Ameliorating culturally based extreme response tendencies to attitude items:
the use of item response models to explore the alternatives

Maurice Walker
Australian Council for Educational Research
Melbourne, Australia
walker@acer.edu.au

Paper presented to ACSPRI methodological conference
Sydney
11-13 December 2006

Introduction

In international comparative educational research such as that undertaken by the IEA or the 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 data collected at the same time. It has certainly been noticed that associations between some background scales and achievement differs markedly *across* countries (e.g. OECD, 2001; Kirsch *et al* 2002; OECD, 2004; Mullis *et al* 2004, Martin *et al* 2004). There have been a number of hypotheses for the causes of such variation. For example, it has been observed that countries with similar patterns of association between background and achievement measures can be clustered variously by similarities in achievement, latitude, language and culture (e.g. Lie and Turmo, 2005; Kirsch *et al* 2002; Mullis *et al* 2004).

There has also been some speculation that these observed differences may in part be attributed to response biases. Paulhus (1991) 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 (*ibid*) the three main types of response bias are:

1. **social desirability**, or the tendency to provide responses that the respondent believes are those which make him or her 'look good' (*ibid* p.17);

2. **acquiescence**, or the tendency to agree rather than disagree with any statement; and

3. **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 have noted cross-cultural differences in response styles (e.g. Choi *et al* 2006; Heine *et al* 2002), some studies have specifically noted that distinct groups tend to exhibit extreme response bias either within studies or when administered a common instrument (e.g. van Herk *et al* 2004; Lee *et al* 2002). 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 it has been said that the use of such items "is most valid for identifying differences within rather than between groups" (Heine *et al* 2002, p914). 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 (see Heine *et al* 2002 for a concise review of this literature). A 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 '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?

There have been several studies which have examined the relative effects of culturally related response bias using different numbers of response option in Likert type scales (e.g. Lee *et al* 2002). But one 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 acquiesce, extreme response bias becomes a non-issue. With this in mind it was decided to administer parallel versions of item batteries for measuring two constructs in the field trial of the PISA 2006 study (an international comparative educational study surveying the scientific, mathematics, and reading literacy of 15 year olds in about 60 countries).

The following research questions were investigated:

- is extreme response tendency exhibited differentially across cultures when answering Likert type items?
- does item response format influence the key outcomes of comparative studies?
- what can be done analytically to ameliorate extreme response bias?

Method

Two parallel versions of item batteries for measuring two constructs the PISA 2006 field trial were administered randomly: one variant was administered as Likert type items; the other as dichotomous.

For this paper, comparative analyses were undertaken of the results from 8 different countries participating in the PISA 2006 field trial, chosen to provide a range of test languages (each country administered the test in a different language), mean science achievement, and geographic location[1]. This chosen range was deliberately wide so as to emphasise the cross-cultural aspect of the investigation.

The data were collected in 2005 and, as it was a field trial, the sample design was based on convenience although attempts were made to select schools that were representative of the school types and study programmes within that country.

Translation of the items from two source languages, English and French, was subject to a process of double blind translation and reconciliation, followed by independent linguistic verification. This does not preclude the possibility however that countries' results were influenced by linguistic nuances of the items.

In addition to comparing the results of the dichotomous items directly with those from the Likert type items, the examination is enhanced through comparing 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).

[1] As the data are not publicly released, the countries are not identified in this paper.

The two constructs administered were 'enjoyment of science', and 'anxiety of science'. Only the former construct is examined in this paper[2]. The English source versions of the item battery for 'enjoyment of science' appears in Table 1. As Likert type items, the responses categories were *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

Enjoyment of Science	
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 constructs in the PISA 2006 field trial, four questionnaires were randomly administered using a rotated design. Two of the questionnaires contained the dichotomous variants of the items; the other two contained the Likert variants. Table 2 provides the sample sizes.

Table 2: Sample sizes

Country	**'enjoyment of science' Dichotomous**	**'enjoyment of science' 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

Results

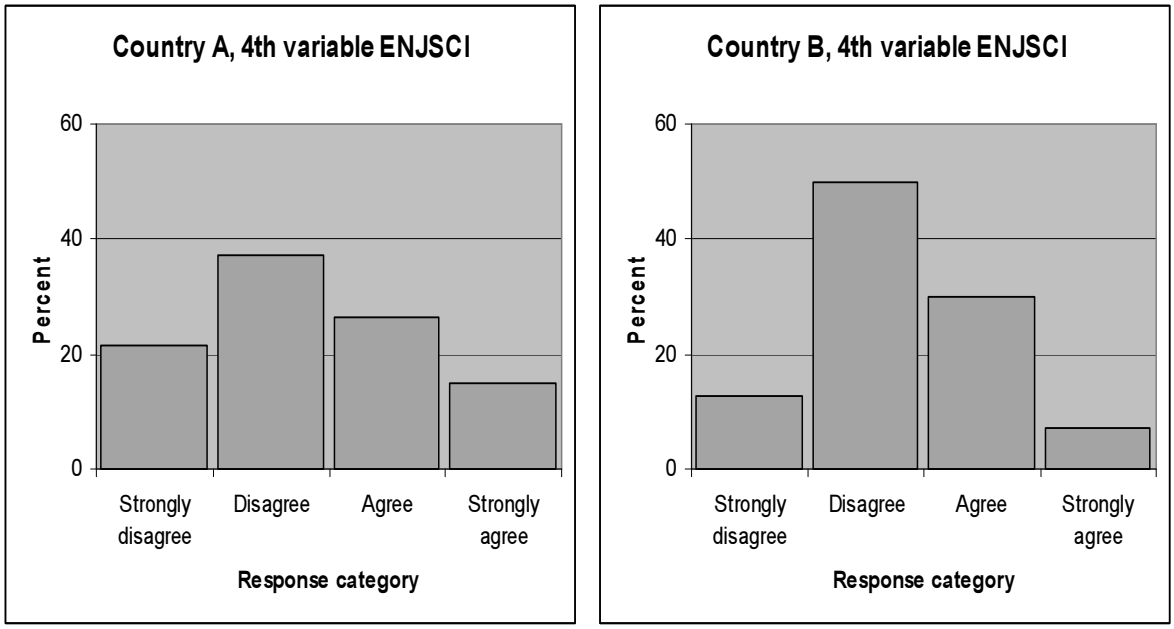
Is extreme response bias exhibited differentially across cultures when answering Likert type items?

Initially the raw frequencies of the responses were examined. Figure 1 presents the proportions of responses for two countries – referred to hereafter as Country A and Country B – for item 4 in the Likert battery. From these simple frequencies it is clear that students 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 in the scale although item 4 was the most extreme case

[2] The 'anxiety of science' construct was found not to be robust and was excluded from the Main Survey. However, all the analyses presented in this paper were replicated for 'anxiety of science' and the results were similar but less definitive. The 'enjoyment of science' construct was chosen for examination in this paper because the results better illustrate the points of discussion.

and therefore it is item 4 that is focussed upon in this paper for illustrative purposes. Similarly, consistent response patterns were seen across the other 6 countries but Country A and Country B were the most extreme and are again focussed upon in this paper for illustrative purposes.

Figure 1: Proportions of responses for the fourth Likert variable in 'enjoyment of science', for Countries A and B



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 and Wright, 1997), which takes the form of

$$P_{x_i}(\theta) = \frac{\exp \sum_{k=0}^{x} (\theta_n - \delta_i + \tau_{ij})}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^{k} (\theta_n - \delta_i + \tau_{ij})} \qquad x_i = 0,1,\ldots,m_i,$$

where $P_{xi}(\theta)$ is the probability of person n to score x on item i. θ_n denotes the person's latent trait, the item parameter δ_i gives the location of the item on the latent continuum and τ_{ij} is an additional step parameter.

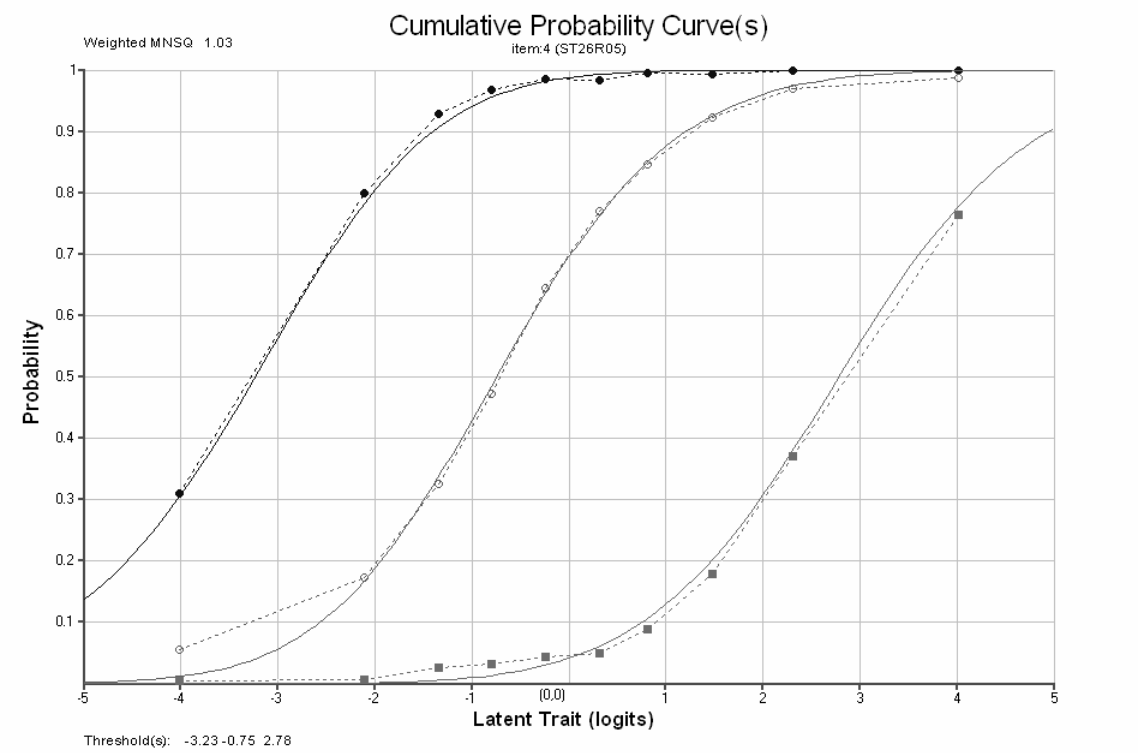
ConQuest software was used for all IRT analyses (Wu, Adams and Wilson, 1997).

Item fit was assessed using the weighted mean-square statistic (infit), which is a residual-based fit statistic. Weighted infit statistics ranged between 0.94 and 1.02 for the item parameters and between 0.90 and 1.00 for the step parameters, which all indicate good fit.

The goodnesss of fit for item 4 can be illustrated by examining the observed data for the pooled sample overlayed on the modelled cumulative probability curve 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 move from a response category to a higher one by the strength of their latent trait (in this case 'enjoyment of science'). Given that the

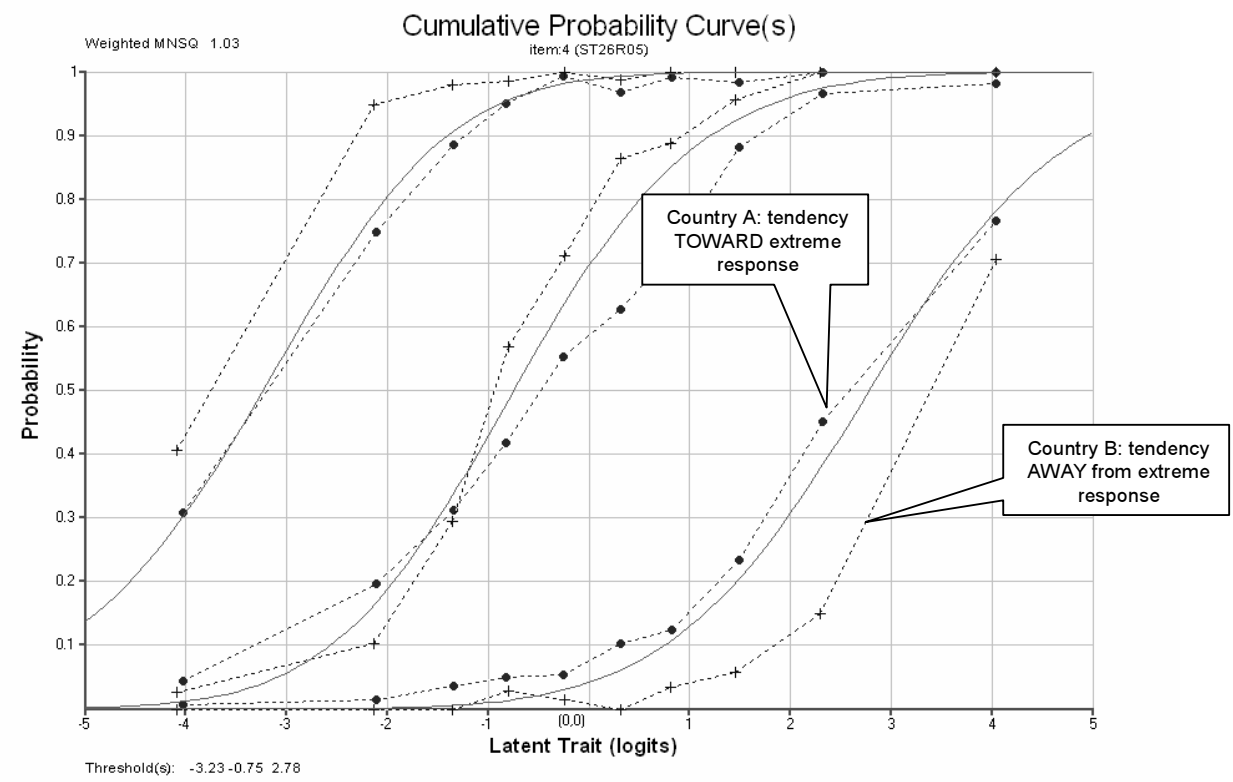
response categories are scored as 0 for *strongly disagree,* 1 for dis*agree*, 2 for *agree,* and 3 for *strongly agree*, the left hand curves represent the probability that a respondent with a particular level of the underlying trait (shown on the x axis) will score at least a 1, the middle set at least 2, and the right hand set, 3.

Figure 2: Cumulative probability score for the fourth Likert variable in 'enjoyment of science', partial credit model and observed data for pooled sample



However when the data was grouped by country and the observed values for each country were plotted alongside the same model, considerable country response differences were revealed. More specifically the country data does not always fit the model. Figure 3 shows the observed data from Countrys A and B overlayed on the modelled cumulative probability curve for the item 4.

Figure 3: Cumulative probability score for the fourth Likert variable in 'enjoyment of science', partial credit model and observed data for Country A and Country B



Of interest are the left and the right hand sets representing the movement between the extreme and middle response categories. Looking at the left hand set it can be seen that a person in Country B, represented by the crossed dotted line, has a much higher probability of moving from strongly disagree to agree (or beyond) than a person in Country A, represented by the dotted line with circles, who has the same level of underlying enjoyment in science. For example, a person in the Country B at -2 logits on the latent trait scale has a probability of around 0.96 that they will move from strongly disagree, whereas this is only about 0.77 for a person in the Country A at -2 logits on the latent trait scale. In other words, it is more likely that a respondent in the Country A will choose a category other than the extreme of strongly disagree.

Looking at the right hand set of curves it can be seen that a person in Country B has a much lower probability of moving to strongly agree than a person in Country A with the same level of underlying enjoyment in science. A person in the Country B at 2 logits on the latent trait scale has a probability of around 0.21 that they will opt for strongly agree, whereas this is about 0.38 for a person in Country A with the same level of latent trait. In other words, it is less likely that a respondent in the Country B will choose the extreme of *strongly agree*.

In summary, the data shows that respondents in the Country B tend to opt for less extreme responses than a person 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 ten variables, 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 Least Estimate for the each country, using the partial credit model for the Likert variant and the Rasch model for the dichotomous variants. A scale ranking of the countries has also been provided. The (attenuated) Cronbach's alpha correlation between the attitude estimates and the score for science achievement (also a WLE) for each country also appears in this table.

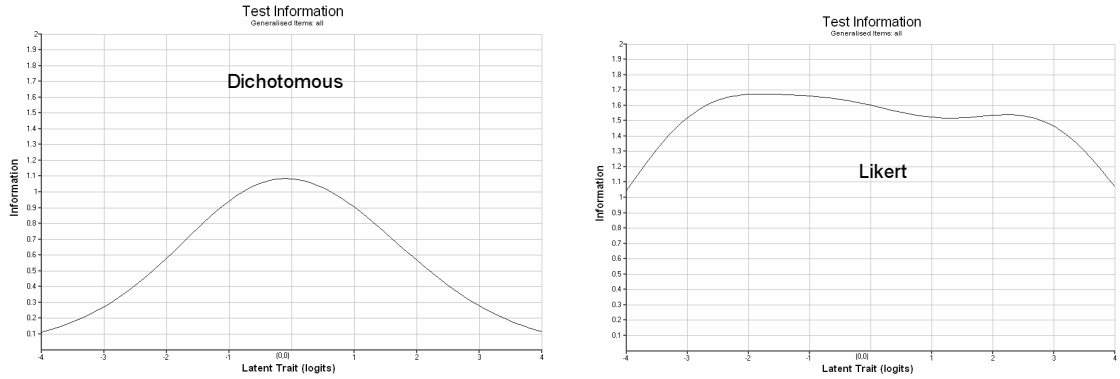
Table 3: Country mean WLE, correlation with science achievement and scale rank for the 'enjoyment of science' construct by method of data collection and treatment

'enjoyment of science'	Dichotomous			Likert			Likert treated as dichotomous		
Country	Mean WLE	Correlation with Science Achievement	Scale rank	Mean WLE	Correlation with Science Achievement	Scale rank	Mean WLE	Correlation with Science Achievement	Scale 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

It can be seen that using simple IRT models, the different methods of data collection and treatment effect only slight differences in the scale rankings, and the correlations with science achievement remain fairly consistent. Thus, on these criteria, no preference clearly emerges for any one variant.

However, because the Likert scale provides relatively more information about the respondents with high or low levels of the latent trait, the Likert variant is preferred. This difference in test information is illustrated in Figure 4. Note that while the information scales can not be directly compared, the shape of the test information curves differ markedly. The dichotomous variant provides relatively less information about the people lying more than one standard deviation from the mean of the scale – those who do not lie between -1 and 1 logit. The test information curve for the Likert variant shows that this scale yields more consistent information across the range of the population – the curve does not drop off until 3 standard deviations from the mean. One reason that the latter type of information curve is preferred is that attitudes such as 'enjoyment of science' are considered educational outcomes and researchers are often interested in examining the characterises students with strong attitudes and the relationship between these attitudinal outcomes and other outcomes such as achievement.

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



Hierarchical analysis of IRT models

Having examined the constructs with the basic Partial Credit model, the following analyses seek to determine the further effects, and their interactions, that might be incorporated into the model to best fit the data. When calibrating the data on the overall pooled sample, for each construct there are three basic effects that can be modelled: the effect of the item; the effect of the item steps (i.e. the step or difference between *strongly disagree* and *disagree*, the step between *disagree* and *agree*, and the step between *agree* and *strongly agree*); and the effect of the country. Interaction between these main effects can also be modelled. Table 4 presents the ConQuest model terms that are employed in the analysis and a description of the effects referred to.

Table 4: Description of ConQuest model terms

ConQuest model term	Description of the effect being modelled
ITEM	A general item effect
STEP	A general step effect
CNT	A general country effect
ITEM*STEP	An effect of the interaction between the item and the step
ITEM*CNT	An effect of the interaction between the item and the country
ITEM*CNT*STEP	An effect of the interaction between the item, the country and the step

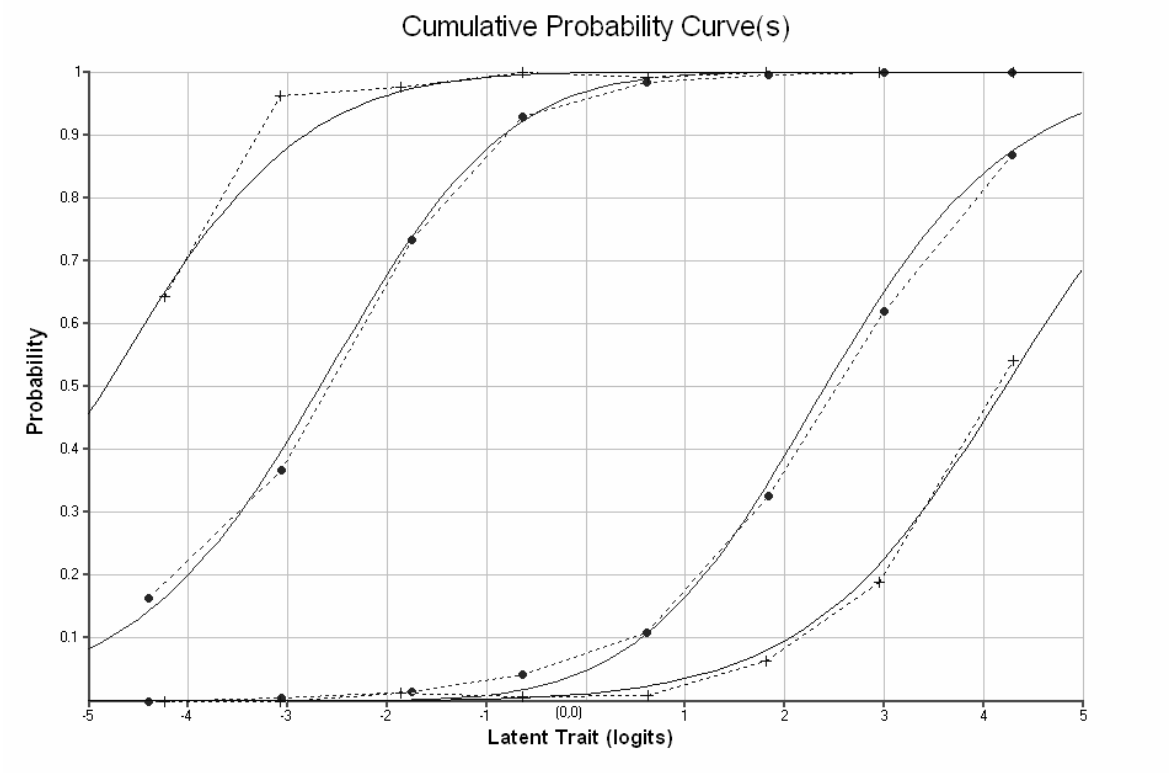
A series of models was tested hierarchically to find the most parsimonious one. In this procedure, one begins by calibrating the pooled sample with a simple model, and proceeds to add effects one at a time, creating increasingly complex models. The deviance statistics that result from each calibration are compared. If the difference in deviation is statistically significant, the data better fits the more complex of the two models being compared. Table 5 presents the seven different models that were compared.

Table 5: Models hierarchically compared in ConQuest for Likert style variants of constructs

	Model statement in ConQuest
Model 1	CNT + ITEM+ ITEM*CNT + ITEM*CNT*STEP
Model 2	CNT + ITEM + ITEM*CNT + ITEM*STEP
Model 3	CNT + ITEM + ITEM*STEP
Model 4 [the Partial Credit Model]	ITEM + ITEM*STEP
Model 5	ITEM + STEP
Model 6	CNT + ITEM + STEP
Model 7	CNT + ITEM + ITEM*CNT + STEP

The results of these analyses reveal that Model 1, the most complex model, is significantly better than all other models (results appear in table A1 in the appendix). Figure 5 illustrates this by plotting the observed observed data from Countries A and B overlayed on the modelled cumulative probability curve for the item 4. Note that the middle set of probability curves have been omitted for clarity in Figure 5 (so only the probability curves for the extreme score categories remain). It can be seen that the data fit the complex model in Figure 5 much closer than the simple model in Figure 3.

Figure 5: Cumulative probability score for the fourth Likert variable in 'enjoyment of science', complex model and observed data for Country A and Country B (scores 1 and 3 only)



In summary, the data better fits a model that includes an interaction effect between the country and the item and the step. It is this effect term that will be used below to help describe the data in terms of extreme response bias.

Examination of step parameter estimates

Having determined the best fitting item response model the final analysis reported in this paper is an examination of the step 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. This suggests that the step parameters need to allow for an interaction both with the country and with the item. In other words, for each item, the distance between steps (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 step parameters.

Table 6 presents the step parameter estimates for Countries A and B, for item 4.

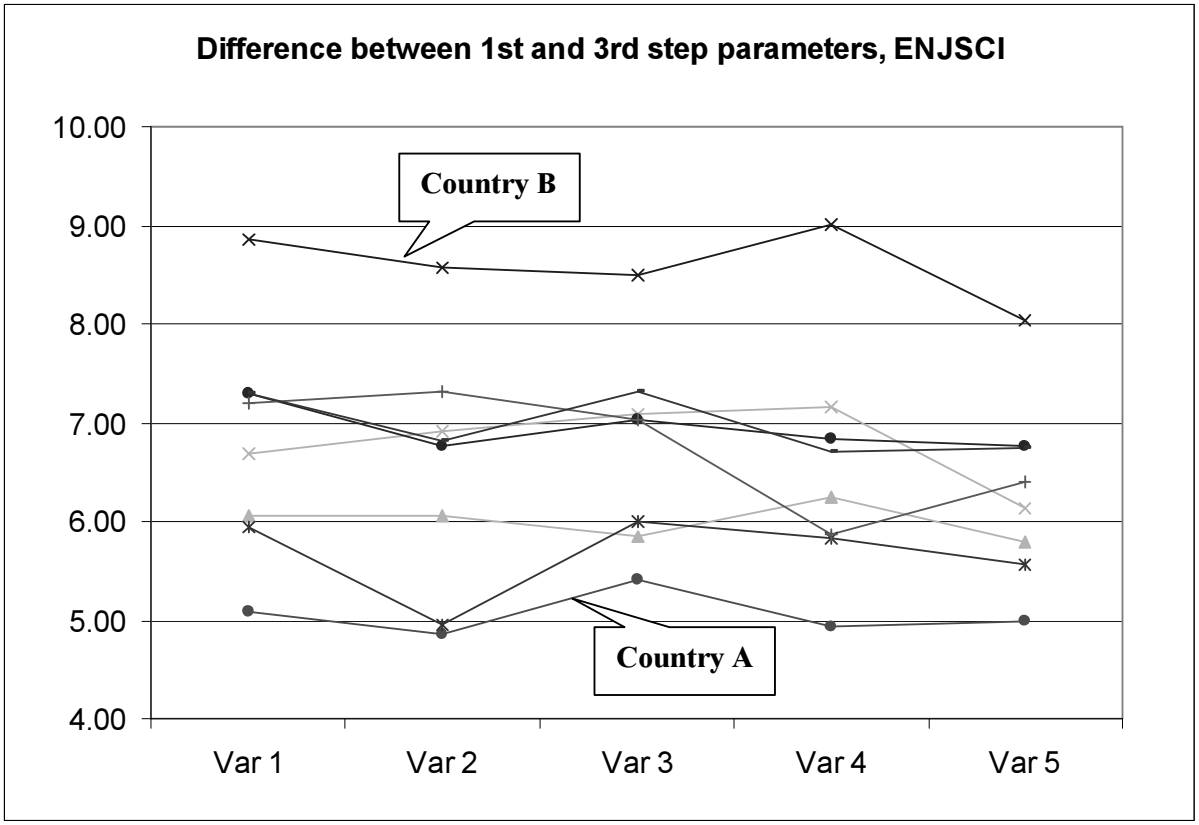
Table 6: Step parameter estimates for the fourth Likert variable in 'enjoyment of science', for Countries A and B

	Parameter estimates	
	Country A	**Country B**
Step 1 (strongly disagree to disagree)	-2.538	-4.263
Step 2 (disagree to agree)	-0.009	-0.341
Step 3 (agree to strongly agree)	2.546	4.603
Difference between Step 3 and Step 1	5.084	8.866

Table 6 also shows the difference between steps 1 and 3. The larger the difference in these step parameters, the less tendency there is for the sample 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 step interactions are represented by these parameter estimates.

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

Figure 6: Difference between 1st and 3rd step parameters, enjoyment of science, by country



This figure demonstrates the fairly consistent patterns of extreme response tendencies for 'enjoyment of science'. Country A, for example, has consistently low differences between the first and third step parameters for all items in the scale, indicating a tendency towards extreme responses. Country B on the other hand has consistently high differences between the first and third step parameters indicating a tendency away from extreme responses.

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 IRT models, the patterns of correlation with achievement and the mean case estimates for each country were very similar across the Likert variant, the dichotomous variant, and the Likert variant treated as though it were dichotomous. 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.

The subsequent hierarchical test of the different facets models, however, indicated that that a more complex IRT model, incorporating an interaction between the country and the step, better fitted the data. Analysis of the Likert style 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. Only two constructs were trialled in parallel Likert/dichotomous forms in the PISA 2006 field trial. One of these constructs, anxiety of science, was not overly robust and hence was not reported in this paper. 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, it can not really be said that a cultural tendency towards extreme response bias really does exist: the effect reported upon here might only be related to the enjoyment and anxiety of science.

What can be concluded from the current examination is that tendencies towards extreme response bias may present more so in some countries than others, at least for some attitudinal constructs. Researchers would be wise to closely investigate all constructs which they may wish to incorporate into explanatory models on a country by country basis. On balance it appears prudent to continue the use of Likert type items for international comparative background measures. If extreme response bias is believed to be present, then treating the items as dichotomous in analyses ameliorates this. Care should also be taken when choosing the analytical model for constructing a scale. 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 step. Importantly however, while the latter option may appeal to the researcher as a rigorous analytical method, essentially such a model masks cultural effects — sometimes the very subject of interest when examining achievement. In other words, incorporating country effects into the model results in scales which essentially compensate for some of the differences between countries: people are measured differently depending on their country, but placed on the same 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.

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.

Appendix

Table A1: Hierarchical test of facets models, 'enjoyment of science'

Model 1	cnt+item+cnt*item+cnt*item*step	79612.7	121
Model 2	cnt+item+cnt*item+item*step	80927.1	51
Model 3	cnt+item+item*step	82846.5	23
Model 4	item+item*step	83040.8	16
Model 5	item+step	83374.7	8
Model 6	cnt+item+step	83183.5	15
Model 7	cnt+item+cnt*item+step	81228.4	43
Models tested for difference	Chi Square	Degrees of freedom	Significance
2 - 1	1314.4	70	0.000000
3 - 2	1919.4	28	0.000000
4 - 3	194.3	7	0.000000
5 - 4	333.9	8	0.000000
5 - 6	191.2	7	0.000000
6 - 7	1955.1	28	0.000000
4 - 6	142.6	1	0.000000
2 - 7	301.3	8	0.000000

References

Choi, Y., Mericle, A. and Harachi, T.W., 2006. Using Rasch analysis to test the sross-cultural item equivalence of the Harvard trauma questionnaire and the Hopkins symptom checklist across vietnames and Cambodian immigrant mothers. In *Journal of Applied Measurement.* V7 n1 pp16-38.

Kirsch, I., de Jong, J., Lafontaine, D., McQueen,J., Mendelovits, J., and Monseur, C., 2002. Reading for change: performance and engagement across countries.

Flaskerud, J.H., 1988. 'Is the Likert scale format culturally biased?'. In *Nursing Research.* V37, n3, pp185-186.

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. In *Journal of Personality and Social Psychology.* V82 n6 pp903-918.

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. In *Journal of Cross-Cultural Psychology* V5 n3 pp 346-360.

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'. In *Journal of Personality and Social Psychology* V82, n6, pp903-918.

Lie, S and Turmo, A., 2005. Cross-country comparability of student's self-reports – evidence from PISA 2003. Paper presented to PISA Technical Advisory Group.

Lee, J.W., Jones, P.S, Mineyama, Y., and Zhang, X.E., 2002. 'Cultural differences in responses to a Likert scale'. In *Research in Nursing and Health*, V25 pp 295-306.

Masters, G. N. and Wright, B. D. (1997). The Partial Credit Model. In: van der Linden, W. J. and Hambleton, R. K. (Eds.). *Handbook of Modern Item Response Theory* (pp. 101–122). New York/Berlin/Heidelberg: Springer.

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.

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.

OECD, 2001. Knowledge and skills for life: first results from PISA 2000. Paris: OECD.

OECD, 2004. Learning for tomorrow's world: first results from PISA 2003. Paris: OECD.

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

van de Vijer, F., and Leung, K., 1997. 'Methods and data analysis for cross-cultural research. In J.W. Bery, Y.H. Poortinga, and J. Pandey (Eds.) Handbook of cross-cultural psychology (2^{nd} ed). Vol 1, pp 257-300.

Wu, M.L., Adams, R.J., and Wilson, M.R. (1997). ConQuest: Multi-Aspect Test Software [computer program manual]. Camberwell: Australian Council for Educational Research.