Comments on *Kreiner 2011: Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment*

Ray Adams, April 19 2011[i]

This paper is concerned with two issues:

- Do the outcomes of PISA – more particularly the rankings – depend on the items chosen?
- Could alternative scaling procedures, produce different outcomes – more particularly the rankings.

A careful examination of these two issues is clearly an important activity that should be the focus of rigorous scholarly investigation. Unfortunately this article presents neither a rigorous nor a scholarly investigation.

In making comments on the article I shall organise them into two sections. First, I shall make some general points about the work that cover the implications for PISA. Second, I shall make some more specific points that relate to the paper's acceptability as an academic piece of work.

General Comments

The fundamental flaw in Kreiner's argument is that he confounds the two primary issues of:

(a) Do the outcomes of PISA depend upon the set of items that are developed and chosen?; and
(b) Does the use of the Rasch model provide misleading results because the data do not *fit* the Rasch model?

The PISA Consortium has undertaken considerable work on both of these issues and the evidence demonstrates that while the items matter enormously, the model matters to a much lesser degree.

In what ways does item choice matter?

One of the goals of PISA is to compare the outcomes of education systems. A fundamental issue, therefore, is what should form the basis of that comparison. PISA has taken the position that the basis of the comparison should be a relatively large basket of items that are representative of the knowledge, skills and understandings that reflect the respective content framework and are valued by the set of countries that participate as well as leading educators. The procedures for preparing this basket of items have been fully articulated in Chapter 2 of each of the PISA technical reports and collections of sample items have been regularly published, for example OECD, 2009a, 2009b.

There are many legitimate discussions to be held concerning the size of the basket of items and the details of the content, but Kreiner raises none of them, save for *fit* to the Rasch model, which is discussed later.

Kreiner takes four small subsets of reading items – three sets of 6 items and one set of 9 items -- and compares the ranks of the countries using the 2006 data. He provides no basis for the selection of the items and he undertakes no technically appropriate statistical testing of the differences in the ranking. The outcome that he reports, not surprisingly, is that country ranks vary across the four small sets.

Is Kreiner really suggesting that he would expect and hope that the outcomes of PISA would not depend upon the items that are chosen? Does it not make sense that there should be consistency between the aspects of reading competence valued by an education system and the aspects of reading required by the PISA items? Does it not make sense that education systems might well be expected to perform relatively worse on some items and relatively better on others as a function of the societal, cultural and educational experiences of their students.

The option of restricting PISA to only those items that do not respond to the specific strengths and weaknesses of educational systems is not one that has been chosen. Adopting such an approach would result in an assessment that, while perhaps objectionable to few, would be of interest to nobody. Such an assessment would not include, as criteria, many of the outcomes valued by participating education systems and leading educators.

Considerable research has been undertaken on the impact of item choice on international comparisons (Adams, Berezner & Jakubowski 2010; Beaton & Gonzales 1997; Beaton 1997; Hencke *et al.* 2009) – Kreiner does not refer to any of this research. The findings of these studies show that it is indeed possible to manipulate the outcomes of international studies through post hoc manipulations of the item pool. The post hoc manipulations require examining the data to find small sets of items advantageous or disadvantageous to specific participants and then selecting those items as the basis for comparison. But, more importantly, the studies show that when subsets of items are selected from the item pools without knowledge of actual student performance on the items, the rankings on the subsets, and other headline results, are not influenced. That is, without an examination of the data, national representatives and reviewers are not able to identify subsets of items that will disadvantage or advantage specific countries.

Conclusions from two of these studies illustrate this point. Adams *et al*. (2010) conclude "*Two lessons can be drawn from this analysis. First, the pool of test items has to be big enough to provide robust comparisons of countries and to accommodate diverse preferences. PISA uses around 100 items for testing major domains and around 30 items for minor domains, which seems to be sufficient to limit the impact of individual countries' views. Second, although this analysis suggests that final rankings can depend upon the choice of items, experts cannot identify items in advance that will advantage or disadvantage their country. This paper shows that this is the case in PISA, because countries in general do not gain or lose from considering their preferred items only. In other words, experts are not able to predict which items can increase their country's chances of improving its ranking position in the final PISA test.*" (p 12). Similarly, Hencke et al. (2009) conclude "*Our findings suggest that a high degree of confidence* *can be placed in the estimated scale scores for all countries assessed during TIMSS 2003 regardless of item selection.*" (*p 111)*

Lack of model fit and focus on significance

Kreiner's line of argument concerning the use of the Rasch model is strongly based upon tests of statistical significance rather than the substance of the effects detected. As Box (1979) reminds us no statistical model will fit data perfectly, but some statistical models are useful. The majority of Kreiner's findings could be summarised with the simple observation that PISA has a large sample.

The sample sizes in PISA are such that the fit of any scaling model, particularly a simple model like the Rasch model, will be rejected. PISA has taken the view that it is unreasonable to adopt a slavish devotion to tests of statistical significance concerning fit to a scaling model.

The more fundamental question is whether the scaling approach that has been adopted is useful. There is nothing in Kreiner's paper that speaks to the issue of the utility of the scaling approach that has been used or the implications of it use.

Kreiner's alternatives

As an alternative to the Rasch model employed in PISA, Kreiner first poses a more general Rasch-type model that permits *dependence* and *DIF* terms – item dependence and DIF are two misfits to the Rasch model that he discusses. At this point he fails to make clear that the dependence term he adds has no influence on the rankings and that the DIF term means he is reducing the set of items upon which comparisons between the two countries in the analysis are being made to the subset of items without DIF terms. But, perhaps, that is not so important since he quickly notes that it will not be feasible to apply such an approach to PISA: "…the search for a GLLRM for all items was abandoned".

So, in a second attempt at a solution, Kreiner suggests data purification, whereby one attempts to identify and eliminate items that do not fit the scaling model and then applies the model only to those items that fit. He finds using his methodology that eight out of 24 reading items can be used to compare the United Kingdom and Denmark and concludes that there is no difference between the scores of the United Kingdom and Denmark on those eight items. Perhaps he should also have reported that the published PISA 2006 reading results, based upon all items and all students, are United Kingdom, mean 495 with a standard error of 2.3 and Denmark mean 494 with a standard error of 3.2. The United Kingdom is estimated to rank between 11 and 17 in the OECD and Denmark is estimated to rank between 11 and 16.

Has Kreiner truly offered a better and viable alternative? For some reason that goes unmentioned in his paper where he analyses data from less than one third of the students who responded to reading items and data from less than one thirteenth of all of the PISA students from the United Kingdom and Denmark. Secondly, he limits himself to just two countries and says nothing of the possibilities or implications of extending his approach to more than 50 participants. Third, his final comparison relies upon just eight of the 28 questions that countries and expert panels have agreed are required to cover the range of competencies that are valued by PISA. In fact it is remarkable that his comparison of the United Kingdom and Denmark produces a result identical to that presented in the PISA 2006 report.

Rasch Theory

The Rasch Model is indeed a very special model that has a range of properties that support a very powerful form of measurement. Kreiner clearly has a deep understanding of and respect for these properties.

In practice, however, there is often a discrepancy between the behaviour of the observed data and the ideals of the model. With samples as large as those in PISA, it is easy to be confident that there are indeed discrepancies between the model and the data.

When discrepancies are observed between an adopted model and observed data a number of alternative courses of action are possible. First, one can reject data and retain only those data that are compatible with the model. Second, one can reject use of the model and move to an alternative model. Third, one can proceed with use of the model on the assumption that the results will still have utility even if the data and model are not fully compatible.

The first approach is what Kreiner has illustrated in his paper. He has purified the data and retained only items that conform to the Rasch model. The outcome was a comparison based upon just eight items, clearly an extreme approach that is not compatible with PISA being a comprehensive assessment of the construct. Commentators such as Goldstein (Goldstein, 2004; Goldstein, Bonnet and Rocher, 2007) would be strongly opposed to such data censoring. The consequential compromise to the validity of the comparisons is simply not tenable.

Second, one could explore alternative models. PISA is more than happy for due consideration to be given to alternative models. Kreiner, has explored one approach but he himself conceded that it was not viable. Goldstein, Bonnet and Rocher (2007) show how alternative approaches can be used for certain analytic purposes. Other large scale studies, for example the TIMSS study since 1999, use the so-called three parameter logistic model. The two- and three-parameter logistic models are more general than the Rasch model, but using Kreiner's criteria they still do not fit PISA data. The two-parameter model, which as Kreiner mentions permits differing item discriminations, has been applied to PISA data and it has been shown that the outcomes are identical to those obtained when fitting a Rasch model (Macaskill, 2008). Further, the dependency between PISA items that Kreiner mentions has also been modelled, and no implications for the rankings have been observed (Macaskill, 2008).

When exploring alternative models one must take into account the full PISA context. In doing so some factors that need to be taken into account when considering alternative models include:

- The need to comprehensively cover the constructs
- The need for analytic techniques that work simultaneously for more than 50 countries;
- The need to integrate with appropriate sampling methodologies;
- The requirement to provide a database that is accessible to and usable by secondary data analysts;
- The need to support the construction of described proficiency scales;

- The need to permit inter-country comparison of overall levels of performance;
- The need to support the construction of scales that retain their meaning over time.

The third option is to proceed with the Rasch model acknowledging that it does not fit – and no model will – but undertaking work to ensure that the violations of the model are properly recognised in terms of their impact on the outcomes. The work on item choice and alternative scaling models described above is exactly that kind of work.

It should also be noted that PISA does implement a form of data purification through its field trial process. Before use in a main survey all PISA items are tested in all participating countries and reviewed for compatibility with the Rasch model. Furthermore, at this stage items with the most substantial DIF are excluded from further consideration and not used in the main PISA surveys.

Whenever it can, PISA reviews the impact of its scaling assumptions on the outcomes. To this date it has been found that the inferences that have been made on the basis of the Rasch model are sound. At the same time others are encouraged to undertake rigorous scientific work that examines the PISA scaling approach, reviews the extent to which it is fit for purpose, and considers alternatives.

Specific Comments

In this next section I make some more technical and specific observations and ask some questions. I do not cover all the errors in the manuscript.

If asked to review this article for a professional journal, I would advise that the article should be rejected on technical grounds.

1. There is insufficient evidence that the author is aware of the literature that is relevant to the various aspects of the paper, and there is no discussion of the literature. For example assertions such as "*Different dimensions of mathematical attainments are known to be very strongly correlated and international surveys measuring such traits should result in similar results if measurements are valid*" are made without any references. Adams, Berezner & Jakubowski 2010; Beaton & Gonzales 1997; Beaton 1997; Hencke *et al.* 2009 would all appear to be relevant to this section.

There is no presentation of the literature related to scaling in these large scale contexts. Apart from the PISA technical reports which go through the scaling step by step a discussion of the literature on scaling in large-scale assessments would appear relevant. Such literature includes Adams & Wu (2000), Adams, Wu & Carstensen (2000), Beaton (1987), Mislevy *et al.* (1992), and Thomas (2000).

The alternative model (1) is presented and posited without a discussion of its position within the literature. For example, there is no acknowledgement that this model can be fit by the scaling software that is currently used for PISA.

The reliance on Kreiner & Christensen (2011a & 2011b) is unfortunate since they are yet to be published. If the methods described in those papers are to be heavily exploited then they must be accessible to the reader.

A number of comments show that the PISA literature, particularly that on DIF, has not been fully explored. For example a list of references on DIF in PISA includes: Le (2006a, 2006b, 2007, 2009a, 2009b) and Grisay et al. (2007).

Which dataset has been downloaded for use? A data source and download date needs to given.

2. An explanation is needed as to why just 28,593 of the available cases is analysed. If I am interpreting Figure 1 correctly, about 210,000 students responded to reading items. Why aren't all of these students used?

3. Can't the methods proposed deal with missing data? This would seem a potentially major issue that needs to be discussed.

4. There are many errors of fact, below are some of them

- The statement "not all items were administered in all countries" just below Figure 1 is an error of fact. All items were administered, and after the fact a very small number of items are omitted due to extreme DIF. All of these omitted items are listed in the technical report.

- "but it is obvious that items have been administered and/or scored in different ways in different countries" is erroneous. Much effort is expended in ensuring that this is *not* the case, including verification of translated coding guides, international coder training meetings and a coder query service by email (see chapter 2 of the technical reports).

- The infit statistic given in (3) is not that used in PISA. The technical report gives the reference Wu *et al.* (2007) to the correct formula.

- Some of the commentary about consistency is incorrect for MML, but it is hard to comment further without access to the Kreiner and Christensen articles.

5. How were the four item subsets that were used formed? Are they random subsets or were they chosen according to some criteria? It is noted that items from different units are split between these sets, which could never happen in practice.

The rankings computed use a raw score methodology. That's fine, but why not do it properly and get the standard errors correct? Three things that need to be taken into account are: sampling weights, sampling design effects and the item clustering - these are not mentioned in the appendix.

6. There is a lack of focus in the paper on the main issue of DIF other topics appear rather randomly. For example, what is the relevance of the material on plausible values?

References

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficient multinomial logit model: A generalized form of the Rasch model. In M. v. Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 57 – 76): Springer Verlag.

Adams, R. J., Wu, M. L., & Carstensen, C. H. (2007). Application of multivariate Rasch models in international large scale educational assessment. In M. v. Davier. C. H. Carstensen (Ed.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 271-280): Springer Verlag.

Adams, R., Berezner, A., Jakubowski, M. (2010) *Analysis of PISA 2006 preferred items ranking using the percent-correct method.* OECD Education Working Papers, No. 46.

Beaton, A. E. (1987). *Implementing the new design: The naep 1983-84 technical report* (report no. 15-tr-20). Technical report, Educational Testing Service.

Beaton, A. E. (1997). Comparing cross-national student performance on TIMSS using different test items. *International Journal of Educational Research, 29(6)*, 529-542

Beaton, A., & Gonzalez, E. (1997). Reporting achievement in mathematics and science content areas. In M. Martin & D. Kelly (Eds.), *Third International Mathematics and Science Study technical report: Vol. II. Implementation and analysis* (pp. 175–185). Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In *Robustness in Statistics* (Edited by R. L. Launer and G. N. Wilkinson), 201-236. New York Academic Press.

Goldstein, H. (2004) International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education*, 11, 319-330.

Goldstein, H., Bonnet, G. and Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioural Statistics 32:* 252-286.

Grisay, A., de Jong, J.H.A.L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007) Translation Equivalence across PISA Countries. *Journal of Applied Measurement 8(3)* 2007, 249-266.

Hencke, J., Rutkowski, L. Neuschmidt, O., and Gonzalez, E. (2009). *Curriculum coverage and scale correlation on TIMSS 2003*, IERI Monograph, 2, 85-112

Le, L (2006a). *Investigating Gender Differential Item Functioning Across Countries and Test Languages for PISA Science Items.* Paper presented at 5th Conference of International Test commission, Brussels, July 2006.

Le, L.T. (2006b). *Analysis of Differential Item Functioning.* Paper presented at the annual meeting of American Educational Research Association, San Francisco CA.

Le, Luc T. (2007). *Effects of item positions on their difficulty and discrimination - A study in PISA Science data across test language and countries.* New Trends in Psychometrics. Proceedings of Conference of Psychometric Society, Tokyo, 2007.

Le, L.T. (2009a). Investigating Gender Differential Item Functioning across Countries and Test Languages for PISA Science items. *International Journal of Testing*, 9, 2, 122-133.

Le, L.T. (200b9). Effects of item positions on their difficulty and discrimination- A study in PISA Science data across test language and countries. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics*, Tokyo: Universal

Macaskill, G (2008). PISA TAG(0809)6a_1.doc: *Alternative Scaling Models and Dependencies*. Available from mypisa.acer.edu.au.

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992a). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*:133

OECD (2009a). *PISA 2006 Technical Report*, OECD, Paris.

OECD (2009b). *PISA Take the Test: Sample Questions from the OECD's PISA Assessments.* OECD, Paris.

Thomas, N. (2000). Assessing model sensitivity of the imputation methods used in the National Assessment of Educational Progress. *Journal of Educational and Behavioral Statistics, 25*:351-371.

Wu, M. L., Adams, R. J., Wilson, M. R., Haldane, S.A. (2007). *ACER ConQuest Version 2: Generalised item response modelling software* [computer program]. Camberwell: Australian Council for Educational Research.

[i] Several individuals have made valuable comments on this document. They include Keith Rust, Rolf Olsen and a number of my ACER PISA colleagues.