The Irish Journal of Education, 2010 *xxxviii*, pp. 94-118.

PISA: ISSUES IN IMPLEMENTATION AND INTERPRETATION

Eemer Eivers
Educational Research Centre
St. Patrick's College, Dublin

Issues with the conceptualization, implementation and interpretation of the OECD Programme for International Student Assessment (PISA) are examined. The values that underpin the project are discussed. What PISA is intended to measure (preparation for life, key competencies, real-life challenges, curriculum independence) is contrasted with what it probably measures. Issues are identified relating to cultural fairness (quality and equivalence of translations, Anglophone origins, response styles, importance accorded to the test) and the representativeness of participating populations (samples, response rates, adjustment for nonresponse).

PISA is a project of the Organisation for Economic Co-operation and Development (OECD), set up to provide member states with 'international comparisons of the performance of education systems' in key subject areas (OECD, 2001, p. 27), more specifically, the reading, mathematical, and scientific literacy skills of 15-year-olds. First conducted in 2000, it runs in three-year cycles. It is one of a number of large international surveys, such as Trends in International Mathematics and Science Study (TIMSS) and PIRLS (Progress in Reading Literacy Study), both organized by the International Association for the Evaluation of Educational Achievement (IEA). PISA differs from its competitors in its frequency and cyclical nature, and in its focus on the 'knowledge and skills that are essential for full participation in society' (OECD, 2007, p. 16) rather than on the outcomes of a curriculum.

Perhaps due to the reasonable performance of Irish students, PISA has not come under the same media and academic scrutiny in Ireland as in some other countries. For example, the unexpectedly low ranks in PISA 2000 of Germany and Denmark resulted in major political, educational and academic responses, and in changes in their education systems (e.g., Dolin, 2007; Rubner, 2006). It has meant that an unusually large proportion of academic criticism of PISA has come from Germany (e.g., Hopmann, Brinek, & Retzl's 2007). This can be contrasted with the largely uncritical response to PISA in Ireland.

Sjøberg (2007) has described the data generated by PISA as a playground for psychometricians. In this paper, I examine how much we can rely on what their playground activities tell us about educational, as opposed to psychometric, issues. An unfortunate byproduct of the complexity of the statistical techniques used in PISA is that few feel qualified to debate what PISA does and what it means, a point noted by Rochex (2006). What debate there is tends to be technical relating to the statistical methodologies used – e.g., the use of a single parameter Rasch model to scale test data – aimed at statisticians, not educationalists. In this paper, an attempt is made to describe some of the broader issues relating to the implementation and interpretation of PISA, with a reasonably broad audience in mind. Five aspects of PISA are examined: the values underpinning PISA; what it is intended to measure; what it probably measures; whether it is a culturally fair assessment; and the representativeness of the participants.

THE VALUES UNDERPINNING PISA

PISA is one of the largest educational surveys in the world. Although initially envisaged as a means of supplying OECD countries with data on which to base policy, more non-OECD than OECD countries took part in PISA 2009. Its size, coupled with the prestige of the OECD name, has led to what Grek (2009) called 'a taken-for grantedness' about the education indicators it produces. However, it is worth remembering that the OECD is, as its name suggests, dedicated to economic growth, co-operation, and development. PISA reflects OECD aims, with an emphasis on economic priorities, and the drive to create efficient education systems, offering value for money, and producing quality outputs. As Bonnet (2002) noted, studies such as PISA are appealing to policy makers because of a belief that countries with effective education systems become successful economies. Bonnet's point is not about the strength (or weakness) of such a relationship, but that political interest in cross-national studies is largely derived from economic, not educational, interests.

An economic perspective is apparent in the selection of reading, mathematics, and science as the key skills or competencies for future life (the corollary of this selection being that subjects such as social science, foreign languages, art and music do not provide students with key life skills). Similarly, the desire to compare education systems and measure value for money or 'added value' can be traced to the economic priorities of the OECD. As is

explicitly stated in its aims and key features, PISA is not simply a student test; it is designed to bring about policy changes in education (OECD, 2009). Countries featuring lower down the PISA national 'league tables' find their education systems under pressure to change so that students perform better on PISA. Moving up the PISA league tables is seen as making a country more attractive for economic and human capital investment (Grek, 2009).

While the OECD has described PISA as a means of comparing the quality of education systems, Uljens (2007) suggested that PISA is intended to homogenize education systems and to create a mentality of competition between countries. He believes PISA has a 'hidden curriculum' that is redefining the goals of education and education systems. Policy makers are moving away from the traditional *Bildung* view (broadly, that learning leads to independence, self-awareness and maturity, and that these are worthy goals in themselves). Replacing this is a neo-liberal perspective, characterized by (i) *economization* (value measured mainly in economic terms), (ii) *privatization* (education as a private commodity; 'goods' rather than something for the public good), and (iii) *productivity* (stimulating economic growth as a core objective). In Uljens' view, PISA is causing countries to move towards a competitive approach to education, where more attention is paid to topping the 'league tables' than to what students actually learn in schools.

While Uljens' perspective on PISA is probably a minority one, contrasting his and any OECD description of PISA highlights how ideological perspectives can influence interpretation of information. Of course, PISA should not be criticized simply because it is based on an ideology; all research has some form of ideological basis. However, it is naïve to assume (as many seem to) that PISA is a politically neutral entity. All elements of the study – from deciding what should be assessed to deciding how to interpret and report data – are underpinned by an ideology. As Sjøberg (2007) noted 'PISA results and advice are often considered as objective and value-free scientific truths, while they are, in fact embedded in the overall political and economic aims and priorities of the OECD'(p. 203).

WHAT PISA IS INTENDED TO MEASURE

In this section, I examine four aspects of PISA that can be considered as key, and discuss how well these aspects are reflected in practice. The four aspects are: how well schools have prepared students for life; choice of key competencies; use of real-life challenges; and, curriculum independence.

Students' 'Preparation for Life'

The OECD selected 15 as an age that was close to the end of compulsory schooling in many countries. However, countries differ in typical school-leaving age (compulsory and otherwise). The typical Irish student continues in school long after his or her 15th birthday, and few Irish schools would claim that their student body was 'prepared for life' by age 15. In contrast, almost half of 15-year-olds in the OECD countries of Turkey and Mexico have already left school, meaning that those targeted by PISA are not representative of 15-year-olds as a whole in those countries. A suitable age at which to assess how Irish schools prepare students for life might be 17 years, while in Turkey, 14 might be a better target age. While 15 is the most reasonable compromise age, given the task definition, using an age-based sample to assess something which varies widely by age is in itself problematic. A more accurate description of PISA might be that it is a comparison of the skills of 15-year-olds still enrolled in schools in participating countries.

Key Competencies

As noted earlier, it could be argued that reading, science and mathematics were picked because of their perceived value as economic assets, not because they represent key competencies for participation in society. Kellaghan and Greaney (2001) believe that key skills in a global economy include higher-order thinking skills, the ability to learn quickly, to manage and process information, and problem-solving. They suggest that 'the achievements that are assessed in most national and international assessments would seem to fall far short of these, for the most part focusing on the core curriculum areas of reading, mathematics and science'. More generally, the assumption of a universal set of competencies is problematic *within* a country, and quite difficult to sustain across countries: 'the notion that there should be any general competency for living in one country let alone across nations seems open to serious questions … within countries there are clearly different expectations in terms of the lives that people are likely to lead and that are required to fill their social, cultural and economic needs' (Goody, 2001, p. 184). To take an extreme example, it is difficult to imagine that the child of a goat herder in Azerbaijan and the child of a Dublin dentist will require the same 'knowledge and skills that are essential for full participation in society', yet it is upon such a one-size-fits-all assumption that PISA is based.

'Real-life' Challenges

There are two main difficulties with this element of PISA. First, and most obviously, PISA does not measure real-life skills. It is a paper-and-pencil test, with no practical component. This criticism is not unique to PISA, but as one of the major 'selling points' of PISA is that it ostensibly measures real-life skills, it is an important point to make. Second, the real-life focus requires a certain style of task, and means that some elements of the three domains are more easily assessed than others. For example, over 28% of PISA 2006 science items could be categorized under biology in the Junior Certificate science syllabus, compared to fewer than 15% under chemistry (Eivers, Shiel, & Cunningham, 2008). The disparity in coverage may be because it is easier to write a meaningful context for a biology item than for a chemistry item[1]. Similarly, Bodin (2007) believes that some fundamental mathematical concepts are excluded from the PISA assessment of mathematical literacy because it is not possible to fit them into the real-life style of a PISA test unit.

Curriculum Independence

This key aspect is intended to distinguish PISA from the curriculum-linked TIMSS. In developing the PISA materials, a group of subject experts for each PISA domain develop a framework for the domain. The framework details the key elements of that domain, defines what students should know, and informs how they should be tested. Theoretically, it is not based on the curriculum of any one country. However, given complaints about the lopsided national compositions of the PISA expert groups and item writers (e.g., Bottani & Vrignaud, 2005; Murat & Rocher, 2004), it is reasonable to say that PISA reflects the curricula (and 'world view') of some countries better than others. Thus, it could be argued that while TIMSS is explicitly informed by the curricula of all participating countries, PISA is implicitly informed by the curricula of some participating countries.

Irish research has found that, as might be expected, students tend to do best on PISA items related to topics covered in the relevant Junior Certificate syllabus (Cosgrove, Shiel, Sofroniou, Zastrutski, & Shortt, 2005; Eivers, Shiel, & Cunningham, 2008; Shiel, Cosgrove, Sofroniou, & Kelly, 2001). In a similar

[1] Despite the heavy emphasis on biology, human reproduction did not feature as a topic in the PISA 2006 science assessment. Eivers, Shiel and Pybus (2008) speculated that this was not due to the difficulty in supplying a context, but because it would have been viewed as an unacceptable topic in some of the participating countries.

vein, students in countries where curricula are similar to the relevant PISA frameworks may have an advantage (on PISA) over students in countries with curricula dissimilar to PISA frameworks. This is why TIMSS explicitly deals with the link between national curricula and national performance. In contrast, PISA's emphasis on a future-oriented, skills-based assessment seeks to underplay the importance of curriculum in defining the domains, and largely ignores the influence of curriculum in interpreting the results. Thus, the influence of the curriculum experienced by students on PISA test performance is at best unexplored, and at worst, treated as a source of 'bias' in items, and something to be removed. Furthermore, the failure to collect information on students' classroom experience means that a potential advantage of an international study to develop a greater understanding of the factors (that vary from country to country) that contribute to differences in student achievement is lost (see Husén & Postlethwaite, 1995; Kellaghan, 2008).

WHAT PISA PROBABLY MEASURES

In this section, I discuss three issues: the effects of the PISA test format; the interrelatedness of PISA domains; and, the assumption of unidimensionality within PISA domains. I argue that PISA does not measure three distinct and unidimensional competencies, but one general literacy-based domain.

To adhere to the 'real-life' theme, PISA requires tasks to be presented in a context. This hampers the validity of the assessment, as the need to describe context means that the science and mathematics assessments have a heavy reading load. This increases the influence of reading skills on the measurement of student performance in these domains, with the strong likelihood that it contributes to the creation of 'excess reliable variance that is irrelevant to the interpreted construct' (Messick, 1989, p.34). The effects of reading load and reading difficulty have been raised by Bodin (2007) who, writing about PISA mathematics items, suggested that it is unclear if item difficulties derive from understanding the associated text rather than the mathematical problem's degree of difficulty. Ruddock, Clausen-May, Purple, & Ager (2006) in comparing items from PISA, TIMSS, the English national curriculum tests at age 14, and the General Certificate of Secondary Education examination taken at age 16 reached a similar conclusion: 'It is the quantity of reading that marks PISA out, not the complexity of the language

…The high reading demand of questions in PISA is often accompanied by a relatively lower demand in the mathematics or science required' (p. 123).

It would be difficult to have a paper-and-pencil assessment of either mathematics or science that did not presume some level of reading skills. If the level of reading skills required is low, relative to what might be expected in the population in question, then the extent of construct irrelevant variance is generally low, but can be significant for certain subgroups in the population (e.g., newcomer students with limited proficiency in English). If the level of reading skills or the amount of reading required is high, then construct irrelevant variance becomes an issue that affects many students. I would argue that PISA science and mathematics assessments – particularly science – require reading skills beyond a level that we can presume to be shared by most or all of the target population. Thus, the test format employed in PISA means that many students' reading skills have an unnecessarily large effect on how well (or poorly) they perform on the science and mathematics assessments.

While there are exceptions (see O'Leary, Kellaghan, Madaus, & Beaton, 2000), countries tend to perform at a similar level on international studies such as PISA, TIMSS and PIRLS, suggesting considerable overlap in what is measured in such assessments. Perhaps of greater interest is the finding that within countries, performances on the scales or domains of an assessment are strongly correlated (e.g., Rindermann, 2007). Correlations are found, not only at the aggregated, national level, but also at the level of individual students. Irish data for PISA 2006 reveal that while there are strong relationships between performance on PISA domains and performance on the equivalent Junior Certificate Examination subjects, the correlations between PISA domains are stronger than the correlations with their equivalent Junior Certificate subject (Eivers et al., 2008). For example, the correlation between individual student performance on PISA reading literacy and Junior Certificate English is .64, compared to a correlation of .86 between PISA reading and scientific literacy. In a related vein, Bodin's (2007) analyses of French PISA 2003 data noted that the correlation between individual PISA reading and mathematical literacy was much higher than that typically found between French students' results in different mathematical areas (e.g., algebra or statistics). In effect, the skills-based, real-life approach adopted in PISA means that 'the focus was shifted from three school subjects to literacy skills in three areas' (Bonnet, 2002).

Difficulties in determining what exactly PISA measures may stem from the statistical methodology and theoretical perspectives upon which it is

based. As with other comparative studies such as TIMSS, PISA uses Item Response Theory (IRT) to analyze student responses[2]. Analysis for each domain is based on the assumption that the domain being measured is unidimensional, and that the chance of a correct response depends only on the student's competence on the assessed domain and the difficulty of the item (see Goldstein, 1995). Thus, although PISA target domains are quite complex, and students in many cultures are assessed, there is no recognition that a domain may be multidimensional or that an item may be easier in some countries than in others. Items that do not 'fit' into the single dimension or that perform differently in different countries are dropped from analyses. The fact that items that perform unexpectedly (e.g., students who generally perform poorly on the domain are performing exceptionally well on the item) may indicate the presence of a second dimension is not considered. Similarly, items whose difficulty varies unexpectedly between countries become a problem of 'cultural bias', something to be removed rather than explored. What remains are items 'leached of intrinsic interest, comprehensibility, and vitality' (Hilton, 2006).

Rather than letting the data reveal the domains, the domains are superimposed on the data, and anything that does not fit is removed (see Goldstein, 1995). In each domain, this leads to the creation of a unidimensional construct, leaning heavily on reading skills. The high level of similarity in performance across domains suggests that PISA is measuring not three separate domains, but one general domain, using three different areas of content. Some commentators have invoked '*g*' as the common factor in all the assessments. 'Once it is found that PISA mainly measures one general factor per examinee, it is hard not to make a connection to the *g* factor of cognitive psychology' (Wuttke, 2007, p. 260).

CULTURAL FAIRNESS

Efforts are made to make PISA test units as authentic as possible, ideally sourced from real life. However, the requirement for a real-life 'context' makes it difficult to develop culturally neutral items. There are limited contexts that might be equally familiar to students in various OECD countries, and fewer

[2] PISA uses a single parameter, Rasch model, whereas TIMSS uses a 2/3 parameter model, which incorporates a 'guessing' parameter for multiple-choice items. Both methods produce very similar 'national' scores, but differ in how student scores within a country are dispersed, particularly in less developed countries.

again if OECD and non-OECD countries are considered. This section outlines four issues related to the cultural fairness of the test and its equivalence across country and cultural differences: the quality and equivalence of test translations; potential anglophone bias underpinning PISA; differences between countries in how students respond to different types of items; and cultural differences in the importance students accord the assessment.

Quality and Equivalence of Translations

If tests are to be comparable across education systems, translated tests should be similar to the originals. Not only should the meaning be retained, but text difficulty, tone, level of comprehension required, and length should approximate the original. PISA includes a large number of cross-checks to control the quality of translation, and provides what are called 'parallel source versions' of the test booklets in English and French. All translations are subject to multiple reviews and, subsequent to testing, individual items are examined for evidence of different characteristics in different countries. Items that prove unexpectedly difficult or unreliable in a given country are examined to see if there is a translation problem, and may be dropped if this proves to be the case. Thus, it seems reasonable to say that PISA includes many checks to ensure high quality translations.

PISA's use of parallel source versions represents an advance on the procedures used in TIMSS and PIRLS. As both English and French versions are developed in parallel, many potential problems and cultural idiosyncrasies are identified and dealt with before the tests are received by participating countries. Nonetheless, some translation problems remain. Grisay and Monseur's (2007) analysis of PISA reading items from the 2000 assessment concluded, following an examination of items with differential item functioning (DIF)[3], that 'translating a test from a source version *had always at least a basic cost in terms of loss of equivalence*, whatever the quality of the translation'. For example, the number of DIF items when German- and French-speaking cantons in Switzerland were compared was larger than when Ireland and New Zealand or the USA were compared. Thus, similarities based on common language exceeded those based on common experiences and location.

[3] DIF occurs when the difficulty level of an item varies unexpectedly across groups (e.g., Irish students perform better on an item than its overall difficulty level would suggest).

PISA recommends that each participating country should use two translators. One should use the French source version and the other, the English version. The two translations are then compared for discrepancies. However, a number of countries had difficulties finding translators familiar with both the linguistic and scholastic requirements of a PISA translation, while, in other cases, the limited time allocated for translation (see Hambleton, 2002) meant that some countries made no more than a nominal effort at double translation. Thus, some countries could not avail of the cross-checking mechanism that dual translation provides. Apart from the cross-checking facility, it is preferable if countries do not rely on one source version only, as there are considerable differences in the lengths of the French and English source versions of PISA test booklets. Countries that translated entirely or largely from the French versions of the 2006 tests started from a base that is almost 20% longer than the typical English version. Almost identical differences in length were also apparent in PISA 2000 and 2003 (Adams & Wu, 2001; OECD, 2005).

It is difficult to establish the length of tests in languages other than French and English, as most of the materials remain confidential. However, some comparisons of the limited number of PISA items released for general review after each cycle are possible. For example, Puchhammer's (2007) comparison of the English and (Austrian) German versions of released PISA 2003 mathematics items found that the German versions were approximately 15% longer than the English version. Furthermore, the German translations used more 'low frequency' words, meaning that the words were, on average, less commonly used in German than the equivalent English words in English. Some of the science items released after PISA 2006 are available in translated form on the Spanish national PISA website (Spain. Ministerio de Educación y Ciencia, 2007). A comparison of the four science units accessible online reveals that the Spanish versions of the test booklets are approximately 11% longer than the English versions.

As Ireland administers both English and Irish language versions of the test booklets, entire booklets (i.e., the quite lengthy instructions, plus all test items) in both languages can be compared. In PISA 2006, the Irish language versions were, on average, almost 11% longer than the English versions. Indeed, the difference would have been much greater if the reading units had been translated into Irish. The PISA national centres in Germany and Finland also supplied total word counts and number of characters for the German and Finnish versions of the PISA 2006 test booklets. This revealed that the

German versions[4] were almost 17% longer than the English versions, while the Finnish versions[5] were approximately 8% longer. Thus, of five languages examined (French, German, Spanish, Irish, and Finnish) only Finnish comes close to the test length of the English booklets.

Test lengths are relevant for many reasons, not only because PISA is a timed test. With longer tests, students' attention may wander or they may get tired or bored. The PISA 2000 Technical Report compared student responses to reading field trial units, categorized by differences across language in unit length (Adams & Wu, 2001). While the overall percent correct was marginally higher among English-speaking than French-speaking students, the French 'disadvantage' tended to increase as the length disparity increased. Although it was concluded that there may be some effect of unit length on performance, subsequent PISA cycles have continued to show considerable differences in the lengths of the English and French versions of test booklets. Further, Adams and Wu do not appear to have examined the effects of overall test length. If a performance difference is apparent for unit length, it seems likely that larger differences would exist for test length.

The two hours allocated to students to complete a PISA test booklet should provide the average reader with ample time to read the text. However, as PISA students are taking a *test*, average reading speed is typically much slower than normal. This, coupled with the fact that poorer readers have much slower reading speeds, means that weak readers taking PISA mathematics and science tests in some of the 'wordier' languages (such as German and French) may be at a disadvantage vis a vie stronger readers taking the tests in a more succinct language (such as English). As noted earlier, the heavy reading load for mathematics and science introduces construct irrelevant variance. Based on this brief review of test lengths in a number of languages, it is likely that the extent of construct irrelevant variance differs by test language. In particular, the scientific and mathematical literacy of students who are poor readers and who do not take the test in English may be underestimated. From an Irish perspective, the corollary of this is most relevant: PISA may overestimate the mathematical and scientific literacies of English-speaking students with good reading skills.

---

[4] Personal communication with U. Schroeder at the Institut für die Pädagogik der Naturwissenschaften, Kiel, August 14, 2008.

[5] Personal communication with T. Karjalainen and P. Arinen at the Center for Educational Assessment, Helsinki, August 13, 2008.

Anglophone Origins

Some commentators have criticized PISA's anglophone origin and orientation (e.g., Bonnet, 2002; Wuttke, 2007). Roughly half the reading items and almost three-quarters of the mathematics items in PISA 2006 were originally written in English. The percentage of (the more-recently written) science items written in English was much lower (35%), perhaps reflecting a growing awareness of the need to source items from a variety of backgrounds. A large percentage of English-origin items might not be a cause for concern to Irish readers, but it is to those in countries with a different linguistic background. Level of difficulty may change as a result of translation, while subtle changes to familiarity of the settings and language may alter how items function. Bottani and Vrignaud (2005) (in a paper commissioned by the French government) have suggested a broader anglo-saxon/anglophone bias in the theoretical underpinnings of the three PISA domains. They noted that of the original expert groups[6] appointed to guide the development and implementation of PISA, only one was French, in contrast to the large representation from those working in English-speaking countries (including the US, Canada, the UK, Ireland, and Australia).

Murat and Rocher's (2004) analyses of PISA 2000 data showed that countries that shared a common language, culture, or were geographically close tended to exhibit similar patterns of performance. Sorting items by the percentage of correct responses for each country, Ireland was in a cluster of countries that included the UK, US, Canada, New Zealand, and Australia. Responses from Japanese and Korean students were clustered together, as were responses from the Nordic countries and from German-speaking countries. Not only did countries cluster together on individual item difficulty, they also clustered on the difficulty of different types of item. For example, anglophone students were particularly good at items requiring constructed responses. Given these factors, it is understandable that non-anglophone countries may be worried about potential bias in the theoretical bases of PISA and in its implementation, as expressed through test items. On a positive note, it is also one of the easier criticisms to address. Indeed, the planning and implementation of PISA 2009 was divided between two consortia, one of which is led by the Dutch-based CITO. The ongoing efforts

[6] PISA has 'Expert Groups' for each PISA domain, as well as a Technical Expert Group. These groups are composed of subject specialists, providing technical expertise in each assessment domain and expertise in relation to assessment generally.

to encourage participating countries to submit items should also increase the pool of languages and cultures from which test items are developed.

Response Styles

Analysis of PISA 2006 items reveals that, across all countries, item response rates were related to item difficulty and type. Almost all students answered multiple-choice items, while fewer answered constructed response items. For example, examining the OECD averages for PISA 2006 science items, well over 95% of students typically attempted multiple-choice items, while response rates for constructed response items typically varied between 80 and 90% (and fell much lower in the case of very difficult items). However, these global patterns of responding hide significant between-country differences.

Apart from the main national difference of interest (the percentage who supply the correct answer), there are differences in the percentage of students who do not reach the end of the test, who do not answer items, and who tick more than one response to multiple-choice items. As noted by Eivers, Shiel, & Cunningham (2008), the percentage of Irish students who offer any answer is typically higher than the OECD average, even for items on which few Irish students answered correctly. This may be a consequence of PISA timing in Ireland (shortly before the Junior Certificate Examination). As part of preparation for the examination, most students would have been told to always attempt an answer, as marks might be gained for effort. Of course, students do not get marks for effort in PISA, but even random guessing for multiple-choice items gains a score 25% of the time. A 'test smart' student is frequently able to isolate one or more obviously wrong answers, further increasing the likelihood of guessing being a successful strategy.

Contrasting countries include Germany and Austria (both above the OECD mean on the science scale) and Turkey and Italy (both below the OECD science mean). Students in these countries were less likely than Irish students to attempt responses. For example, on science items used in PISA 2006, the percentage of students in these countries who did not answer or provided uncodable answers (e.g., two answers to a single multiple-choice item) to an item was close to twice that of Irish students. Wuttke (2007) and Murat and Rocher (2004) have highlighted the fact that while students in most countries have little difficulty with the multiple-choice format, a significant minority of students in Germany, France, Austria, and Luxembourg supply more than one answer to a multiple-choice item, meaning that they are marked as incorrect.

They believe these national differences are related to familiarity with the multiple-choice format. Thus, Irish students' performance on PISA may be slightly (albeit marginally) assisted by their familiarity with multiple-choice questions and their willingness to guess an answer.

The use of a multiple-choice format also creates difficulty from a language perspective. Generally, Hambleton (2002) has noted that multiple-choice can be problematic for translators, as the organization of subject, verb, and object varies across languages. Specific to PISA, Grisay and Monseur's (2007) analyses of PISA 2000 reading items indicate that multiple-choice items did not function in four of the five participating Asian countries in the way they did in countries using Western languages. They suggested that multiple-choice may be 'more sensitive to large linguistic differences affecting syntax, order of sentence, or direction of writing' and proposed that reviews of the types of problem that lead to differential item functioning should inform future translation guidelines.

Importance Accorded to the Test

The value students place on performance in international studies such as PISA varies considerably, not only between students, but across cultures. Sjøberg (2007) gives the following example of a TIMSS session in Taiwan: '….pupils and parents were gathered in the schoolyard before the big event, the TIMSS testing. The director of the school gave an appeal in which he also urged the students to perform their utmost for themselves and their country. Then they marched in while the national hymn was played' (p. 221). The PISA experience in Irish schools is rather different. In fact, Ireland has one of the highest *student-level* non-participation rates, partly because some students choose to exempt themselves from what they see as a pointless test.

Student lack of interest affects the validity of assessment test results, particularly in low-stakes tests. 'When low-stakes assessment tests are used, the underestimation of student proficiency can be substantial. All low-stakes assessment programs are vulnerable to this threat' (Wise & Demars, 2005). Boe, May and Boruch (2002) examined TIMSS data for cross-national differences in student persistence. They found clear national differences in the questionnaire responses; furthermore, national differences in persistence scores were linked to national variation in test performance. In other words, cultural differences expressed in the persistence shown in completing the *questionnaire* were related to performance on the TIMSS science and mathematics *assessments*.

Recognition of possible effects from the value placed on the test by students led to the inclusion of 'effort thermometers' in PISA test booklets. Students rated their effort on PISA on a 10-point scale, and then rated the effort they would have invested were the test to count in their school marks. Butler and Adams (2007) used PISA 2003 data to analyse student effort, concluding that 'expenditure of effort is fairly stable across a majority of countries'. However, somewhat confusingly, they also concluded that effort should be examined when interpreting trends, as improved German performance in PISA 2003 might be attributable to more effort on the part of German males.

In fact, there are quite large differences in self-reported effort by country in both PISA 2003 and 2006. Even among OECD countries, the average test effort reported ranges from approximately 7 out of 10 for Japan, France, and Norway to closer to 9 out of 10 for Turkey and Mexico. Ireland is fairly average in terms of the effort students report investing in the test. Butler and Adams (2007) found little cross-national differences because they focussed on the difference between the effort students reported making for the test and the effort they would make for school marks, rather than reporting test effort as a standalone indicator. As they note, there are differences in rating styles by country. For example, Japanese students do not give a high effort mark for either test or school effort, while Turkish and Mexican students give high effort marks to both. Thus, a focus on the difference should minimize cultural differences in response patterns. However, even if – like Butler and Adams – we examine only the effort difference, countries still differ to a greater extent than they suggest. In both 2003 and 2006, Norwegian and Japanese students had much larger gaps between average effort reported for test and for school marks than students in most countries. In contrast, Finland was one of the countries where students reported least difference in effort between the two scenarios.

It should be noted that the effort thermometer is a not always a good indicator of the effort students invest. First, there is evidence that students with poor reading achievement had difficulty understanding what was required (Butler & Adams, 2007). Second, Butler and Adams note that 17.5% of students did not complete the effort thermometer, suggesting that the very large missingness might be due to readability issues, although a more plausible explanation might be that the less enthusiastic students did not bother to complete it. Third, the *test versus school marks* difference adds an extra cultural dimension – the extent to which school marks are important. In countries with external high-stakes examinations, internal school marks do not matter in the same way as they do in countries where progression is

linked to school-based assessment. Thus, the effort difference is not cross-culturally comparable as it means different things in different countries.

REPRESENTATIVENESS OF PISA PARTICIPANTS

Debates about the theory/theories underpinning PISA, or how PISA results are interpreted, receive considerably more attention than *who* participated in the survey. However, if participants are not representative of the targetted population, other debates become somewhat irrelevant. In this section, I discuss issues related to sampling methods, response rates, and procedures for dealing with nonresponse, all of which relate to the extent to which the students who do participate truly represent 15-year-olds students in their country. The section draws on PISA Technical Standards (statements of the criteria for acceptable completion of various elements of the assessment). The use of explicit Technical Standards in PISA is a useful tool for countries, as it makes clear precisely what is acceptable and what is not. However, the success of the standards is constrained by the extent to which they are achieved in practice.

Samples

The target population for PISA is all 15-year-olds enrolled in grade 7 or higher in educational institutions in the country[7]. A sample of schools is first selected, followed by a sample of students within each school. Subject to negotiation with PISA sampling experts, national centres may make a small number of exclusions. However, exclusions must not exceed 5% of the target population. Exclusions can take place at either the school-level (e.g., if the school is geographically inaccessible) or the student-level (e.g., because the student has 'an intellectual disability' that would preclude him or her from taking the test).

Given the two-tier approach (schools are selected, then students), sampling needs to be based on accurate information about both school size *and* number of 15-year-olds. Unfortunately, this is not always the case, even in OECD countries. For example, according to the PISA 2006 population coverage details, a rather impressive 110% of the total population of 15-year-olds in both Germany and Italy were enrolled in schools (OECD, 2007). To

[7] Ireland classifies special schools as primary-level institutions. As grade 7 is equivalent to first year in a post-primary school, students in special schools are not included in the target population. This is one of the reasons that school and student-level exclusion rates in Ireland are relatively low.

some extent, this is an excusable problem, as PISA national centres are limited in how much they can compensate for inadequacies in the availability of school- and student-level information in their countries. However, there are also quite clear differences between countries in the extent to which students are excluded from the assessment. Again, some are more excusable than others (e.g., Azerbaijan excluded almost 6% of the population from PISA 2006 as they were in 'occupied regions"). However, in other instances, there are clear breaches of the exclusion upper limit which seem to escape comment. Over 6% of students in both Canada and Denmark were excluded from PISA 2006, yet data for both countries appear in the final report. Exclusion rates for both countries also exceeded the 5% cut-off in PISA 2003.

At least Canada and Denmark declared their breach of the PISA technical standards. In PISA 2000, students (especially males) from a particular type of Austrian vocational school were underrepresented in the sampling frame, and among the subsequent participants. This resulted in inflated Austrian scores (Neuwirth, 2006). However, the error was only revealed when the subsequent 'drop' in Austrian performance in PISA 2003 led to an investigation by the new Austrian government.

Response Rates

Ensuring that a representative sample of a nation's schools and 15-year-olds are selected to take part in PISA is only the first step. It is equally important that most of those selected actually take part. For this reason, it is a PISA requirement that, in each country, at least 65% of the schools initially invited to participate do so, and that following the introduction of replacement schools, the response rate reaches at least 85 percent. Furthermore, a minimum of 80% of sampled students must take part. In 2000, The Netherlands was excluded from analyses for failing to reach these standards, while the UK was excluded for similar reasons in 2003. Other countries have not been excluded, though perhaps they should.

The US has *never* achieved the required school-level participation rate, yet US data have been included in all three cycles of PISA[8]. The PISA Technical Report for 2006 simply states that 'The [US] National Centre provided a detailed analysis of school non-response bias, which indicated no

[8] US data for reading literacy were excluded from the PISA 2006 analyses, not due to failure to reach the required response rates, but because a printing error affecting pagination was deemed to make the reading data invalid.

evidence of substantial bias resulting from school non-response' (OECD, 2009, p. 281). The 2003 Technical Report provides a little more detail, noting that two investigations conducted on the US data concluded that they should be included in the full range of PISA reports. This was despite the use of a second testing period, outside of the approved test window[9], and exclusion rates that far exceeded the 5% cut-off. It is useful to compare the treatment of the US with that of the UK. In 2003, the UK had an initial school response rate of 64.3% (almost identical to the US rate of 64.9%), rising to 77.4% after replacements (considerably higher than the US rate of 68.1%). Only on final weighted student-level response rate was the UK response rate (77.9%) exceeded by that of the US (82.7%) to any notable extent, and that was with the assistance of a higher exclusion rate which broke the cut-point of 5% specified in the technical standards. Wuttke's (2006) pithy comment, 'note: the USA contributes 25% of the OECD's budget' seems apposite.

Until PISA 2006, females were significantly underrepresented among Korean participants (44.1% of Korean participants in PISA 2000 were female, falling to only 40.5% in PISA 2003)[10]. Only in PISA 2006 has the gender composition of the participating Korean students approximated an even split, suggesting that only 2006 data should be considered as representative of Korean students. The gap cannot be attributed to biases in school or student response rates, as both have consistently approached 100% in Korea. Furthermore, while selective abortions did create a slight gender imbalance in Korean birth rates, it is not nearly sufficient to account for the missing females. Korea is not the only country to submit a gendered dataset (e.g., in PISA 2006, four of the 30 OECD countries had female participation rates of under 48%). Nonetheless, Korea's gender gap merits attention because of its magnitude and because Korea is one of very few countries where national performance has shifted significantly between cycles. The OECD and the Korean authorities (OECD, 2007) attribute a 22-point increase on reading literacy between PISA 2003 and 2006 to a new curriculum. However, Korean females have outperformed Korean males on PISA reading literacy in all three assessments (the gap was 35 points in 2006). Thus, much

[9] Three PISA Technical Standards relate to the test period in each country. It must be no longer than 6 weeks and must not coincide with the first 6 weeks of the academic year (unless otherwise agreed) and it must be inside the overall official 'test window'. The US also broke all three of these standards in PISA 2006.

[10] These data were obtained using the open-access PISA online databases for 2000, 2003, and 2006 at www.pisa.oecd.org

of the improvement could be due to increased female participation, rather than to curriculum change.

The purpose of monitoring sampling methods and response rates is to ensure that those that participate are representative of the population. If the characteristics of the participants do not reflect the characteristics of the national population, then the results cannot – or, at least, should not – be used as indicators of national performance.

Adjustment for Nonresponse

Methods for dealing with nonresponse are particularly relevant to Ireland, as Irish students are more likely than students in most other countries to 'opt out' of PISA. For example, Ireland's weighted response rate at the student level was the fifth poorest among participating OECD countries in PISA 2000, the second poorest in PISA 2003, and the third poorest in 2006 (Adams & Wu, 2001; OECD, 2005, 2007). PISA data are weight-adjusted, using stratifying variables, to take account of nonresponse at both the school and student level. In both PISA 2000 and 2003, nonresponse adjustments assumed that school and student non-participants were similar to school and student participants, within weighting classes. In Ireland, this assumption holds at the school, but not the student level. For example, Cosgrove (2005) linked Junior Certificate Examination performance to PISA 2000 and 2003 data. While there was no evidence of nonresponse bias at the school level (i.e., the mean Junior Certificate performance of students in schools that did and did not participate in PISA was similar), there were considerable differences for *student* nonresponse. Her analyses suggest that this led to an overestimation of the achievement of Irish students on reading literacy in 2000 and on mathematical literacy in 2003.

Nonresponse adjustments may also be affected by the extent of between-school variance or by differential participation rates by gender and/or by grade (Monseur, 2007; Monseur & Wu, 2002). This means that in countries such as Ireland (where schools do not differ as much from each other as is the case in countries such as Germany), scores may have been slightly inflated. In relation to differential rates by gender or grade, Monseur (2007) estimated that nonresponse adjustments were biased for 9 of the 32 countries that participated in PISA 2000. The difference between the reading estimates reported by the OECD and his adjusted estimates ranged from -4.96 for Luxemburg to +7.42 for Portugal. (Monseur's adjusted reading score for Ireland was 1.8 points lower than that originally reported by the OECD.) A

member of the PISA Technical Advisory Group, Monseur's analyses have led to changes in the computation of the student nonresponse adjustment in PISA. Such changes, while welcome, do not alter the fact that the estimates reported for Ireland in PISA 2000 and 2003 are likely to slightly overstate the achievements of Irish students.

CONCLUSION

Ireland's reported performance on PISA is likely to be slightly inflated due to the heavy reading load, the fact that we are an anglophone nation, our high rate of student-level nonresponse, and (in PISA 2000 and 2003) by the nonresponse adjustment methods used. However, many of the education systems against which we are likely to compare ourselves are also anglophone or perform well on PISA reading (e.g., England, Northern Ireland, Canada, US, Australia, New Zealand). Thus, we might reasonably consider that PISA provides Ireland with better comparative data for these countries than for culturally and linguistically different countries.

Comparisons, however, have to be considered in the context of the issues identified in this paper. A further, and more fundamental question that arises is: does PISA provide a fair assessment of what education systems accomplish? PISA was created to provide OECD member states with comparative information about the performance of their education systems, on which judgments could be made about the quality of the education that is provided. In addition to issues already raised about aspects of the assessment (cultural fairness, participation rates, and so on), there are a number of reasons why PISA may not be a good measure of the performance of education systems.

First, PISA only attempts to assess three domains of achievement. Although the domains are core subjects, it would be a poor education system that only taught students reading, mathematics, and science. As well as scholastic achievement across a variety of subjects, the traditional view of a 'good' education system encompasses a range of outcomes, often described as 'soft' skills and considered by employers and economists as very important in gaining employment: attitudes, values, motivation, oral presentation skills, the ability to work with others. None of these elements is reflected in PISA.

A second issue in deciding if PISA fairly evaluates education systems is the fact that the study is cross-sectional, not longitudinal, which limits the

inferences that can be made about the extent to which even the limited range of students' scholastic achievement assessed can be attributed to their formal educational experience. As Goldstein (2004) has noted 'To make comparisons in terms of the effects of educational systems, it is necessary (although not sufficient) to have longitudinal data and it remains a persistent weakness of all the existing large-scale international comparative assessments that they make little effort to do so'. A cross-sectional design cannot distinguish the 'net' impact of students' formal educational experiences, which represents outcomes directly attributable to those experiences, from the 'gross' impact which reflects, in addition to net impact, other influences on student achievement (e.g., the value placed on education in a society, students' preparedness for school, the support and assistance provided in the home and community, and participation in 'shadow' education systems).

In conclusion, we may note the observation of Rochex (2006) of a shift in the relationship between research and politics, whereby researchers have changed from adopting a critical stance to becoming politicized 'experts' (although they may view themselves as apolitical, neutral authorities). This may be obscured by the complexity of the statistical methodology which underpins large-scale studies such as PISA, in turn discouraging the less technically inclined from engaging with the project. If Rochex's observation is correct, there is all the more reason why these studies should be subjected to critical analysis. Given the influence that PISA has had on educational policy (especially in countries such as Germany), it seems imperative that educationalists, and indeed the general public, as well as statisticians should be able to engage with the project's assumptions, methods and interpretation.

REFERENCES

Adams, R., & Wu, M. (2001). *PISA 2000 technical report*. Paris: OECD.

Bodin, A. (2007). What does PISA really assess? What does it not? A French view. In S. Hopman, G. Brinek, & M. Retzl (Eds.), *PISA according to PISA: Does PISA keep what it promises?* (pp. 21-56). Wien: LIT.

Boe, E., May, H., & Boruch, R. (2002). *Student task persistence in the Third International Mathematics and Science Study: A major source of achievement differences at the national, classroom, and student levels.* (Research Rep. No. 2002-TIMSS1). Philadelphia: University of Pennsylvania, Graduate School of Education, Center for Research and

Evaluation in Social Policy. Retrieved August 22, 2008 from http://www.gse.upenn.edu/cresp/pdfs/CRESP%20RR%202002-TIMSS1.pdf

Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment. *Assessment in Education, 9,* 387–399.

Bottani, N., & Vrignaud, P. (2005) *La France et les évaluations internationales. Rapports établis à la demande du Haute Conseil de l'Évaluation de l'École. Rapport No. 16*. Paris: Ministère de l'Éducation Nationale.

Butler, J., & Adams, R.J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement 8*, 279-304.

Cosgrove, J. (2005). *Issues in the interpretation of PISA in Ireland*. Doctoral dissertation, NUI Maynooth.

Cosgrove, J., Shiel, G., Sofroniou, N., Zastrutzki, S., & Shortt, F. (2005). *Education for life: The achievements of 15-year-olds in Ireland in the second cycle of PISA*. Dublin: Educational Research Centre.

Dolin, J. (2007). PISA – An example of the use and misuse of large-scale comparative tests. In S. Hopman, G. Brinek, & M. Retzl (Eds.), *PISA according to PISA: Does PISA keep what it promises?* (pp. 93-126). Wien: LIT.

Eivers, E., Shiel, G., & Cunningham, R. (2008). *Ready for tomorrow's world? The competencies of Irish 15-year-olds in PISA 2006. Main report.* Dublin: Stationery Office.

Eivers, E., Shiel, G., & Pybus, E. (2008). *A teacher's guide to PISA science*. Dublin: Educational Research Centre.

Goldstein, H. (1995). *Interpreting international comparisons of student achievement.* Paris: UNESCO.

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

Goody, J. (2001). Competencies and education: Contextual diversity. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 175–190). Göttingen: Hogrefe & Huber.

Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy, 24,* 23-37.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation, 33,* 69-86.

Hambleton, R. (2002). Adapting achievement tests into multiple languages for international assessments. In A.C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 58-79). Washington DC: National Academy Press.

Hilton, M. (2006). Measuring standards in primary English: Issues of validity and accountability with respect to PIRLS and National Curriculum test scores. *British Educational Research Journal, 32,* 817-837.

Hopman, S., Brinek, G., & Retzl, M. (Eds). (2007). *PISA according to PISA: Does PISA keep what it promises?* Wien: LIT.

Husén, T,. & Postlethwaite, T.N. (1996). A brief history of the International Association for the Evaluation of Educational Achievement (IEA). *Assessment in Education, 3,* 129-141.

Kellaghan, T. (2008). IEA studies and educational policy. In W. Harlen (Ed.), *Student assessment and testing* (pp. 394-412). London: Sage.

Kellaghan, T., & Greaney, V. (2001). The globalisation of assessment in the 20th century. *Assessment in Education, 8,* 87-102.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 13-103). New York: Macmillan.

Monseur, C. (2007). *An exploratory alternative approach for student non-response weight adjustment.* University of Liège, Belgium. Retrieved August 22 2008 https://mypisa.acer.edu.au/images/mypisadoc/monseur.pdf

Monseur, C., & Wu, M. (2002). *Imputation for student nonresponse in educational achievement surveys.* Paper presented at International Conference for Improving Surveys, Copenhagen, August 25-27. Retrieved August 22, 2008 from http://www.icis.dk/ICIS_papers/E2_5_2.pdf

Murat, F., & Rocher, T. (2004). On the methods used for international assessments of educational competencies. In J.H. Moskowitz & M. Stephens (Eds.), *Comparing learning outcomes. International assessment and educational policy* (pp. 190-214). London: Routledge Falmer.

Neuwirth, E. (2006). PISA 2000. *Sample weight problems in Austria.* OECD Education Working Papers No. 5. Paris: OECD.

OECD (Organisation for Economic Cooperation and Development). (2001). *Knowledge and skills for life: First results of PISA 2000*. Paris: Author.

OECD. (2005). *PISA 2003 technical report.* Paris: Author.

OECD. (2007). *PISA 2006: Science competencies for tomorrow's world (Vol. 1).* Paris:

OECD. (2009). *PISA 2009 technical report.* Paris: Author.

OECD. (2009). *PISA – the OECD Programme for International Student Assessment.* Retrieved April 27, 2009 from http://www.pisa.oecd.org/dataoecd/51/27/37474503.pdf

O'Leary, M., Kellaghan, T., Madaus, G.F., & Beaton, A.E. (2000). Consistency of findings across international surveys of mathematics and science achievement: A comparison of IAEP2 and TIMSS. *Education Policy Analysis Archives, 8*(43).

Puchhammer, M. (2007). Language-based item analysis. Problems in intercultural comparisons. In S. Hopman, G. Brinek, & M. Retzl (Eds.), *PISA according to PISA: Does PISA keep what it promises?* (pp. 127-138). Wien: LIT.

Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality, 21,* 667–706.

Rochex, J.-Y. (2006). 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.

Rubner, J. (2006). How can a country manage the impact of 'poor' cross-national research results? (A case study from Germany). In K.N. Ross & I.J. Genevois (Eds.), *Cross-national studies of the quality of education: Planning their design and managing their impact* (pp. 255-264). Paris: International Institute for Educational Planning.

Ruddock, G., Clausen-May, T., Purple, C. & Ager, R. (2006) *Validation study of the PISA 2000, PISA 2003 and TIMSS 2003 international studies of pupil attainment.* Nottingham: Department for Education and Science.

Shiel, G., Cosgrove, J., Sofroniou, N., & Kelly, A. (2001). *Ready for life? The literacy achievements of Irish 15-year olds with comparative international data.* Dublin: Educational Research Centre.

Spain. Ministerio de Educación y Ciencia: Secretaria General de Educación. Instituto de Evaluación. (2007). *PISA 2006: Programa para la Evaluación Internacional de Alumnos de la OCDE. Informe Español.* Madrid: Subdirección General de Información y Publicaciones. Retrieved August 25, 2008 from http://www.mepsyd.es/mecd/gabipren/documentos/files/informe-espanol-pisa-2006.pdf

Sjøberg, S. (2007). PISA and 'real life challenges': Mission impossible? In S. Hopman, G. Brinek, & M. Retzl (Eds.), *PISA according to PISA: Does PISA keep what it promises?* (pp. 241-263). Wien: LIT.

Uljens, M. (2007). The hidden curriculum of PISA – the promotion of neo-liberal policy by educational assessment. In S. Hopman, G. Brinek, & M. Retzl (Eds.), *PISA according to PISA: Does PISA keep what it promises?* (pp. 265-294). Wien: LIT.

Wise, S. L., & DeMars, C. W. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 1-17.

Wuttke, J. (2007). Uncertainties and bias in PISA. In S. Hopman, G. Brinek, & M. Retzl (Eds.), *PISA according to PISA: Does PISA keep what it promises?* (pp. 265-294). Wien: LIT.