Paper prepared for the 35th Annual Conference
International Association for Educational Assessment
13–18 September 2009
Brisbane, Australia

A new role for item aficionados in high-stakes testing programs

Gabrielle Matters
Australian Council *for* Educational Research

INTRODUCING THE ITEM AFICIONADO

Devotee, enthusiast, adherent, fanatic, addict, admirer to/of the rituals of test design and item analysis

The purpose of this paper is to point teachers, test analysts, and users of test results to the significance of student responses at the item level and considering what it is that each item purports to measure and actually measures before taking the evidence of a low score on a test – just a score derived from a collection of items – and coming to the seemingly obvious but not necessarily accurate conclusion that the student has no knowledge or understanding of the domain being tested. An item aficionado does not approach test items and test results at the level of abstraction of ability. Rather, the role is to identify what students can do and what they cannot do or have trouble with, to understand sources of item difficulty, and to critique tests.

This new role is pertinent to specific forms of external standardised tests such as the 'Programme for International Student Assessment' (PISA) (see <http://www.pisa.oecd.org>), which has gained significance in all corners of the globe. PISA assesses the reading, mathematics and scientific literacy skills of 15-year-old students in 3-year overlapping cycles. A curious by-product of the release of comparative data from PISA is the almost palpable performance anxiety at the level of participating countries. Even more curious is the not-infrequent spectacle, at conferences and other national and international gatherings, of countries defining themselves in terms of their PISA results. This phenomenon is observed in low- as well as high-performing countries.

In a test that is conducted in 57 countries or economies (as PISA was in 2006), defines its construct in terms of real-life applicability, and emphasises a real-life context at the unit level, context is likely to be problematic. This paper traces consequences of the PISA construct, scientific literacy, especially how its emphasis on 'real-life' skills might contribute to various kinds of difficulty that students generally encounter in sitting for a test. The 107 items in the PISA science main study were analysed. Twenty-three of these items are in the public domain. Readers who are not familiar with the structure and content of PISA science units should study Appendix 1 before proceeding. The unit, 'Acid Rain', its items and other relevant information abut the items are used throughout this paper for illustrative purposes.

It is necessary to state at the outset that I consider the term 'real life' to be overemphasised in PISA testing even though officials and participating countries accept its centrality in the test development process. Furthermore, I believe there is a question about what has been sacrificed in the pursuit of 'real-life' contexts (the issue being with the meaning of 'real' as opposed to 'unreal'). I acknowledge that this personal view (perhaps to some a biased view) is a limitation of the paper.

CONSTRUCT, CONTENT, CONTEXT

PISA does not test science as such but scientific literacy. According to Thomson and Bortoli (2007), the definition of scientific literacy 'distinguishes knowledge *about* science from knowledge *of* science. Knowledge of science refers to *knowledge of the natural world* across the major fields of physics, chemistry, biological sciences, Earth and space science, and science-based technology. Knowledge about science refers to *knowledge of the means* (scientific enquiry) *and the goals* (scientific explanations) of science. Elements that 'underscore students' knowledge about the characteristic features of science' are added to emphasise knowledge about science as an aspect of science performance.

Thus PISA does not purport to test the common curriculum in science across nations. It is the notion of real-life skills, rather than documented curriculum, that drives the construct as a whole. The PISA notion of the real-life contexts in which it is essential for students to apply scientific knowledge and skills is presented in Figure1 (OECD, 2006). The two elements for classifying item context are *area of application* and *setting*. The categories for area of application are listed in the left-hand column of Figure 1: health, natural resources, environment, hazard, and frontiers of science and technology. The categories for setting (or situation) make up the top row in Figure 1: personal (self, family and peer group), social (the community), and global (life across the world).

For example, the area of application for all three items in 'Acid Rain' (refer Appendix 1) is deemed to be *hazard*, the setting for the first item is *social*, and for the other two, *personal*. Presumably the 'hazard' associated with acid rain relates to inanimate objects not people, the 'social' setting relates to the voice of an observer in 'Acid Rain' Item 1, and the 'personal' relates to the voice of a student empathising with the student doing the experiment described in 'Acid Rain' Items 2 and 3. To me the link between item and categorisation as documented by the test developers does not declare itself.

Figure 1.2 ■ **Contexts for the PISA 2006 science assessment**

	Personal (Self, family and peer groups)	**Social** (The community)	**Global** (Life across the world)
Health	Maintenance of health, accidents, nutrition	Control of disease, social transmission, food choices, community health	Epidemics, spread of infectious diseases
Natural resources	Personal consumption of materials and energy	Maintenance of human populations, quality of life, security, production and distribution of food, energy supply	Renewable and non-renewable, natural systems, population growth, sustainable use of species
Environment	Environmentally friendly behaviour, use and disposal of materials	Population distribution, disposal of waste, environmental impact, local weather	Biodiversity, ecological sustainability, control of pollution, production and loss of soil
Hazard	Natural and human-induced, decisions about housing	Rapid changes (earthquakes, severe weather), slow and progressive changes (coastal erosion, sedimentation), risk assessment	Climate change, impact of modern warfare
Frontiers of science and technology	Interest in science's explanations of natural phenomena, science-based hobbies, sport and leisure, music and personal technology	New materials, devices and processes, genetic modification, weapons technology, transport	Extinction of species, exploration of space, origin and structure of the universe

Figure 1: Contexts for the PISA 2006 science assessment

The proportion of items in each of the available contexts over all of the units on PISA science 2006 is shown in Table 1.

Table 1: Frequency of contexts in the 2006 PISA science assessment (main study)

Context	Environment	Hazards	Frontiers	Health	Natural resources	Other	Total
Frequency (~ %)	19	14	24	24	16	3	100

The PISA approach to context for testing the knowledge and skills required for adult life or real life appears to envisage a future largely in terms of stewardship of the environment and natural resources, at the personal, social and global levels. Notwithstanding the plausibility of this position for many, it is fair to say that today's notions of the future always seem more plausible than yesterday's, and to acknowledge that not all scientists would locate their work in such an immediate and direct relation to environmental applications.

The PISA approach to content can be seen from one point of view as overriding differences in school curricula: the domains are covered 'not so much in terms of mastery of the school curriculum, but in terms of important knowledge and skills needed in adult life' (OECD, 2006: 8). Stated in this way, the approach suggests a construct that rises above disparities between local curricula in a potentially equitable way. When considering how this approach might have an impact on the differential difficulty of items for various cohorts, however, it is necessary to recall that some curricula already embody PISA-like 'important knowledge and skills needed in adult life' to a greater extent than others. The apparently equitable supra-curriculum will resemble countries' actual curriculum to varying degrees. Where there is a lack of alignment between the PISA approach and the national approach there are obvious ramifications for test preparation and test-wiseness.

DIFFICULTY – EMPIRICAL, PERCEIVED AND IMPOSED

PISA results are reported as mean scores that indicate average performance and various statistics that reflect the distribution of performance. School and student variables are also provided. PISA attaches meaning to the performance scale by providing a profile of what students have achieved in terms of skills and knowledge. The performance scale is divided into levels of difficulty referred to as 'described proficiency levels'.

Students at a particular level not only typically demonstrate the knowledge and skills associated with that level but also the competencies required at lower levels. Proficiency levels are derived from the test data and defined in the various PISA reports (for example, Thomson & De Bortoli, 2007). There are six proficiency levels, from 1 (lowest) to 6 (highest). For example, the three released items in 'Acid Rain' come in at Levels 4, 2, and 6, respectively.

It can be deduced from Table 2 that the composition of PISA science 2006 is less than one-third open-ended items.

Table 2: Distribution of item types, by format, PISA science, 2006

Item type	Multiple choice	Complex multiple choice	Open response	Closed constructed response	Total
No. of items	38	30	37	5	110

In a test such as PISA science, which is time-limited, pen-and-paper, and of composition less than one-third open-ended items, it is possible to test the competencies *Explaining*, *Using* and *Identifying* at second-order level only, except where students are given the opportunity to provide written responses demonstrating skills such as explaining to others. PISA documentation lists the

competencies for the three items in 'Acid Rain' Items 1–3 as *Explaining*, *Using* and *Identifying*, respectively. Presumably being competent in the skill of identifying scientific issues required by 'Acid Rain' Item 1 is evidenced by the ability to recall information about the oxidation of sulfur and nitrogen and/or the ability to go beyond the data presented; being competent in the skill of explaining phenomena scientifically as required by 'Acid Rain' Item 2 is evidenced by the ability to interpret the meaning of words especially in the genre of scientific modelling; and, being competent in the skill of using scientific evidence as required by 'Acid Rain' Item 3 is evidenced by the ability to justify the steps in the design of an experiment. Again the link between item and categorisation as documented by the test developers does not declare itself to me.

Also, analyses that include competency as a variable ignore valuable information about the higher-order cognitive skills that are brought into play (e.g. the skill of hypothesising wrapped up with both *Explaining* and *Identifying* as noted in Appendix 1). In terms of a unit of analysis, the grain of competency, with only three categories, is too coarse to be the basis for reporting sub-group differences.

Sources of difficulty

Apart from the intrinsic difficulty of an item, student perceptions and design impositions have a role in explaining the differential difficulty by individual student or country on context-bound items.

Self-imposed difficulty, a function of a particular student's mindset on viewing the test item, might be influenced by features such as content and context. The notion of self-efficacy (Bandura, 1977) is important here. Matters & Burnett's (2003: 241–242) investigation of self-efficacy in relation to the propensity to omit items in tests has relevance to wider considerations of difficulty in tests:

> [Self-efficacy] is a cognitive mechanism consisting of beliefs concerning one's capacity to perform tasks successfully. These expectations are hypothesised to affect the initiation of coping behaviour, the expenditure of effort, performance accomplishment, and persistence in overcoming obstacles. Because self-efficacy relates to mastery and persistence, which, in turn, are precursors to academic success, self-efficacy could provide a possible answer to the question of who omits test items.

For PISA, with its varied clientele and dedication to real-life contexts, this kind of difficulty might be very significant.

The scoring rule that applies to multiple-choice items on PISA implies that under no circumstances is it better for test takers to omit an item rather than guess the correct response. According to Wood (1991: 23): 'Failing to answer [MC] questions can be due to believing that guessing is frowned upon (often quite mistakenly) or to running out of time'. PISA candidates are aware of the time limit and the scoring rule. If it can be assumed that all PISA test-takers are 'rational' (Budescu & Bar-Hillel, 1993) and strive to maximise their test scores, then the incidence of an omitted response should be zero. But it is not.

As is the case for MC items, failing to respond to an open-ended item can be due to running out of time or maybe to the belief systems that prevent students from attempting open-response items especially when they are not located at the end of the test. The partial credit scoring rule implies that under no circumstances is it better for test takers to omit an item rather than to 'have a go' and write something. But they do not.

Something else beyond scoring rules and rational behaviour might be pertinent to analysing omissions – students' initial perceptions of a test item affecting their inclination to respond. A student's perception of success is influenced by surface features of the stimulus material, which, in the case of PISA units sets the real-life context. Taken to extremes, this self-imposed difficulty might prevent the student from even attempting a response to an item. A high omit rate on an item could signal cultural bias. In trawling through Adams and Wu (2002), I came upon the

characteristics of a multiple-choice item (not in science and not located at the end of the test) with a facility of 79.35%, and an omit rate of 4.30%. This omit rate is not inconsiderable; its magnitude is more like that reported elsewhere for items where students have to generate a response rather than simply select the best response from (typically) four options. What made that item difficult? I cannot answer this question because the individual items were not available. Is it possible that the particular context of the item masked its intrinsic difficulty (easy) and created an impression of difficulty in students' minds so they omitted it? We can only speculate.

Problems with real-life applicability

Skills that can be identified as applicable to real-life are deemed to be worth assessing on PISA. Moreover, the test items themselves manifest contexts that are intended to be real-life, but which do not carry with them such a load of local content or fad/fashion that students in different locations would be disadvantaged by them. The aim is for 'contexts that are as realistic as possible and reflect the complexity of real situations', while 'bias due to the choice of contexts is minimised' (OECD, 2006: 37). The incorporation of 'real life' in PISA is both an assessment strategy (to do with how things get assessed) and inherent in the prior conception of what it is that gets assessed.

But PISA stands on problematic ground that lies between two positions: First, what is being tested derives its educational value from its embeddedness in daily (inevitably, to some extent, local) realities; and, second, performance on a test item is not to be skewed by local contexts. The one position implies experiential multiplicity; the other, an abstracted (underlying or overarching) commonality. The two positions are most reconcilable when seen within a 'we are-all-one-world' perspective. And the world of PISA is indeed more uniform than the actual globe. The 57 countries (and non-whole-country-economies) serviced by PISA in 2006 represented 90 per cent of the world's economy (OECD, 2006: 9), but they did not represent 90 per cent of the world's diversity. Society for PISA is explicitly the 'knowledge society': PISA 'aims to measure how far students approaching the end of compulsory education have acquired some of the knowledge and skills essential for full participation in the knowledge society' (OECD, 2007). The daily realities that are intended to validate the PISA construct are the relatively uniform, *sans-frontières* realities of that knowledge society. (Even within a single country, the concept 'knowledge society' filters out some quotidian reality.)

This arrangement could be seen as strength or a weakness. Some such notion of commonality must underlie the very process of comparative international reporting. However, Goody (cited in Rochex, 2006: 177) questioned (and found wanting) the 'notion that there should be any general competency for living in one country let alone across nations'.

Approaches to contextualisation

In responding to the challenge of a single test of life-related skills in a diverse range of life situations, PISA seems to adopt different approaches, *reduction* and *exclusion*, at different stages of the testing cycle. At the stage of formulating the construct to be assessed, the variety is reduced to an underlying commonality: All countries have different situations, but beneath those differences these are the skills that everyone needs in their own situations. This approach does not include local particularities in the construct, but allows for the underlying, common skills to be manifested in local ways. Rochex criticises the conception of skills as context-independent. He undertook a secondary analysis of the PISA 2000 Literacy test. The French students' results showed that, for a great number of them, the assumption that skills are steady and well-grounded and would turn up whatever the items are, is far from accurate.

> Many of the PISA literacy tests required students to mobilise various fields of reference and various registers of resources and to combine and organise the elements that they could draw from these fields and registers into a hierarchy. The issue of hierarchy was all the more the case given that the goal of the PISA designers was to assess 'the skills to carry out tasks that belong to real-life situations', rather than specific knowledge, and that their themes were often close to the social and cultural references and experiences of the young people taking the test. (Rochex, 2005: 185).

One of the conclusions of the study of students' methods (part of the larger study) was that, 'for a great number [of students], these methods varied more in relation to the texts and contexts, topics, and type of tasks or question formats than to their sole text treatment and reading and writing competencies – what was supposedly being assessed' (Rochex, 2005: 204). Rochex's finding has implications for the preparation of students for international surveys and also for national and state tests of generic or cross-curriculum skills, where skills that have been developed through the experienced curriculum (the study of several academic subjects) are then tested in unfamiliar contexts.

The other approach to contextualisation, exclusion, seems to occur at the level of item development. Here, local particularities that might jeopardise the general accessibility of the test across a range of countries seem to be removed. Another possible approach, appropriation – in which diversity is exploited – is not employed.

The test developers' intention to minimise bias is apparent. However, if a student's psychological reactions to an unfamiliar context can constitute a form of self-imposed difficulty, then the function of the test – to allow inferences to be made about the student's ability – may be compromised. Two issues arise here. The first one is a validity issue: Is there a disjuncture between the construct being assessed, which is grounded in (presumably manifold) reality, and the context of the test items, which seems to aspire towards unreal neutrality? The second issue relates more directly to self-imposed difficulty: Does a neutral context have the same effect on the mindset of all students in the range? Later in this paper is a discussion of the issues of authenticity, contrivance and novelty as ways of approaching context. But first, the point about so-called neutral context is made by referring to a 'neutral' country called Zedland which appears in the PISA mathematics framework (OECD, 2006: 9, 94). To be fair this is not PISA science but the philosophy is a shared one. Either way, the example well illustrates the conundrum of neutral contexts generally.

Zedland's currency is the zed. Calculations based on zeds are apparently taken to be more equitable – less subject to localised advantage and disadvantage – than those based on dollars, pesos, dinars or euros would be. Does this invented neutrality really serve a purpose? Will any students successfully perform a calculation in zeds that they would be unable or unwilling to perform in dollars or euros? A more important question might be: Does shared unreality really ensure equal access? Has a stringent effort to remove the distraction of local reality introduced a greater distraction, that of unreal neutrality? Is the unreality of the zed more of a barrier for some students than a possibly unfamiliar reality (say, a peso in New Zealand) would have been? How authentic is the task?

It would appear that students' perceptions of difficulty might contribute to empirical difficulty.

Design-imposed difficulty

Students experience a test not just at the item level but at the unit (or batched-item) level. Design-imposed difficulty originates in features of the item that occur as a result of the test developers' fulfilment of one or more of the design criteria. PISA documentation includes a classification scheme for application to test items. There are five criteria as exemplified in Table 3 (OECD, 2006: 120).

Table 3: Illustration of the application of the PISA item classification scheme

Criterion	Class
Item type	Open-constructed response
Competency	Explaining phenomena scientifically
Knowledge category	Earth and space systems (Knowledge of science)
Application area	Natural resources
Setting	Global

There are, therefore, possible sources of difficulty that derive from the design of the test; for example, mode of response (item type) and knowledge category (domain). And then there is the turbulent experience for students as they move from one unit to another or from one context to another, 'changing intellectual gears' as they go.

To the list containing five criteria for classifying PISA items (Table 3), other criteria could be added; for example, 'approach to contextualisation' (three categories, reduction, exclusion and appropriation as discussed above) and 'epistemic content' (five categories as discussed below).

Epistemic content

Phenix (1964) elaborates a system of epistemic areas, which are of relevance in the design of tests of skills that are not subject-specific in the same way that PISA tests science knowledge and skills (e.g. the scientific method) that do not belong exclusively to chemistry or physics or biology or Earth and space. Part of the context in a test of generic skills is provided by the epistemic area from which the stimulus material is extracted. For a test that assesses performance in a particular field of knowledge as does PISA science, a cursory glance at a summary of Phenix's (1964) 'realms of meaning' (Table 4) might effectively rule out many of these epistemic areas for use in PISA science stimulus material. However, disciplinary skills can be assessed via epistemic areas that do not at first glance correspond to the discipline in question.

The entries in italics in Table 4 refer to epistemic content already used in PISA science for item context thus revealing yet another criterion that could be used to classify PISA items.

Table 4: Summary of epistemic content (after Phenix, 1964)

Epistemic content	Examples of subject for context
Symbolics	*ordinary language* *mathematics* non-discursive symbolic forms
Aesthetics	music visual arts *arts of movement* literature
Synnoetics	*personal knowledge* *ethics*
Empirics	*physical science* *life sciences* psychology social science
Synoptics	*history* religion politics philosophy

The importance of classification systems cannot be overstated at the test assembly stage. Content validity requires a balance in the proportion of various dimensions and a suitable range of item characteristics. The classification of items on multiple criteria allows test developers to construct a matrix so that they can reflect on the composition of the test both as a valid instrument and as a set of experiences for test-takers. These two perspectives on test composition ensure that decisions based on the reliability imperative (such as this item 'worked' at trial) do not result in a test that students experience as a set of items with a certain sameness about them.

Another phenomenon that is a function of the test design and observed in PISA science units is now discussed. There appear to be items on PISA that could be answered by savvy 15-year-olds including typical Australian 15-year-olds who see, hear or subliminally absorb the news on commercial television. These young people are attuned to popular analyses of topics for mass consumption; and there is a preponderance of such topics on the PISA science test. Could this phenomenon possibly account for differences between countries? If so, are there implications here for bias or is there a discussion to be had about test-wiseness?

Many items can be answered without an understanding of atoms and molecules, cells and organs, forces and waves and so on. Therefore, they are not really testing the application of science to the so-called real world. Although there is no law that says that they should be so, the fact that they are not would be a surprise to many policy makers and journalists. It is almost as if students are being rewarded for knowing (read 'being aware of') the application without studying the underpinning theory.

Some fragments in the definition of scientific literacy do not appear to figure at all. For example, 'scientific knowledge to understand the natural world and participate in decisions that affect it' surely implies that one cannot just think but, instead, must think about something – which brings us back to the beginning of this chapter: knowledge of the natural world across the major fields of physics, chemistry, biological sciences, Earth and space science, and science-based technology versus knowledge of the means (scientific enquiry) and the goals (scientific explanations) of science. The point here is not to privilege one or other (knowledge *of* and knowledge *about*) but to attempt provide a partial answer to the question of what has been sacrificed in the real-life approach.

It would be an interesting experiment to compare the performance of Australian students on PISA and a test of scientific literacy where the context was not so heavily dependent on topics discussed in the mass media. Or to compare the performance of Australian students and students from other countries (say the former Soviet states) on items related to the sorts of topics mentioned above.

After all, PISA literacy scores improved in Poland after an intervention based on the research finding that the 'literate environment' was one of the factors contributing to differential performance (see Bialecki, 2008: 91).

It would appear that aspects of test design might contribute to item difficulty.

Authenticity, contrivance and novelty – some approaches to context

Sometimes plausibility of context is taken to be sufficient for authenticity of assessment. It can certainly be a part of it, as when the assessment task is 'real' in the sense that students experience it as it could be carried out in a non-school environment (assuming that such an environment is part of real life). Other elements also contribute: the range of response modes and the presence of skills developed in other subject areas. Authentic assessment involves students in the use of relevant and useful knowledge, thinking and practical skills. Real-life and authentic can, however, be differentiated.

The context of the *non*-PISA items in Appendix 2, for example, cannot claim surface plausibility because it is hard to believe in trained ants in another galaxy. The context, however, does allow the assessment of mathematical skills (using Pythagoras, applying a progression of steps, performing calculations, analysing) in a way that might have real-world applicability. These items adopt the approach of *declared contrivance* rather than the pretence of authenticity. Just as declared contrivance is not incompatible with authenticity, nor is *novelty* of context. A context may be novel to the candidates – not only unfamiliar in real life but unrehearsed in an assessment situation – but still be real and still embody real-life skills. For example, a unit set in the context of preparing for a night at the opera (a real-life situation that would be novel to most candidates) could be the context for assessing deductive skills that would be required in, among other real-life situations, scientific investigation (see Appendix 3). I am not of course suggesting that this stimulus material would find a place in a test of scientific literacy but I am presenting it as an extreme example of *unrehearsed context* for testing a particular skill or set of skills. Such items are not to be found in PISA where the choice of context seems largely circumscribed by the range of contexts that would *readily be agreed to be* of the real life.

Most working definitions of 'real-life importance' [in this context] would leave ample room for knowledge and skills that are worth learning but lack real-life importance. It would presumably be possible to identify skills with real-life importance (however defined) and then assess them without using real-life contexts. Equally, it would be possible to identify skills that do not have a real-life importance but assess them within real-life contexts. A further refinement exists – assessing real-life skills in a real-life context that is not the one in which those skills would most readily be expected (as in the opera items above). To illustrate these variations Kelly (2008) devised a grid of seven possibilities (Figure 2).

<table>
<tr><td></td><td></td><td colspan="2">Importance of knowledge & skills</td></tr>
<tr><td></td><td></td><td colspan="2">Real-life</td><td>Not real-life</td></tr>
<tr><td rowspan="4">Item context</td><td rowspan="2">Real-life</td><td>Same</td><td>1</td><td rowspan="2">5</td></tr>
<tr><td>Different</td><td>2</td></tr>
<tr><td>Contrived</td><td colspan="2">3</td><td>6</td></tr>
<tr><td>None</td><td colspan="2">4</td><td>7</td></tr>
</table>

Figure 2: Contextual basis of knowledge and skills, test items

An analysis of the PISA documentation would suggest that most PISA items are located in Cell 1 or Cell 2. In fact, the context is often so neutralised as to make Cell 4 a more accurate description of PISA items. Cell 1 belongs to the item type in which real-life skills are assessed in a real-life context that is the same as the one in which those skills would most readily be expected. Cell 2 belongs to the item type in which real-life skills are assessed in a real-life context that is not the one in which those skills would most readily be expected. Cell 4 belongs to the item type in which real-life skills are assessed without using real-life contexts or even contrived contexts. Context might be non-existent.

The downside of presenting Figure 1.2 of OECD (2006) (refer to Figure 1 earlier in this paper) is that it would be tempting for test setters to stay with the examples given and overlook other

exciting possibilities. Nuclear weapons and interrogation methods (the physics and physiology of) come to mind but using these as a context would involve some element of moralising, just as there has always been in testing – another limitation of the pursuit of real-life contexts.

GENDER DIFFERENCES

If it is the case as (Hawkins, 1983) that masculine thinkers (often boys but including girls) outperform feminine thinkers when confronted with the esoteric, the abstract, and the visuo-spatial, then why the remarkable gender neutrality overall in PISA science?

PISA is a reading-dependent type of testing. The PISA Reading results from the same students and the gender differences that emerge on the separate scientific competencies are consistent across the OECD PISA countries; namely, girls outperforming boys in *Identifying*; boys outperforming girls in *Explaining*; and no gender differences in *Using*.

Aligned with the reading-dependent nature of PISA science testing is the breakdown of PISA units by epistemic content of stimulus material (see Table 4). For half of the units in the main study the stimulus material is in the form of 'ordinary language' (Symbolics); for the other half it is mainly 'life sciences' and 'physical sciences' (Empirics). These two opposing influences on real and perceived difficulty could be admissible as explaining the gender neutrality in PISA science results.

Table 4: Epistemic content of PISA science units, main study 2006

Epistemic content	**Units in main study**	
	No.	**Specific area**
Symbolics	18	Ordinary language
Empirics	14	Life sciences Physical sciences
Aesthetics	1	Visual arts
Synoptics	4	History
Synnoetics	0	–

Some history of gender issues in education is necessary. The following discussion, which provides some background to *changing differences* in gender differences over the past 30 years or so draws heavily on Matters, Allen, Gray and Pitman (1999), who argue that the feminisation of education is, to a large degree, for the underachievement of boys relative to girls (reported in Australia in the 1990s), which in turn is related to the issue of contextualisation of test items.

Questions about the underachievement of boys have replaced questions from the 1970s about whether tests were biased in favour of boys (e.g. Adams, 1987). The post-1960s movement towards gender equality (read improvements in female participation and achievement) occurred on several fronts (at least in Australia): the status of women in society generally was raised; the proportion of females holding executive positions in curriculum/assessment agencies increased rapidly; female teachers gradually became the dominant force in the secondary school; and there was deliberate intervention to exploit the learning styles of girls within content relevant to their predilections.

These interventions had a significant impact on curriculum content and assessment practices. Examples of the subtle changes included a decreased emphasis on technical correctness in English, concentration on the local rather than the global (as in geography), redefinition of mathematics (as it loses its general abstract power so does it alienate), cutting out topics in which boys traditionally excel (e.g. solid geometry, astronomy), and test items especially in mathematics and science stemming from the need to contextualise.

Arguably, at least two of these observations are pertinent to the analysis of PISA science items given the serious issue of *item context*), which is a focus of this paper. Particular aspects of context that are illuminated here include the following polarities: the everyday versus the esoteric; abstract versus concrete; and verbal stimulus material versus visuo-spatial stimulus material. In PISA, esoterica is virtually ruled out by the real-life emphasis in the definition of scientific literacy. In much the same way, abstraction is ruled out in favour of what is 'real'. As well, the proportion of science items in the PISA main study that required spatial reasoning skills was minuscule. Given these features why would gender neutrality be expected? Given that there *is* gender neutrality, does this raise questions about the construct? Is it the case that the three competencies are the most appropriate 'baskets' of PISA items (30% *Using*, 22% *Identifying*, and 48% *Explaining*) for analysing gender differences? Otherwise one might be led to conclude that eminent researchers in the area of gender-related abilities (e.g. Stage, 1994; Willingham & Cole, 1997) are wrong. Or it might be that the remarkable gender neutrality overall in PISA is symptomatic of a neutered context … or a safe haven for all.

CONCLUSIONS

The problematic nature of real life as a context

A study of the PISA 2006 science items in the light of the related issues of context and difficulty points to an inherently complicated situation. On the one hand, the items exemplify an attempt at minimising bias caused by contexts that may be more familiar to some students than to others (inevitably an issue in a test that services such a range of locations). On the other hand, the neutralising of contexts raises its own issues: (i) the possibility that students will respond differentially to neutral contexts; and (ii) the potential weakening of the link between the test and the construct (the construct being based on real-life applicability). Furthermore, it would seem that the range of available options for contexts is not being fully utilised at the unit level and that some validity issues need to be confronted at the test level.

The explanatory power of design-imposed and self-imposed difficulties

Aspects of test design and students' perceptions of item difficulty help explain the differential difficulty of PISA science items. Contributors to self-imposed difficulty and design-imposed difficulty include epistemic content of stimulus material, items requiring the generation of a response, confusion in the un-cued movement between formats and contexts and from 'knowledge of' items to 'knowledge about' items, and in ambiguous wording of test items. Context *per se* may not be the outstanding source of difficulty after all.

Of use to test developers and test users would be the addition of an extra layer of review once the test has been assembled. I envisage the direction of program funds to a process for detecting semantic anomalies in the test items.

The issue of validity − the functional and explanatory perspectives

This paper has raised some questions about validity. It would seem that these issues need to be confronted at test *and* item level to ensure that the integrity of the data upon which many generalisations about countries and systems are based, is not compromised. Misconceptions about what students are supposed to be able to do and how students are asked to demonstrate these skills in the test would be circumvented by precise language for explaining the construct of scientific literacy. Not included here but work in progress is a short critique of language usage in the PISA explanation of scientific literacy. It includes a translation from PISA documentation to operational definitions.

Students are most likely to do well on tests of things they have been taught. International comparisons include countries whose curricula emphasise the sort of thinking rewarded by PISA

to greater and lesser degrees. The issue is not whether they should or should not be emphasised, should or should not be rewarded, but whether the manifold technical reports are read against this reality.

According to Cronbach's (1988), *functional* perspective on validity, we are required to look at the test through a lens that focuses on the test's antecedents and consequences and, in drawing conclusions about the worth of the test, ask questions about the consequences of the test on various players and the sorts of behaviours that might occur before the test. For PISA one possible answer is: As a consequence of this test, certain countries' education systems are vilified in the media; and an antecedent to the test is the pressure on students from some countries who have not experienced tests other than content-specific, multiple-choice tests containing items that are not contextualised much less located in an unfamiliar context.

According to Cronbach's (1988) *explanatory* perspective on validity, we are required to look at the test through a lens that focuses on interpretations of test data and, in drawing conclusions about the adequacy and appropriateness of the conduct, analysis and interpretation of PISA results.

The other three perspectives, *political*, *operationist* and *economic*, although capable of generating powerful questions about the test's validity, are not elaborated on here. Suffice it to say they have all been alluded to: absence of bias and fairness across countries/economies (the political perspective); test design and item properties (the operationist perspective); and the relevance and utility of test statistics (the economic perspective).

AND, FINALLY …

This paper has raised some methodological and ethical concerns about PISA. However, comments with an evaluative tone are not directed at PISA alone but refer to aspects of other standardised tests in various parts of the world, and aim to enhance standardised testing per se rather than attribute any negativity to PISA itself.

In the current environment of increased use of student performance data for international comparisons and national accountability purposes, it is timely to be concerned about the design of tests, the interaction of students with individual items and the analysis of test data. It is also timely to encourage teachers and users of test results to gain a better understanding of how a collection of items becomes a test. The stakes are higher than ever, and the requisite demands on reliability and validity, which are at the core of the methodological perspective, are extremely high. Equity and quality are inextricably linked. Therefore, Cronbach's (1988) 'disputatious community' is to be encouraged to partake in the pursuit of fairness and excellence in external standardised tests.

REFERENCES

Adams, R., & Wu, M. (Eds.). (2002). *PISA 2000 Technical Report*. Paris: OECD.

Adams, R. (1987). *Sex bias in ASAT?* Melbourne: ACER.

Bandura, A. (1977). *A social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Bialecki, I. (2008). 'Assessment measurement and evaluation of literacy levels and of basic competencies in Poland'. Paper delivered at UNESCO Regional Conference (Europe) in Support of Global Literacy, Baku, Azerbaijan.

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement*, *30*(4), 277−291.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity*. Hillsdale, N.J.: Lawrence Erlbaum.

Goody, J. (2001). Competencies and education: Contextual diversity. In D. S. Rychen & L. S. Salaganik Eds.). *Defining and selecting key competencies*, 175−190. Bern: Hogrefe & Huber.

Hawkins, B. L. (1983). 'Agency and Communion: An alternative to masculinity and femininity'. Paper presented at the annual convention of the American Personnel and Guidance Association, Washington, DC.

Kelly, D. J. (2008). 'Authenticity, contrivance and novelty: Some approaches to context'. Unpublished manuscript. Brisbane: ACER.

Matters, G. N., Allen, J. R., Gray, K. R., & Pitman, J. A. (1999). Can we tell the difference and does it matter? Differences in achievement between girls and boys in Australian senior secondary education. *The Curriculum Journal*, *10*(2), 283–302.

Matters, G. N., & Burnett, P. C. (2003). Psychological predictors of the propensity to omit short-response items on a high-stakes achievement test. *Educational and Psychological Measurement*, *63*(2), 239−256.

OECD. (2007). *Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006*. http://www.pisa.oecd.org/dataoecd/63/35/37464175.pdf

Phenix, P. H. (1964). *Realms of Meaning*. New York: McGraw-Hill.

Rochex, J.-Y. (2005). Social, methodological, and theoretical issues regarding assessment: lessons from a secondary analysis of PISA 2000 Literacy Tests. *Review of Research in Education*, 30, 163–212.

Stage, C. (1994). 'Gender differences on the SweSAT: A review of studies since 1975'. Department of Educational Measurement, Umeå University, EM No. 7.

Thomson, S., & De Bortoli, L. (2007). *Exploring Scientific Literacy: How Australia measures up. The PISA 2006 survey of students' scientific, reading and mathematical literacy skills*. Camberwell: ACER Press.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Princeton, NJ: Lawrence Erlbaum.

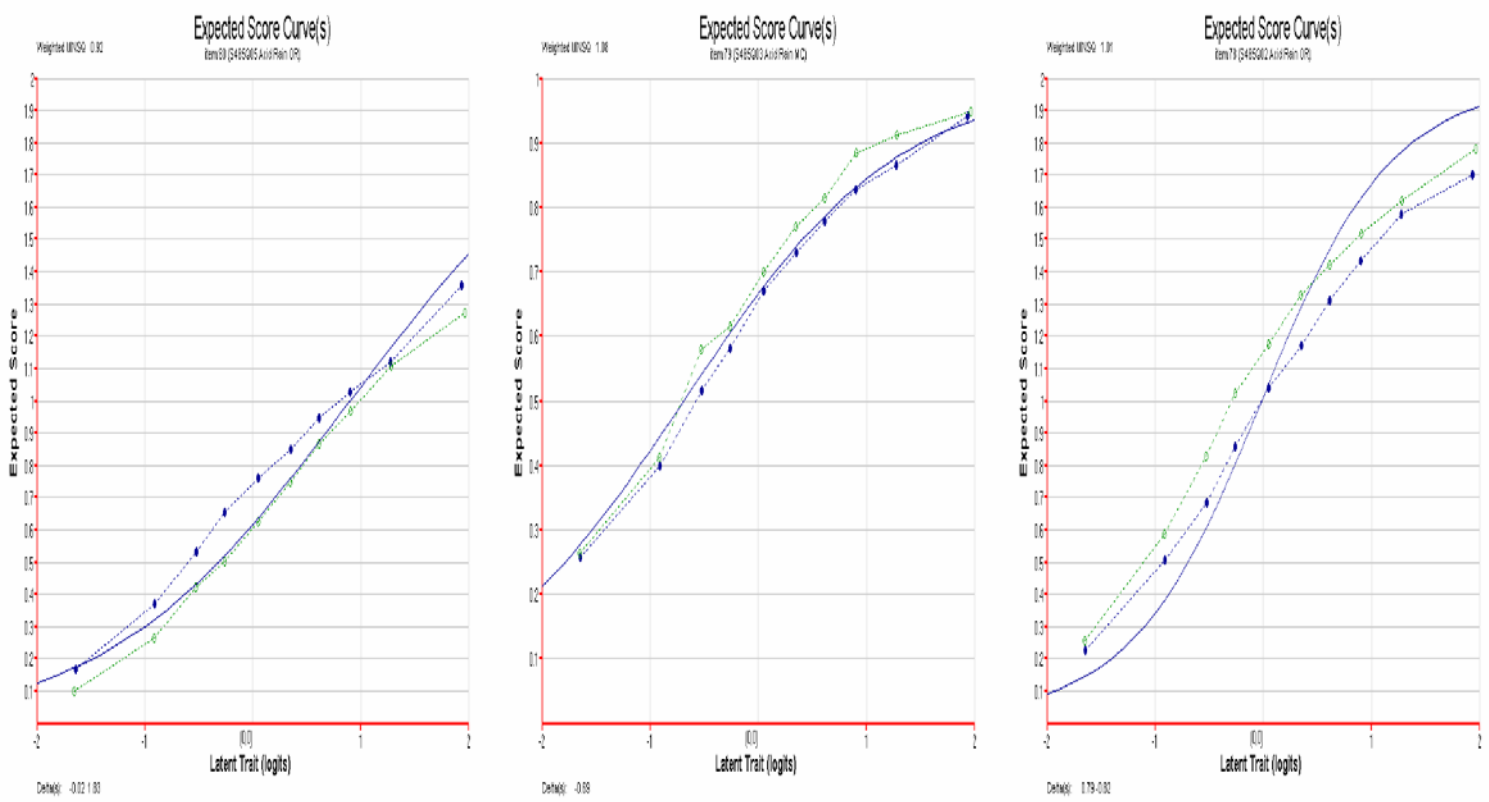
Wood, R. (1991). *Assessment and testing: A survey of the research*. Cambridge, England: Cambridge University Press.

\

APPENDIX 1

ACID RAIN Items 1, 2 and 3 (diminished version of original presentation)

Annotations: **area of application**; (**finer grained skill**); **difficulty** (derived from proficiency level); key (*) (multiple choice); exemplar response from PISA report (open-ended and closed response), item characteristic curve by gender (blue = F; green = M) (Main Study Items, 2006, Science Expert Group Meeting, Lyon), cryptic comment on each



ACID RAIN

Below is a photo of statues called Caryatids that were built on the Acropolis in Athens more than 2500 years ago. The statues are made of a type of rock called marble. Marble is composed of calcium carbonate.

In 1980, the original statues were transferred inside the museum of the Acropolis and were replaced by replicas. The original statues were being eaten away by acid rain.

ACID RAIN 1 – **Explaining phenomena scientifically (Extrapolating) - Medium difficulty**

Normal rain is slightly acidic because it has absorbed some carbon dioxide from the air. Acid rain is more acidic than normal rain because it has absorbed gases like sulfur oxides and nitrogen oxides as well.

Where do these sulfur oxides and nitrogen oxides in the air come from?

.....

.....

Any one of car exhausts, factory emissions, burning fossil fuels such as oil and coal, gases from volcanoes or other similar things.

Burning coal and gas

Oxides in the air come from pollution from factories and industries.

Volcanoes

Fumes from power plants [taken to include plants that burn fossil fuels]

They come from the burning of materials that contain sulfur and nitrogen.

The effect of acid on marble can be modelled by placing chips of marble in vinegar overnight. Vinegar and acid rain have about the same acidity level. When a marble chip is placed in vinegar, bubbles of gas form. The mass of the dry marble chip can be found before and after the experiment.

ACID RAIN 2 – **Using scientific evidence (Interpreting the meaning of words and other symbols) – Easy**

A marble chip has a mass of 2.0 grams before being immersed in vinegar overnight. The chip is removed and dried the next day. What will the mass of the dried marble chip be?

A. Less than 2.0 grams*
B. Exactly 2.0 grams
C. Between 2.0 and 2.4 grams
D. More than 2.4 grams

ACID RAIN 3 – Identifying scientific issues (Explaining, Justifying) – Very difficult

Students who did this experiment also placed marble chips in pure (distilled) water overnight.

Explain why the students included this step in their experiment.

.....

.....

To show that the acid (vinegar) is necessary for the reaction

To make sure that rainwater must be acidic like acid rain to cause this reaction

To see whether there are other reasons for the holes in the marble chips

Because it shows that the marble chips don't just react with any fluid since water is neutral

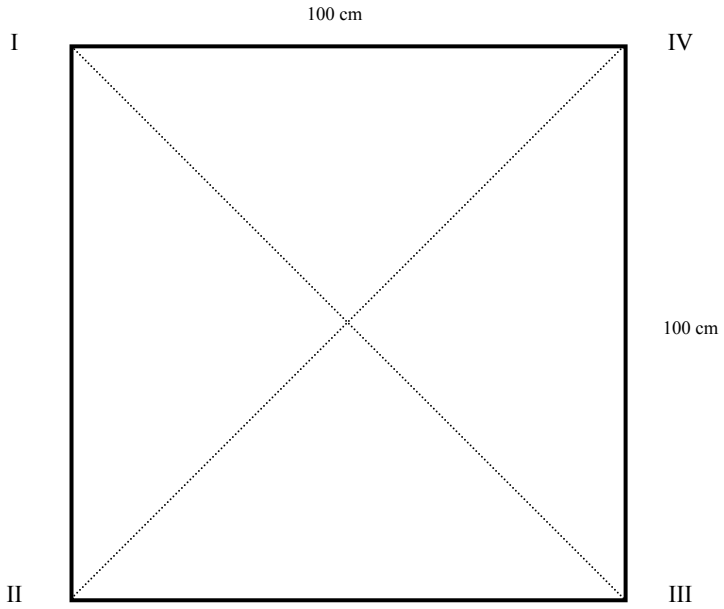
---

APPENDIX 2

On the planet Archid, in a galaxy far away, lives a species of intelligent ants that can be trained by Archidians to crawl in straight lines at constant speeds. The constant crawling speed of one ant may, however, differ from that of another.

An Archidian selects four trained ants, labelled I, II, III and IV, and places them on the comers of a square board of side exactly 100 centimetres.

Figure 1 shows the board, its diagonals, and the positions of the four ants.



ITEM 9 [***]

Ants I and III are released simultaneously from their initial positions (in Figure 1) and crawl towards each other along the diagonal. Ant I crawls twice as fast as Ant III.

What distance does Ant I crawl before the two ants meet?

Show your working. .....

.....

ITEM 10 [**]

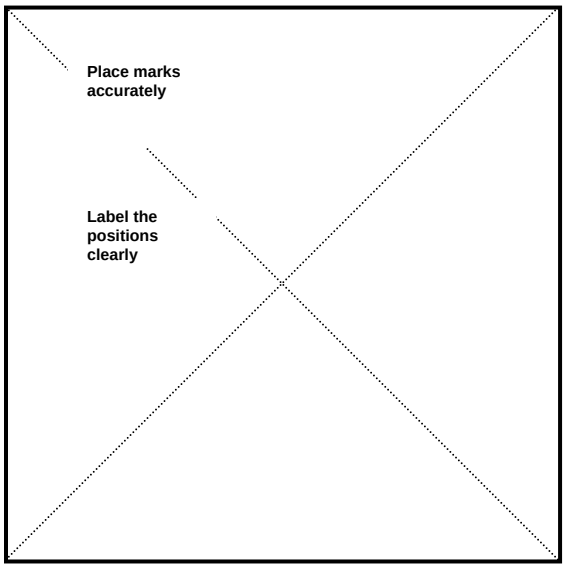
Point P, not shown on the diagram, is reached by Ant IV 20 seconds after its release.

How many times does Ant IV pass point P in the 350 seconds after its release?

Give a number only. ……

ITEM 11 [***]

The diagram below is drawn to scale. On this diagram, mark the positions of Ants II and IV 50 seconds after their release. Label these positions II and IV respectively.



Source: The 1993 Queensland Core Skills Test Paper 3

APPENDIX 3

See next page (third item in a set of three in a unit in short-response format.

Source: The 1994 Queensland Core Skills Test Paper 3

20

ITEM 16 [****]

The reader of a plot summary cannot expect to obtain all the information that the audience would acquire at an actual performance of a scene.

This opera concludes with Count di Luna ordering the execution of Manrico. As Manrico dies, Azucena[2] shouts to di Luna, "He was your brother! Mother, you are avenged!"

Note: [2]Azucena has by then been revealed to be the gypsy daughter mentioned in both guides.

Make a considered judgment as to which plot summary better enables the reader to anticipate the ending.

Tick one.

The plot summary in Guide $\mathcal{V}$ ❑

The plot summary in Guide $\mathfrak{M}$ ❑

Demonstrate that the weight of evidence favours this judgment.

Focus on significant differences in the plot summaries.

Fresh working copies of the plot summaries are provided on the next page.