachi, T. W. (2006). Using Rasch analysis to test the cross-cultural item equivalence of the Harvard trauma questionnaire and the Hopkins symptom checklist across Vietnamese and Cambodian immigrant mothers. *Journal of Applied Measurement, 7*(1), 16-38.

Flaskerud, J. H. (1988). Is the Likert scale format culturally biased? *Nursing Research, 37*(3), 185-186.

Grisay, A. (2002). Translation and cultural appropriateness of the test and survey materials. In R. J. Adams and M. Wu (Eds.), *PISA 2000 technical report* (pp. 57-70). Paris: OECD Publications.

Heine, S. J., Lehman, D. R., Peng, K., and Greenholz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference group effect. *Journal of Personality and Social Psychology, 82*(6), 903-918.

Kirsch, I., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J., and Monseur, C. (2002). *Reading for change: Performance and engagement across countries. Results from PISA 2000.* Paris: OECD Publications.

Lee, J. W., Jones, P. S., Mineyama, Y., and Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in Nursing and Health, 25*, 295-306.

Lie, S., and Turmo, A. (2005). *Cross-country comparability of student's self-reports: Evidence from PISA 2003*. Oslo, Norway: University of Oslo.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago: MESA Press.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., and Chrostowski, S. J. (2004). *TIMSS 2003 international science report*. Boston: TIMSS and PIRLS International Study Centre.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., and Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report*. Boston: TIMSS and PIRLS International Study Centre.

OECD. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris: OECD Publications.

OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: OECD Publications

Paulhus, D. L. (1991). Measurement and control of response bias. In *Measures of Personality and Social Psychological Attitudes* (Vol. 1, pp. 17-59). San Diego: Academic Press.

van Herk, H., Poortinga, Y. H., and Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 5*(3), 346-360.

Warm, W. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-445.

Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalized item response models.* Unpublished masters thesis, University of Melbourne, Melbourne, Australia.

Wu, M. L., Adams, R. J., and Wilson, M. R. (1997). ConQuest: Multi-aspect test software [Computer program]. Camberwell, Australia: Australian Council for Educational Research.