Policy Futures in Education
Volume 12 Number 7 2014
www.wwwords.co.uk/PFIE

Policy Futures in Education

Issues with PISA's Use of its Data in the Context of International Education Policy Convergence

DAN MURPHY
Melbourne, Australia

ABSTRACT In its analysis of data it collected in 2006, the Programme for International Student Assessment (PISA) survey by the Organisation for Economic Co-operation and Development (OECD) used a variety of statistical methods to arrive at the 'key findings' that school autonomy, school competition and the public posting of student achievement data were associated with higher student performance. The application of one of the methods PISA used – cross-country correlations – to a wider group of school organisation policies found that the consequential accountability measures of using student achievement data to evaluate teachers and allocate resources to schools were associated with worse student performance. PISA applied the cross-country correlation method more consistently to its 2009 data but did not include in its analysis the relationship between using student achievement data to evaluate teachers and student performance. Secondary analysis revealed significant negative relationships between the two. This selective use of data over 2006 to 2009 raises questions about the objectivity of the OECD and PISA and bolsters the views of critics who believe that PISA actively favours convergence around certain schooling policies.

Introduction

In the polarised debates over education policy in many countries, evidence is highly valued. The triennial Programme for International Student Assessment (PISA) of the Organisation for Economic Co-operation and Development (OECD) is widely considered to provide high-quality evidence-based research and policy advice. Its scoring and ranking of how countries perform on its test attract interest around the world, and policy recommendations arising from PISA's analysis of the data it collects carry great weight.

The growth in large national and international education data-gathering exercises such as PISA invites secondary analysis. This can result in original research using existing data sets. Secondary analysis can also be used to apply scrutiny to existing research, as in this study.

Employing high levels of technical expertise with a large budget and staff, and publishing many volumes of detailed analysis, PISA has great influence within the increasingly globalised processes that determine schooling policies. Key to PISA's credibility is a widely held perception that it conducts its work with great rigour and to high scientific standards. As demonstrated by the secondary analysis carried out for this study, PISA's 2006 and 2009 reports lacked consistency in applying statistical methods and reporting results.

Convergence around Particular Schooling Policies

Education policy convergence is a reality in this era of intensifying globalisation. The form that this convergence takes is heavily influenced by influential actors and dominant policy paradigms. International organisations (IOs) with an interest in education are important players in this process, particularly the OECD and the World Bank (WB).

 http://dx.doi.org/10.2304/pfie.2014.12.7.893

Sahlberg (2007) uses the term *global education reform movement* to refer to a set of school policies that have increasingly been adopted in many parts of the world since the 1980s. These are:

1. The standardisation of education – the setting of performance standards and testing to evaluate their attainment through frequent external assessment measures;
2. Increased focus on literacy and numeracy – basic knowledge and skills are considered most important and are also more suited to standard testing and comparison;
3. Consequential accountability – the tying of school performance to the processes of accrediting, promoting, inspecting and rewarding or punishing schools and teachers.

Other policies that Sahlberg states are integral to the globalisation of education policy are decentralisation, privatisation and the drive for increased efficiency in school systems (Sahlberg, 2004).

This global convergence in schools policy has been widely noted. Forsey (2007) writes of a new 'global education policy consensus' in favour of school autonomy linked to efficiency and cost saving and the focus on pre-specified outcomes and measurable competencies. Robertson (2008) describes changes to education systems as a re-mandating which forces teachers and schools to demonstrate their effectiveness through national and global testing systems, use funds more efficiently, increasingly seek private sources of income and accept performance-based 'merit' pay systems. Steiner-Khamsi (2004) notes how of all the possible school reform models, a select few – school choice and competition, outcomes-based education and standards – continue to surface in different parts of the globe.

Carnoy and Rhoten (2002) write of an ideological package consisting of decentralisation and privatisation, choice and accountability and testing and assessment having an impact on education systems worldwide. Hargreaves (2003) and Ball (1998) refer to a new orthodoxy of education to describe this convergence. Rinne et al (2002) observe a new type of consensus among the political elite in favour of decentralisation, goal steering, accountability, managerialism, evaluation, choice, competition and privatisation (p. 643). Hargreaves (2003, 2006); Ball (2006); Green (2006); Grubb and Lazerson (2006); Moutsios (2009), Kamens and McNeely (2010); and Ravitch (2010) are others to have described the spread of education policy with the characteristics described above.

A 2007 survey by the United Nations Educational, Scientific and Cultural Organization (UNESCO) found national assessment regimes were spreading across all regions of the world. Between 1995 and 2006 the number of countries with standardised national testing regimes more than doubled from 28 to 57, with the trend greatest in the developed nations of North America and Western Europe (Benavot & Tanner, 2007). PISA's triennial basis allows for the monitoring of education policy trends over time. For the period between the 2006 and 2009 surveys more than half the participating countries reported a reduction in restrictions on school choice. Twelve OECD countries reported the creation of new more autonomous school models and 10 had introduced new funding mechanisms to promote school choice and competition (OECD, 2010b).

The fullest embrace of these policies, involving well-developed education markets and strong consequential accountability systems, has taken place in England, New Zealand, the United States, Chile, France, Colombia and parts of Australia and Canada. Portugal and Sweden are also proceeding in this direction (Ball, 1998; Sahlberg, 2004). In the United States the No Child Left Behind Act of 2001 required all schools in the country that receive federal government funding to make adequate yearly progress on state-run standard tests with the aim of having every student in the country proficient in reading and mathematics by 2014. Failure to make adequate yearly progress triggers interventions, including requiring the school to facilitate parents' choice to take their children to a school making the required progress, replacing all or most staff, or handing control of public schools to private operators (Hamilton et al, 2007).

The existence of pervasive common global education reforms is thus well established and documented. Not all countries are implementing this agenda and some are implementing parts of it. Nevertheless, there has been indisputable convergence around these policies over the past 30 years.

Debate over Policies around which Policy is Converging

Arguments in favour of these policies are usually advanced in terms of improving educational quality and student outcomes, and making schools more efficient. Sahlberg (2004) writes:

> Standardization-oriented reforms ... are based on the assumption that in competitive economic and social contexts the quality of education ... can best be improved by setting high performance standards for teaching and learning then measuring whether these standards have been met. (p. 72)

Publishing of test results and other indicators places pressure on schools and teachers to improve measured outcomes. Jacob (2005), in a study of test scores in Chicago after an accountability policy was implemented in 1996, found maths and reading achievement as measured by standard tests increased sharply.

Making test information public also enables parents to make an informed choice in the context of a competitive schools market (Apple, 2004). In the early 1990s in the United States, Chubb and Moe's influential *Politics, Markets and America's Schools* (1990) urged policy makers to see choice as a panacea that would make schools more accountable. Competition, it was argued, improves schools' efficiency and quality, as it does in the business sphere. Analysing Sweden's experience with increased school choice, Sandström and Bergström (2002) found school results improved as a result. In Britain, Bradley and Taylor (2002) found strong evidence that the quasi-market in schooling that developed there in the 1990s led to a substantial improvement in efficiency.

Arguments and findings in favour of these pervasive schooling policies are summarised in Hanushek and Wößmann's 2008 paper for the World Bank, *Education Quality and Economic Growth*:

- Consequential accountability including merit pay for teachers will improve student performance.
- Countries with a larger share of private schools tend to perform better.
- Choice and competition puts pressure on teachers and schools to perform, lifting student achievement. Their introduction in the Czech Republic and United States has had positive effects on student performance.
- School autonomy is essential to establish an incentive system and has been shown to improve student learning outcomes.
- Strong accountability systems must accompany choice and autonomy and can lead to better student performance.

Common criticisms of standard testing and consequential accountability systems are that placing such importance on test results leads to teachers and schools teaching to the test and narrowing their curriculum around those areas that are tested. Rather than increased diversity and responsiveness to local needs, what results is greater uniformity. Also neglected are education's broader goals (Sahlberg, 2004).

Ravitch (2010) writes that in the United States, claims by advocates of consequential accountability that it leads to improvements in learning do not stand up to scrutiny. What has emerged, however, is 'gaming' of results through excluding struggling students from tests, selective admission procedures and cheating. The plateauing of results several years after the introduction of national tests in England is, according to Thrupp and Hursh (2006), evidence that teachers adjusted to teach to the test but this approach delivered diminishing returns. They go on to state that 'reforms such as target setting divert attention from more fundamental solutions, they are themselves deeply damaging' (p. 653).

Sahlberg (2004) writes that there is no evidence that school autonomy increases student achievement, and experiences from New Zealand, Chile and the United States where privatisation and competition have been introduced do not support claims that they improve learning. Chile's experience was also examined by Carnoy and McEwan (2000), who found promotion of a schools market did little or nothing to improve academic performance.

Marsh et al's (2011) survey of New York's Schoolwide Performance Bonus Program found no evidence of improved student performance related to the tying of school bonuses and penalties to students' test results. Amrein and Berliner's (2002) analysis of national test results in 18 states in the

USA found student learning remains at the same level or goes backwards when high-stakes testing is introduced.

Analysing experiences in New Zealand and England, Apple (2004) found that school choice and accountability through standard testing are likely to exacerbate differences in access and outcomes based on race, ethnicity and class. He believes that schools rather than families end up exercising choice. That is, schools admit students they believe will enhance their test scores and prestige at the expense of ethnic minorities and students with special needs.

Ball (2006) writes that constant appraisal of their work and tying of salaries to student test results have been criticised for creating demoralisation, insecurity and guilt among teachers, and making the profession less attractive.

Role of International Organisations

The OECD is one of several IOs that increasingly define the process of globalisation, along with the World Bank, the International Monetary Fund and the World Trade Organisation (Lingard & Grek, 2007; Robertson et al, 2007; Moutsios, 2009). For Moutsios (2009), transnational institutions are now the main contexts which define the major educational aims for most countries. IOs, particularly the WB and the OECD, driven in part by what Resnik (2006) calls an' intrinsic tendency to increase their power and resources' (p. 178), have taken a central role in debates over education policy (Kamens & McNeely, 2010). What Grek (2009) and Steiner-Khamsi (2004) call externalisation – the setting of national policy in reference to international examples or norms propagated by IOs – is now a reality in much of the world.

The World Bank

Established after World War II to assist development through loans, the World Bank took a greater interest in education from the 1960s, and by the early 2000s it was the single biggest funding provider for education programmes (Heyneman, 2003). This role gives the bank influence over education policy in countries receiving loans. In the 1980s the bank's 'structural adjustment loans' required recipient countries to implement a policy prescription of deregulation, competitiveness and privatisation, including in education (Robertson, 2008). By the late 1980s, national testing and assessment regimes were bank policy (Robertson et al, 2007) and its 1990 report *Improving Primary Education in Developing Countries* (cited at length in Kamens & McNeely, 2010) was a strong argument in favour of testing and monitoring. The bank's 2008 report *Education Quality and Economic Growth* (Hanushek & Wößmann, 2008) insisted on the importance of the former to the latter, and called for choice and competition, school autonomy, consequential accountability for schools and performance pay for teachers.

The Organisation for Economic Co-operation and Development (OECD)

Ostensibly a policy advisory and research body, the OECD exerts great influence over its 34 member countries and indeed non-members, many of which aspire to developed-country status and membership of the organisation (Porter & Webb, 2007). It has taken an increasing interest in education since the 1980s, and has done so from a strong human capital standpoint. In its own words, 'the prosperity of countries now derives to a large extent from their human capital and, to succeed in a rapidly changing world, individuals need to advance their knowledge and competencies through their lives' (OECD, 2007b, p. 3).

According to Woodhall, 'The concept of human capital refers to the fact that human beings invest in themselves by means of education, training, or other activities, which raises their future income by increasing their lifetime earnings' (1997, p. 219). After being first articulated in the early 1960s, this concept became of great use to education economists and planners as it allowed them to carry out cost–benefit analysis of education funding and calculate returns on investment, both private and social (Woodhall, 1997). Also, in the early 1960s, the United Nations General Assembly adopted a resolution stressing the importance of education for economic development (Resnik, 2006). Today the main multilateral organisations influencing global education policies – the World

Bank, the OECD, UNESCO – subscribe to a human capital view, meaning they see not only education as a measurable investment but consider that its main purpose should be to generate human resources that will provide an economic return to the individual and the nation (Moutsios, 2009). According to Rizvi and Lingard (2006), the OECD is a key promoter of the concept of the knowledge economy – that is, the concept that knowledge is the 'central intrinsic component of economic production and activity' – and accordingly the economic goals of education are given priority over its social and cultural processes (p. 252).

In line with this standpoint, in the 1980s, basic skills and the collecting and comparison of data on inputs and outcomes became the focus of the OECD's education policy research and analysis (Papadopoulos, 1994). The first volume of *Education at a Glance: OECD indicators* was published in 1992 and the first PISA survey conducted in 2000.

The scope and apparent meticulousness of PISA give it credibility and clout (Steiner-Khamsi, 2004; Lingard & Grek, 2007). PISA benefits from being perceived as autonomous and above local politics, and in the words of Grek (2009), it has 'achieved a brand that most regard as indisputable' (p. 25).

The OECD cannot, however, be characterised strictly as a research organisation publishing disinterested analysis. Its reports mix dry analysis with normative writing that contains many assumptions (Rizvi & Lingard, 2006; Robertson et al, 2007; Kamens & McNeely, 2010). For example, on the issue of school choice:

> When parents and students can choose schools based on academic criteria, schools then compete for students, which in turn, may prompt schools to organise programmes and teaching to better respond to diverse student requirements and interests and so reduce the costs of failure and mismatches. (OECD, 2010c, p. 72)

In many observers' eyes the OECD seeks to steer debate, not merely inform it. The measures PISA chooses to collect data on, and conduct and publish analysis of, influence debates, making certain policies visible and obscuring others (Porter & Webb, 2007). PISA is widely considered to be a powerful force for policy convergence around favoured OECD education policies of measurement indicators, basic skills, performance management and accountability (Steiner-Khamsi, 2004; Rizvi & Lingard, 2006; Porter & Webb, 2007; Grek, 2009).

Other Factors Influencing Education Policy Convergence

New Public Management

In many countries, over the past 30 years public-sector management and service provision have been restructured under the guise of what is called 'new public management'. This entails the introduction of business-type models of organisation to the public sector, a push for greater efficiency, empowerment of management at the expense of traditional bureaucratic procedures and ensuring compliance with policy and efficiency goals through targets and benchmarks. The OECD has been a strong advocate of new public management in all areas of public provision (Rizvi & Lingard, 2006; Lingard & Grek, 2007).

Preference for Scientific and Quantitative Modes of Analysis

Kamens and McNeely (2010) write of increasing demand for so-called evidence-based decision making, something that favours quantitative research and analysis over other forms. The United States' No Child Left Behind Act of 2001, possibly the fullest legislative expression of consequential accountability schooling policies anywhere in the world, contains more than 100 references to scientifically based research (Lauer, 2006). This preference feeds into what Martens (2007) calls the comparative turn in education policy formulation. That is, analysis using quantitative indicators invites comparison, from which models of best practice are identified and transplanted into other contexts in a form of 'government by comparison' (p. 54).

Neoliberal Ideology

Neoliberalism is the label commonly applied to a cohesive set of political economic ideas that have strongly influenced policy making in many countries over the last 40 years and that Harvey (2005) sums up as:

> A theory of political economic practices that proposes that human well-being can best be advanced by liberating individual entrepreneurial freedoms and skills within an institutional framework characterised by strong private property rights, free markets and free trade. (p. 2)

It strongly favours the provision of social needs by private interests rather than publicly run services and believes that competitive markets ensure this happens most efficiently. It favours low taxes, particularly on business, and is opposed to organised labour, which prevents workers being paid a market rate. After the economic crisis of the 1970s, neoliberalism replaced Keynesianism as the dominant economic paradigm in developed countries and within the major international organisations, including the World Bank and the OECD (Robertson, 2008; Turner, 2008; Grant, 2009; Crouch 2011).

Neoliberal theorists and advocates have well-developed views on education. In *Capitalism and Freedom* (1962), Friedman wrote of the need for schools to be subject to competition by providing parents with vouchers which could be spent at any school. His schools market would encourage diversity, generate efficiency and pay teachers according to 'merit' and market forces. Ravitch (2010) writes that Friedman's writing underpins the widespread support for school choice in the United States. Kumar and Hill (2009) describe the neoliberal plan for education as marked by:

- Markets and 'parental choice';
- Privatisation;
- Cutting state subsidies.

Writing of the United States context, Hursh (2007) comments, 'Neoliberal ideals, although rarely explicitly stated, form the basis for most of the education reform proposals since… 1983' (p. 497).

PISA

PISA has been administered every three years since 2000 by the OECD. Randomly sampled students complete a two-hour test designed to assess their competency in science, mathematics and reading literacy. PISA claims not to test students' retention of curriculum-based knowledge but instead assesses 'in terms of the acquisition of broad concepts and skills that allow knowledge to be applied' (OECD, 2006, p. 11).

By quantifying individual students' level of proficiency based on their performance on the test, PISA is able to calculate mean national scores for the three subject domains. These national means are the basis of the country rankings or 'league tables' that generate much interest when released every third December. PISA collects a great deal of contextual information on students' backgrounds and school system characteristics. This is quantified to facilitate analysis using a wide variety of descriptive and inferential statistics.

PISA is one of the largest non-experimental research exercises the world has ever seen and the usual cautions that must be attached to the study of phenomena as they occur in the real world apply. In non-experimental research, 'causation should not be inferred from correlation' (Gorard, 2001, p. 153), and findings are usually presented in qualified terms and any action recommended by the researcher based on their findings is commonly limited to further investigation (Johnson, 2001).

Secondary Analysis

Secondary analysis has a well-established place within educational research and has an important role in an era of large data-gathering surveys such as PISA and the Trends in International Mathematics and Science Study (TIMSS). Cook (1974) defines secondary analysis as 'attempts to use existing data from basic research or evaluations to assess the degree of empirical support for major assumptions underlying present practices or alternatives for future policy' (p. 159). Hakim (1982) describes secondary analysis as 'any further analysis of an existing dataset which presents

interpretations, conclusions, or knowledge additional to, or different from, those presented in the first report on the inquiry as a whole and its main results' (p. 1).

Secondary analysis is valuable for the new perspectives and freedom it can bring to analysing and rephrasing questions and asking new ones. It can also bring role independence free of any explicit or implicit influence the funder of the original research may have on what is done with the data (Cook, 1974). The availability of PISA data since the first survey in 2000 has led to greater secondary analysis among educational researchers. Examples include Gorard and Smith's (2004) use of PISA contextual data to investigate school segregation in Europe, Smith's (2007) analysis of PISA results of low-income background students to refute the view that Britain has a long tail of underachievement and Gorard et al's (2006) use of information from the principals survey to investigate patterns of teacher supply.

Secondary analysis is not without its pitfalls. Cook (1974) is critical of Hanushek and Kain's use of data from the landmark 1966 Coleman report *Equality in Education* on African American twelfth graders in northern states of the USA. While they reached findings that differed from those in the original report, Cook says this is due to sampling bias, or the fact that they selected a sub-sample of the data that gave them a dramatically different result from the report's general finding. Cook states that such sampling is suspect when no compelling reason for it is advanced. There is also the criticism that original data used in secondary analysis may not be value neutral. How survey questions were phrased and even what questions weren't asked are relevant matters, and the secondary analyst needs to interpret data with a good understanding of how and why the original data collection was carried out (Dale et al, 1988).

PISA Operations and Data Collection

Fifty-seven nations/education systems participated in PISA 2006, comprising all 30 OECD member countries and 27 partner countries. Although most participating countries sample students from their entire eligible population, China only does so in three discrete areas: Hong Kong, Macao and Shanghai (2009 onwards). Around 400,000 students participated in the 2006 survey, representing approximately 20 million students in the eligible age bracket of 15 years 3 months to 16 years 2 months at the time of assessment. For 2009 participation expanded to around 470,000 students in 65 countries/education systems, representing about 26 million school-goers (OECD, 2010b). This age bracket is selected to provide a sample of students who have completed at least six years of formal schooling regardless of their system's policy on starting age (OECD, 2006). PISA believes it 'monitor(s) the outcomes of education systems in terms of student performance' and 'assesses the extent to which students near the end of compulsory education have acquired some of the knowledge and skills that are essential for full participation in society' (OECD, 2007b, pp. 3, 16).

PISA also collects extensive detailed contextual information from participating students and schools. PISA's test results and collection of contextual data provide the opportunity to assess performance between countries and to investigate relationships between performance and school and social contexts within and between countries. The OECD carries out substantial analysis using the PISA data. For PISA 2006 this analysis was published in one volume and for 2009 it was spread across six volumes. The data are also made public through the PISA website for use in secondary analysis.

Administration

PISA activity is directed by a governing board (PGB) comprising representatives of all OECD member nations. Contracts for the design and implementation of PISA surveys are awarded to international consortia comprised of public and private educational research bodies (OECD, 2011).

The 2006 PISA report was subtitled 'Science Competencies for Tomorrow's World', and the majority of questions tested students' scientific literacy (54%). Reading and mathematical literacy were also examined (15% and 31% respectively). For 2009, PISA returned its focus to reading, as it had been for the initial survey in 2000 (OECD, 2009).

In each country PISA appoints a national project manager (NPM). Working from a national centre (NC), the NPM implements PISA procedures and coordinates school-level activities with

school co-ordinators (SCs), usually a staff member. PISA is overseen within schools by test administrators (TAs), who cannot be science, reading or mathematics teachers of students being assessed and preferably not staff members of any participating school. The test itself lasts two hours and students receive one of thirteen different questionnaire booklets. Tests are required to be administered to students in a uniform manner to the point of the introduction and explanation given by the TA being read from a prescribed text (OECD, 2009, 2011).

Translation and Cultural Appropriateness

PISA is taken in each participating country in the local language of instruction. This includes some regional languages/dialects spoken among sub-sections of the population. For PISA 2006, 87 national versions of the test materials in 44 languages were used and in 2009, 101 versions in 45 languages. This required translation from the source languages, English and French, in accordance with strict procedures. Formatting of national versions of the test booklets has to follow as far as possible the layout in the source versions (OECD, 2009, 2011).

Quality Assurance (QA)

Official operations manuals detail how everyone involved in PISA must perform their function. National-level planning documents are developed by each country from the operational manuals and their content has to be agreed between the NPM and the consortium before being posted on the PISA website. This aspect of PISA's QA process is continually being refined, and for 2009 new spreadsheets and report forms were introduced at this stage to identify issues for follow up and check whether corrections identified earlier had been implemented. Despite this, students' answers to a number of items on particular national versions of the test had to be omitted for the computation of national scores for reasons including mistranslations, confusing translations and poor printing (OECD, 2009, 2011).

In 2006, National Centre Quality Monitors appointed by the consortium visited all 57 participating national centres in the month preceding testing. In 2009 this visiting was limited, with a focus on newly participating countries. PISA Quality Monitors (PQMs) based in participating countries visit a sample of schools (15 per country in 2006, 7 or 8 in 2009) mostly unannounced, to check and record that all aspects of test administration are carried out correctly (OECD, 2009, 2011).

Sampling

The national target population from which PISA draws its samples is all students in the target age bracket (15 years 3 months to 16 years 2 months) who are enrolled in an educational institution. A sample design and size is determined for each country to, in PISA's words, 'maximise sampling efficiency for student-level estimates' (OECD, 2007b, p. 24). As students are sampled in clusters, PISA calculates design effects introduced by the between-school variance and measurement error for participating countries, and from these, effective sample sizes. This explains why countries with similar target populations can require markedly different sample sizes. In 2006 Sweden with a target student population of 127,036 sampled 4443 students while the Czech Republic with a target of population of 122,764 sampled 5932 students. For reading, the design effect, or inflation of the total variance due to measurement error and complex sampling design, was 5.44 for Sweden and 8.38 for the Czech Republic (OECD, 2009).

The target school population in each country is stratified to improve sampling efficiency and ensure adequate representation of specific groups within a population. Stratification variables include school size, public/private, school type (i.e. academic, vocational) and urban/rural. A minimum of 150 schools has to be selected in each country, or all schools if there are less than 150 eligible. School sampling is conducted by the PISA consortium from a sampling frame of all schools containing eligible students supplied by NPMs and stratified according to agreed criteria. Once a school is sampled the SC sends a list of all eligible 15-year-old students to the NPM. From this list,

students are randomly selected to fulfil the target cluster size (TCS), typically 35, or all of those in the target age bracket if the school had fewer eligible students than the TCS (OECD, 2009, 2011).

Maintaining Precision in the Sample

Exclusions at school level are permitted where remoteness, size or operational factors would create unreasonable difficulty. Within schools students can be excluded on the basis of intellectual disability, physical incapacity, psychological unsuitability or limited proficiency in the language of the test. For PISA 2006, the overall exclusion rate was below 4% in 51 of the 57 countries and below 6% in all countries except Denmark (6.07%) and Canada (6.35%) (OECD, 2007b). For PISA 2009, the overall exclusion rate was below 4% in 53 of the 65 countries and below 6% in all countries except Canada (6%), Luxembourg (8.15%) and Demark (8.17%) (OECD, 2011, Table 11.1).

Within a country, 85% of selected schools are required to take part, although replacement schools are allowed if a selected school does not respond. Schools in which less than 25% of selected students participate are regarded as non-respondents and the data from students are not used. Schools with a student participation rate of between 25 and 50% are not considered participating schools but the data from students in these schools are retained and used. A response rate of 80% of all selected students within a country is required (OECD, 2009, 2011).

The PISA Test, Coding and National Mean Scores

The PISA test lasts two hours. Each student receives one of 13 different questionnaire booklets, with no more than three students in any school group of 35 receiving the same booklet. Item formats vary between multiple choice and those requiring a constructed response ranging from a simple word or number to more detailed explanation. For PISA 2006, the 13 different questionnaire booklets were designed such that science questions comprised 54% of all assessment, mathematics 31%, and reading 15%. This was achieved by assigning questions to one of thirteen item clusters (seven science, four mathematics and two reading). Each booklet comprised four item clusters. In 2009 the booklets were comprised of four clusters assigned from seven reading clusters, three maths clusters and three science clusters (OECD, 2009, 2011).

Science

PISA developed a comprehensive scientific literacy framework for the 2006 survey and summarises it thus: 'scientific literacy requires an understanding of scientific concepts, as well as the ability to apply a scientific perspective and to think scientifically about evidence' (OECD, 2007b, p. 21). For the 2009 survey, PISA used the science framework developed for 2006 and selected test items from those used in 2006 (OECD, 2011).

Science questions were framed so as to be relevant and posited within contexts students are familiar with – namely, health, natural resources, environmental quality, hazards and frontiers of science. The three competencies required of students were the ability to identify scientific issues, explain phenomena scientifically and use scientific evidence. Questions were designed to test knowledge both of science and about science and were drawn from the major science fields of physics, chemistry, biology, Earth and space science, and technology (OECD, 2007b). In PISA a unit begins with a text and/or visual stimulus, which is followed by a number of questions. For questions requiring an open-constructed response, full and partial credits can be awarded (OECD, 2007b).

Proficiency Levels and Question Difficulty

PISA's assessment of a student's level of proficiency is based on a scale of performance which, for science assessment, is known as the science scale. All science questions have a position on this scale depending on their degree of difficulty. This allows each student to be given a score which is based on the highest level at which they could be expected to answer correctly a majority of the time.

The scale has six levels and is constructed to have a mean score of 500 and a standard deviation of 100 for OECD countries (OECD, 2007b). The six levels of the science scale correspond to levels of proficiency in the competencies outlined in the science framework. Table I details the six levels of the science scale developed for PISA 2006.

Level	Lower score limit	Percentage of students able to perform tasks at each level or above
6	707.9	1.3% of students across the OECD can perform tasks at Level 6 on the science scale
5	633.3	9.0% of students across the OECD can perform tasks at least at Level 5 on the science scale
4	558.7	29.3% of students across the OECD can perform tasks at least at Level 4 on the science scale
3	484.1	56.7% of students across the OECD can perform tasks at least at Level 3 on the science scale
2	409.5	80.8% of students across the OECD can perform tasks at least at Level 2 on the science scale
1	334.9	94.8% of students across the OECD can perform tasks at least at Level 1 on the science scale

Table I. Six proficiency levels on the PISA science scale 2006.
Source: OECD, 2007b, p. 43.

The relative difficulty of questions is estimated by considering the proportion of students getting each question correct (OECD, 2007b). It is possible to estimate the location of individual students on the scale by giving them a score according to the hardest task they could be predicted to perform with a given probability. Students are placed on the scale at the point where they have a 62% chance of correctly answering questions located at the same point (OECD, 2009).

Reading

Based on the framework developed for PISA 2000, the reading literacy competencies tested for are the ability to:

- Retrieve information;
- Interpret texts; and
- Reflect on and evaluate texts (OECD, 2007b).

Of the 131 reading items used in PISA 2009, 94 were newly developed, with the remainder having been used in earlier surveys (OECD, 2011). PISA's combined reading proficiency scale has five proficiency levels, with a mean score of 500 points and a standard deviation of 100 (OECD, 2007b).

Mathematics

PISA defines mathematical literacy as 'an individual's capacity to identify and understand the role that mathematics plays in the world, to make well founded judgements and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen' (OECD, 2007b, p. 304). As with science and reading, PISA tests students' literacy in mathematics in an applied context and is not simply concerned with whether they can perform basic maths functions. The scale used to assess question difficulty/student proficiency in mathematics has six levels and a mean of 500 points with a standard deviation of 100. This scale and the framework for assessing mathematical literacy were developed ahead of PISA 2003, which had mathematical literacy as its focus (OECD, 2009).

Coding of Students' Answers

Uniformity in coding student responses to open-ended items is crucial to PISA's validity and considerable effort is made to achieve this. NPMs have to translate national coding guides and submit them to the consortium for verification. They then have to recruit and train coders, who must have a good understanding of either mid-secondary-level mathematics or science or the language of the test and need to be familiar with ways in which secondary-level students express themselves (OECD, 2009, 2011).

Completed test booklets are randomly allocated to coders. Each test booklet contains four clusters. A coder is allocated a cluster and then codes it in four different booklets before moving on to another cluster. The design for allocating clusters to coders means that each student's booklet is coded by four different coders (one for each cluster), a way of minimising the effects of any systematic leniency or harshness (OECD, 2009, 2011).

Weighting

For the purposes of calculating distributions of students' proficiency levels and national mean student scores, it is necessary to weight each student's score so that they represent the correct number of students in the total population. While students are chosen randomly, they do not have the same probability of being chosen. Reasons for this include over-representation of schools with a particular interest (i.e. ethnic or indigenous groups) in the sample design, schools turning out to be larger or smaller than expected (affecting the probability of being sampled) and non-response. With the correct weightings applied, scores of all participating students within a country can be used to calculate a national mean score for each country (OECD, 2009).

Student Contextual Information

Following a short break at the end of the two-hour test, students complete a 30-minute context questionnaire. This asks them to supply information about their background, including their parents' occupation and education level as well as whether their homes contain specified indicators of wealth and educational resources (OECD, 2007b). Indices are constructed for these variables and standardised to have a mean value for the OECD student population of zero and a standard deviation of one (OECD, 2007b).

In its 2006 analysis PISA utilised a combined index of economic, social and cultural status (ESCS) to help define student background. This comprised the index of parents' highest educational level (father or mother), the index of highest international socio-economic status of occupational status (father or mother) and the index of home possessions based on the survey of household possessions and educational resources. The ESCS is used in PISA's analysis of the impact of socio-economic background on student performance. In 2006 it found that across OECD countries, an increase in one standard deviation on the ESCS index was associated with a science score that was 40 points higher (OECD, 2007b). In 2006 PISA also carried out a hypothetical adjustment of mean national science scores assuming the mean ESCS was equal across all OECD countries. This adjusted scores downwards for countries with an above-average ESCS and upwards for countries with a below-average ESCS (OECD, 2007b).

For PISA 2009 the national average ESCS ranged from -1.22 in Mexico to 0.47 in Norway and 0.72 in Iceland within the OECD and from -1.55 in Indonesia to 0.42 in the United Arab Emirates and 0.51 in Qatar among partner countries (OECD, 2009). PISA carried out an adjustment of national mean student reading scores, but whereas for 2006 this adjustment (to science scores) was made if the mean ESCS would be equal in all countries, for 2009 a predicted reading performance for a student with a socio-economic background equal to zero (the OECD average) was calculated for each country.

School Contextual Information

Some basic information such as school size or proportion of females enrolled is contained in enrolment data submitted by the school. Other information sought by PISA is obtained through a questionnaire completed by principals and includes:

– Admittance policies: the degree to which factors such as residence in the local area, academic record and parents' endorsement of the school's religious philosophy determine students' admittance

– School type: public, private, or government-dependent private.

> – Autonomy: Whether decisions about resources and curriculum were made by principals and teachers, a school board or regional or national authorities.
>
> – School Choice: whether the school competes with others in the local area for students, and if so the number.
>
> – Accountability Arrangements: whether achievement data were tracked over time by an administrative authority, whether such data were used in the evaluation of the teachers' or principal's performance, whether the data was posted publicly and whether such data were used in decisions about instructional resource allocation to and within the school. (OECD, 2007b)

Issues with PISA Methodology

High levels of expertise and great technical capacity are applied to the world's largest study of schooling. In its analysis and technical volumes, PISA explains in great detail how it quantifies student performance and contextual information to enable international comparison. PISA is confident that its methods are sound and its measures of student performance are valid, and is strident in defence of these (Adams, 2003).

At the same time, PISA advises a general caution in interpreting its results as many important contextual factors cannot be captured by surveys of this kind and thus cause and effect cannot be firmly established. The learning environment of a 15-year-old at the time they undertake PISA is not necessarily the same as that which shaped their learning earlier in their education, and as such the contextual data collected are an 'imperfect proxy for the cumulative learning environments of students' (OECD, 2007b, pp. 214-215). Smith (2007), while using PISA data for secondary analysis, sounds a general caution on the limits of cross-national education surveys such as PISA, as:

> [t]he virtual absence of cross national educational variables that predict student achievement suggest it is difficult to separate the effect of educational policies and instructional practices from the contexts in which they are developed and implemented. (p. 36)

PISA also warns of the need to take into account the limitations of the contextual data gathered through principals' surveys. This is because on average only 300 principals were surveyed in each country in 2006 and 264 in 2009, and the data gathering relies on a principal's assessment of school resource matters such as class size, for which they may not have accurate or up-to-date data (OECD, 2007b). Caution in regard to the results of the principals' survey should surely extend further to note that they are based on principals' subjective responses to being questioned as to whether they had 'considerable responsibility' for determining such things as course content. In its 2009 analysis PISA acknowledged wider limitations on the value of data collected through its principals' survey (OECD, 2010c).

The age of the target population (15 years and 3 months to 16 years and 2 months) is criticised on the grounds that, particularly in developing countries, many students have already left schooling (Prais, 2003) and the ability of abstract reasoning is still developing (Wuttke, 2007). Prais (2003) writes that mathematics education in countries such as Germany is likely to be more conceptual at this age compared with Britain, where maths is compulsory only to age 16 and likely to be taught to 15-year-olds with greater emphasis on real-life applications of the kind PISA assesses, and that using age rather than schooling grade to define the target population means children could be spread across up to three grades and at very different stages of development. Adams (2003) defends this choice on the grounds that cross-country differences in pre-school provision and school starting ages mean different grades would have to be selected in different countries, creating a greater comparability problem.

In discussing variation in PISA 2003 student response rates of up to 15% across countries, Wuttke (2007) writes that 'non-response bias can be considerable because the propensity of school principals and students to partake in testing is likely to be correlated with the potential outcome' (p. 9). Prais (2003, 2004) agrees with this likely non-response bias and raises concerns with PISA's method of calculating its response rate, in which replacement schools taking part are included in the calculation of a response rate based on the number of schools that ended up participating as a proportion of the initial school sample size. For Britain in 2000, the initial school response rate was 61%. Fifty-five per cent of replacement schools approached then also took part. PISA gave its

response rate as 80%, while Prais believes it should be defined as 57%. Prais (2003) is also critical of PISA's exclusion of schools with 25-50% participation from school-level calculations while the students' individual results were used in the calculation of mean national scores. Adams (2003) acknowledges that the response rates of Britain, whether calculated before or after replacement, were of concern but claims that the direction and magnitude of non-response bias cannot be known and analyses the academic attainment of non-responding British government schools, finding no evidence that they were on average lower achieving. PISA 2003 response rates in Britain and the USA were, according to Smith (2007), too low for findings to be generalised to the larger population.

In the 2009 survey, school response rates after replacement were below the required level of 85% in the United States and Panama, and the student response rate was below the PISA-determined 80% requirement in Canada. Despite this, and after what it terms additional analysis, PISA retained results from these countries in the final database and used them in its analysis (OECD, 2011).

Another target of scholarly criticism is the survey's reliance on item response theory – the assumption that the probability of a correct response depends only on the differences in students' competence and the item's degree of difficulty – with insufficient consideration of the impact of non-competence measures, such as items being easier for one subpopulation than another (Wuttke, 2007). The related problem of cultural and linguistic bias is raised regularly (Prais, 2003; Goldstein, 2004; Dohn, 2007; Hopmann & Brinek, 2007; Wuttke, 2007).

Goldstein (2004) questions the level of insight provided by the orientation towards mean national scores and country rankings based on relative, not absolute, measures of achievement. Dohn (2007) criticises PISA's failure to verify that statistical levels on its scales of difficulty correspond to actual ones. Prais (2003, 2004) calls for PISA to publish country figures on the correct average scores for each question for a clearer view of countries' educational weaknesses and strengths and to make PISA's scaling technique more transparent. Wuttke (2007) calls PISA's proficiency levels 'arbitrary' and the creation of its scales 'ill documented renormalisation' (p. 17).

Other criticisms include gender imbalances in PISA 2003, when only 40.5% of participants in South Korea were girls, and subjective judgement of, and widely varying levels of, exclusions due to intellectual impairment (Wuttke, 2007).

PISA 2006

Student Performance and School and System Characteristics

As part of its study of school and system characteristics and student performance in science, PISA collected data on the implementation of many school organisation matters, including resources (i.e. class sizes) and grouping policies, and investigated the relationships between these and student results on the PISA test. Relationships between school organisation policies and student performance were investigated using a variety of statistical methods.

Included as 'Key Findings' in PISA 2006's Executive Summary were the following observed relationships:

- 'Across countries, having a larger number of schools that compete for students is associated with better results, over and above the relationship with student background';
- After other factors are taken into account 'there still remains a significant positive association between schools making their achievement data public and having stronger results';
- 'Students in countries where autonomy is more common tend to do better in the science assessment, regardless of whether or not they themselves are enrolled in relatively autonomous schools' (OECD, 2007a, pp. 7-8).

School Competition

PISA uses individual, school and system/country-level data in its analysis. In its 2006 analysis of school competition effects, PISA determined a school with a high level of competition as one competing with one or more other schools for students, based on responses to the principals'

survey. Levels of school choice, defined in this way, varied from 95% of students in Indonesia attending competing schools to 34% in Norway. The OECD average was 75% (OECD, 2007b).

Through a comparison of means, PISA found students at schools with a high level of competition performed on average 17.9 points better on the science scale than those enrolled in non-competing schools. This difference, however, was found not to be statistically significant after accounting for demographic and socio-economic background factors. In analysing school competition at the national level, PISA looked at the change in national mean science scores per countries having an additional 10% of competing schools. After accounting for demographic and social factors, this difference was found to be 6.7 points and statistically significant. (Note: when analysing the effect of system-level factors using this method, PISA considered results statistically significant at the country level if the *p*-value was below 0.1.) This result is the basis of the PISA 2006 finding on school competition (OECD, 2007b).

School Accountability Policies

Posting school data publicly was one of nine accountability policies whose relationship with science performance PISA investigated. Based on answers to the principals' survey, schools were classified as posting achievement data publicly or not. The change in score for schools posting was 14.7. After accounting for demographic and social factors this difference was 6.6 and remained statistically significant (OECD, 2007b). The same analysis method was used to investigate the relationships between science results and whether schools implemented a variety of accountability policies. These were:

- School informing parents of children's performance relative to other students in the school;
- School informing parents of students' performance relative to other schools;
- School informing parents of students' performance relative to national benchmarks;
- School using achievement data for evaluating principals;
- School using achievement data for evaluating teachers;
- School using achievement data for allocating resources;
- School with achievement data tracked over time;

For all these variables, no significant relationship was observed.

School Autonomy

The principals' survey asked whether schools had considerable responsibility, whether both schools and regional and/or national education authorities had considerable responsibilities, or whether only regional and/or national education authorities had considerable responsibilities for the following aspects of school management:

- Selecting teachers for hire;
- Dismissing teachers;
- Establishing teachers' starting salaries;
- Determining teachers' salary increases;
- Formulating the school budget;
- Deciding on budget allocations within the school;
- Establishing student disciplinary policies;
- Establishing student assessment policies;
- Approving students for admission to the school;
- Choosing which textbooks are used;
- Determining course content;
- Deciding which courses are offered.

Based on answers to these survey questions, PISA developed three indices of school autonomy: school autonomy in staffing; school autonomy in budgeting; and school autonomy in educational content. In analysis at the school level, autonomy in educational content and budgeting showed no significant relationship with student performance. A significant positive relationship between

autonomy in staffing and student performance was negative (although not significant) after demographic and social factors were accounted for (OECD, 2007b, pp. 252-253).

PISA also investigated the relationship between autonomy and student performance at the system level. Cross-country correlations were calculated between the percentage of schools in countries having considerable responsibility ('school only' and 'school and government') for each of the above and student performance in science. As Table II shows, the correlations were positive and statistically significant in the case of every variable except establishing and determining teachers' salaries. This predominance of statistically significant positive correlations between these various aspects of school autonomy and science performance is the basis for the finding that 'the data suggest that in those countries in which principals reported, on average, higher degrees of autonomy ... the average performance in science tends to be higher' (OECD, 2007b, p. 249).

Aspect of school management	Cross-country correlation between the percentage of schools having considerable responsibility ('school only' and 'school and government') and science performance
Formulating the school budget	0.47*
Dismissing teachers	0.32*
Selecting teachers for hire	0.43*
Establishing teachers' starting salaries	0.20
Determining teachers' salary increases	0.22
Deciding on budget allocations within the school	0.54*
Establishing student disciplinary policies	0.41*
Establishing student assessment policies	0.43*
Approving students for admission to the school	0.27*
Choosing which textbooks are used	0.51*
Determining course content	0.52*
Deciding which courses are offered	0.58*

*Correlation coefficient is statistically significant at the 5% level ($p<0.05$).

Table II. Correlations between degree of countries' adoption of selected school autonomy measures and student science performance, PISA 2006.
Source: OECD, 2007c, Table 5.10, pp. 175-178.

At the system/country level, PISA also calculated the effect of a one-standard-deviation increase in the system average of the three school autonomy indexes in relation to student performance, finding significant positive relationships in the case of budgeting and educational content autonomy (OECD, 2007b, p. 253).

Secondary Analysis using PISA 2006 Data

PISA 2006 collected student, school and system/country-level contextual data and its findings are based on analysis of data collected at all of these levels. In the case of autonomy, school-level analysis had found no significant relationship with student performance but country-level analysis, including correlations, did, enabling PISA to make a 'key finding'. Multi-level analysis – school and system/country – was also carried out to investigate the relationship between school competition and student performance. Although no significant relationship was observed at the school level (after accounting for demographic and social factors), school systems with a higher proportion of competitive schools had significantly higher student performance.

For reasons not explained in the PISA 2006 *Analysis* volume, this multi-level analysis was not used to investigate the relationship between accountability arrangements and student performance. For accountability arrangements, PISA found a positive relationship involving one policy – posting achievement data publicly – analysed at the school level and left it at that. No cross-country-level analysis was carried out, or if it was, it was not published.

Investigation of the PISA 2006 *Data* volume and the PISA website revealed the availability of data on the degree to which countries implemented a number of school organisation policies. This

was expressed as the percentage of students within a country studying at schools which operate under a given policy, based on school principals' answers to survey questions. This availability enabled the calculation of cross-country correlations between the degree of implementation of selected school accountability and other policies and performance in science, reading and mathematics. This would expand the use of cross-country correlations very similar to those used to investigate the effect of a range of school autonomy variables on student performance to a broader range of schooling policies:

- The percentage of children in schools posting achievement data publicly;
- The percentage of children in schools where the governing board exerts direct influence over decision making about staffing. This is a measure of local autonomy and independence from centralised control. Other measures of autonomy are not included, as PISA published cross-country correlations between these and student performance.
- The percentage of children in private schools;
- The percentage of schools competing for students with two or more other schools in the same area;
- The percentage of students in schools where the principal reported that achievement data are being used in decisions about instructional resource allocation to the school. This is a measure of the prevalence of consequential accountability measures targeted at schools;
- The percentage of students in schools where the principal reported that achievement data are being used for evaluation of teachers' performance. This is a measure of the prevalence of consequential accountability measures targeting teachers, including performance pay.

Relationships between these policies and student performance were analysed in all three subject domains and using the PISA-derived mean science scores if the mean ESCS would be equal in all OECD countries. Correlations were tested for by calculating the Pearson correlation coefficient (r) between the levels of implementation of the selected policies and student performance. A significance level of 0.05, the same level applied by PISA, was used to test the significance of correlations, although it is also reported where results are significant at the $p<.01$ level.

Results of Secondary Analysis using 2006 Data

Results of correlations can be seen in Table III. The significant positive cross-country correlations between performance and percentage of schools competing with two or more other schools were consistent with PISA's system-level finding on competition.

	Science performance	Reading performance	Maths performance	Science performance if ESCS was equal across all OECD countries
Percentage of students in schools where:				
Student results posted publicly	-0.09	-0.14	-0.06	-0.12
Governing board exert influence over staffing decisions	0.21	0.16	0.21	0.26
School competing	0.32*	0.31*	0.31*	0.33*
Achievement data used for teacher evaluation	-0.54**[*]	-0.56**	-0.52**	-0.47**
Achievement data used for resource allocation	-0.41**	-0.37**	-0.40**	-0.37**
Percentage of students in private schools	0.21	0.25	0.23	0.29*

*Correlation is significant at the .05 level; **correlation is significant at the .01 level.

Table III. Correlations between level of implementation of school organisation policies in participating countries and student performance, PISA 2006, all 57 participating countries.
Data source: OECD, 2007b, p. 184; OECD, 2007c, pp. 165, 168, 173-174, 180.

The correlations between the implementation of the two consequential accountability policies and student performance are negative and significant for all measures of student performance. School-level analysis of the effect of these had shown no significant relationships.

Summary

PISA 2006 employed a variety of statistical methods and used data aggregated at different levels and different significance tests to arrive at its key findings on relationships between school and system characteristics and student performance. Not adequately explained in PISA's official documents are the reasons for the variation in statistical methods and level of data used.

This secondary analysis applied the cross-country correlation method PISA used with a range of autonomy variables to a wider range of schooling policies and their relationship with student performance. These included consequential accountability policies, the introduction of which is often controversial and vigorously debated. The use of student achievement data for teacher evaluation would encompass performance pay systems which are commonly forced on teachers against their will. The use of student achievement data for resource allocation would include a lot of measures contained in the United States' No Child Left Behind Act, which has been the subject of much scholarly debate.

If PISA had applied its analysis methods consistently to its 2006 data, using its own criteria for what constitutes a key finding, it would have been able to reveal two more such findings:

- Students in countries where the use of student achievement data for teacher evaluation is more common tend to do worse; and
- Students in countries where the use of student achievement data for resource allocation is more common tend to do worse.

2009

PISA's Analysis and Results

Reading was the main focus of the 2009 survey, with science and mathematics also tested. Again, students and school principals completed the questionnaires that provided the contextual information used by PISA in its analysis.

In its analysis of relationships between school organisation policies and learning outcomes, PISA made greater use of cross-country correlations with the 2009 data. In doing so it carried out a number of analyses using the 2009 data that I had first carried out using the 2006 data and presented as part of a master's thesis seminar at the University of Oulu in November 2008. The PISA 2009 Table IV.2.1, *Correlations between System-level Characteristics and Educational Outcomes* (OECD, 2010c), can be considered an expanded version of Table III in this article, used to present results of correlations using PISA 2006 data.

In making broad use of the method I had employed and discontinuing many of its more detailed statistical analyses, PISA 2009 achieved a consistency in its methods that was notably absent in 2006. PISA expresses its confidence in the 'robustness and sensitivity' of its findings and states that 2009 system-level correlations were found to be consistent with those from 2006, which were arrived at using more sophisticated statistical techniques (OECD, 2010c, p. 30).

As can be seen in Table IV, PISA carried out correlations between reading performance and countries' average index of school responsibility for curriculum and assessment and average index of school responsibility for resource allocation.

The index of school responsibility for curriculum and assessment is a composite measure of schools' autonomy over establishing student assessment policies, choosing textbooks, determining which courses are offered and the content of those courses. The index of school responsibility for resource allocation is a composite measure of schools' autonomy over appointing and dismissing teachers, establishing teachers' starting salaries and salary raises, formulating school budgets and allocating them within the school. These indices are created to have a mean of zero and a standard deviation of one for all OECD countries and are calculated using principals' questionnaire answers.

From these, PISA calculated national average scores on the indices and used these in its cross-country correlations with student reading performance.

	Reading performance – all countries	Reading performance – OECD countries
School governance		
Average index of school responsibility for curriculum and assessment	0.46*	0.45*
Average index of school responsibility for resource allocation	0.25	0.02
Percentage of students in private schools	0.14	0.05
Percentage of students in schools that compete with other schools in the same area	0.16	0.06
Assessment and accountability policies		
Percentage of students in schools that assess students with standardised tests	-0.05	0.15
Percentage of students in schools that use assessment or achievement data to:		
Compare the school with other schools	-0.21	0.02
Allocate resources	-0.47*	-0.08
Monitor teacher practices	-0.46*	-0.17
Post achievement data publicly	-0.11	0.04
Monitor progress over time	-0.29*	-0.09
Have their progress tracked by administrative authorities	-0.36*	-0.14

*Significant at the 5% level ($p < .05$).

Table IV. Selected results from PISA 2009 Table IV.2.1,
Correlations between System-level Characteristics and Educational Outcomes.
Source: OECD, 2010c, Annex B1

PISA calculated that the correlation coefficient between reading performance and average index of school responsibility for curriculum and assessment was 0.46 across all countries and 0.45 across OECD countries, and that in both cases it was statistically significant at the 5% level ($p<0.05$). In 2009, PISA introduced calculations of each of the cross-country correlations after accounting for the effect of gross domestic product (GDP) per capita on reading performance in each country – an adjustment based on average national income rather than using national student average ESCS, for which PISA had collected data. After accounting for national income, correlations between average index of school responsibility for curriculum and assessment and reading performance remained positive and significant for all countries and within the OECD. This, PISA found, was a 'clear relationship' that enabled it to find that '[i]n countries where schools have greater autonomy over what is taught and how students are assessed, students tend to perform better' (OECD, 2010b, pp. 14, 41). No significant relationship was observed between greater autonomy to allocate resources and student performance (OECD, 2010b).

Significant negative correlations were observed between reading performance and the percentage of students attending schools that use student assessment or achievement data for the following accountability purposes across all participating countries:

- To monitor progress over time ($r = -0.29$);
- To have their progress tracked by administrative authorities ($r = -0.36$);
- To allocate resources ($r = -0.47$);
- To monitor teacher practices ($r = -0.46$).

The correlation between schools using student assessment or achievement data to allocate resources and student performance remained significant after accounting for national income. When calculated for OECD countries only, these relationships were not significant (OECD, 2010c, Annex B1).

Secondary Analysis using PISA 2009 Data

Through the six PISA 2009 volumes and the PISA website, data were made available that allowed the same secondary analysis to be carried out as was done using PISA 2006 data. The same

correlations as in 2006 were calculated between the selected school organisation variables and performance in the three subject areas (national average scores in reading, maths and science) and a PISA-predicted reading score for a student with an ESCS equal to zero for each country. There was one exception, with data on the proportion of students in schools where the governing board exerts direct influence over decision making not being available. PISA's adoption of cross-country correlation to analyse relationships between school organisation policies and student performance, in addition to vindicating my use of the method, meant that my secondary analysis repeated some correlations PISA had already carried out. In all cases bar one the results were identical: across all countries, correlation between reading performance and percentage of students in schools that compete varied by .01.

Results of Secondary Analysis using 2009 Data

The results of secondary analysis using PISA 2009 results are collated in Table V.

	Science performance	Reading performance	Maths performance	Predicted reading score for a student with ESCS = 0
Percentage of students in schools where:				
Student results posted publicly	-0.11	-0.11	-0.08	-0.18
School competes with other schools in the same area	0.16	0.17	0.14	0.23
Student achievement or assessment data used for teacher evaluation	-0.49**	-0.50**	-0.49**	-0.46**
Student achievement or assessment data used for resource allocation	-0.48**	-0.47**	-0.48**	-0.41**
Percentage of students in private schools	0.14	0.14	0.14	0.16

*Correlation is significant at the .05 level; **correlation is significant at the .01 level.

Table V. Correlations between level of implementation of selected school organisation policies and student performance, PISA 2009, all participating countries.
Data source: OECD, 2010a, Annex B1, Tables I.2.3, I.3.3, I.3.6; OECD, 2010b, Annex B1, Table II.3.2; OECD, 2010c, Annex B1, Tables IV.3.8a, IV.3.9, IV.3.13.

	Science performance	Reading performance	Maths performance	Predicted reading score for a student with ESCS = 0
Percentage of students in schools where:				
Student results posted publicly	-0.03	0.03	-0.11	-0.02
School competes with other schools in the same area	0.08	0.06	0.00	0.16
Student achievement or assessment data used for teacher evaluation	-0.30	-0.31	-0.39*	-0.12
Student achievement or assessment data used for resource allocation	-0.19	-0.08	-0.32	-0.04
Percentage of students in private schools	0.04	0.05	-0.01	0.09

*Correlation is significant at the .05 level.

Table VI. Correlations between level of implementation of selected school organisation policies and student performance, PISA 2009, OECD countries.
Data source: OECD, 2010a, Annex B1, Tables I.2.3, I.3.3, I.3.6; OECD, 2010b, Annex B1, Table II.3.2; OECD, 2010c, Annex B1, Tables IV.3.8a, IV.3.9, IV.3.13.

The cross-country correlations between the percentage of students in schools using student achievement data for teacher evaluation and resource allocation and student performance were negative and significant for all measures of performance. When repeated for OECD countries only,

as shown in Table VI, the correlation between the use of achievement data for teacher evaluation and mathematics performance was negative and significant.

Summary

By adopting the broad use of cross-country correlations to analyse relationships between school organisation policies and student performance, PISA achieved a consistency that it lacked in 2006. PISA reported on how resources, policies and practices are related to student performance in Volume IV of the six 2009 analysis volumes. In the case of school autonomy over educational matters, positive significant correlations with reading performance across all countries and within the OECD, both before and after accounting for national income, form the basis of the finding that such autonomy is related to better student performance. In its discussion of this finding, PISA references the results table IV.2.1 in Annex B1 of Volume IV, that contains the relevant correlations.

The decision to leave the results of other correlations involving school organisation policies in that table within the annex makes less visible the negative correlations across all countries between implementation of certain school accountability policies and reading performance, which remain in the results table, unremarked upon. As they involve some of the accountability policies around which there is global convergence and much debate, arguably the significant results should have been given more prominence and discussion in Volume IV.

With PISA's broad adoption of the cross-country correlation method I first used on 2006 data, repeating the secondary analysis using some of the same variables involved crossover with PISA's work, although with one notable exception. PISA did not perform the cross-country correlation between the use of student achievement data for teacher evaluation and reading performance, or if it did, it did not publish the results. This is unsatisfactory as PISA collected the data to enable this analysis to be done and it could have been added to the list of correlations in Table IV.2.1. This table included a correlation between reading performance and the use of student achievement data to monitor teacher practices, a policy which was not mentioned in PISA's 2006 *Analysis* volume and which does not attract the same interest from researchers and educators as use of achievement data for evaluating teachers, which encompasses the use of teacher performance/merit pay. The negative and significant correlation between implementation of this policy and all measures of performance across all countries, and the significant negative correlation with mathematics performance within OECD countries revealed by the secondary analysis, suggest that, where it is implemented, students may tend to perform worse. At the very least, the results suggest that this policy should be treated with caution and that its relationship with learning outcomes should be subject to more study.

Conclusion

Despite its many critics, PISA retains an influential position in shaping schooling policy around the world. Its findings and recommendations carry great weight in national policy debates, where they are usually cited absent the cautions on interpretation contained in PISA reports.

This secondary analysis of PISA data shows that consistent application of its cross-country correlation method to school organisation policies using 2006 data would have revealed two more 'key findings': that the consequential accountability policies of using student achievement data to evaluate teachers and to allocate resources were associated with worse student performance.

In PISA 2009, the more consistent use of cross-country correlations to look at relationships between school organisation policies and student performance was an improvement. Concerns still remain, however, with the manner in which results were reported, with negative correlations involving accountability policies – including the use of student achievement data to allocate resources – left unremarked on in tables in an annex. The decision to leave the use of student achievement data to evaluate teachers out of published correlations is inexcusable, considering that the relevant data had been collected. Secondary analysis showed this policy to again be associated with worse student performance across all participating countries, and within the OECD in the case of mathematics. These consequential accountability policies are often controversial where

implemented. The reporting by PISA of negative relationships with student performance for 2006 and 2009, as would have occurred had the cross-country correlation method been applied consistently and the results reported in a clear and uniform manner, would have consequences for the take up of these policies.

Inconsistent use of statistical methods in its 2006 analysis and selective reporting of its 2009 analysis, and the fact that over both surveys attention has been drawn to positive findings for school autonomy, competition and accountability measures while negative results for consequential accountability policies have been hidden or played down leave PISA and the OECD open to criticism that their work does not meet the highest standards of objectivity and favours certain schooling policies.

References

Adams, R.J. (2003) Response to 'Cautions on OECD's Recent Educational Survey (PISA)', *Oxford Review of Education*, 29, 377-379. http://dx.doi.org/10.1080/03054980307445

Amrein, A. & Berliner, D. (2002) High-stakes Testing, Uncertainty and Student Learning, *Education Policy Analysis Archives*, 10(18). http://epaa.asu.edu/epaa/v10n18/ (accessed 6 February 2012).

Apple, M. (2004) Creating Difference: neo-liberalism, neo-conservatism and the politics of education reform, *Educational Policy*, 18(12), 12-43. http://dx.doi.org/10.1177/0895904803260022

Ball, S. (1998) Big Policies/Small World: an introduction to international perspectives in education policy, *Comparative Education*, 34(2), 119-130. http://dx.doi.org/10.1080/03050069828225

Ball, S. (2006) Performativities and Fabrications in the Education Economy: towards the performative society, in H. Lauder, P. Brown, J. Dillabough & A. Halsey (Eds) *Education, Globalization and Social Change*, pp. 692-701. Oxford: Oxford University Press.

Benavot, A. & Tanner, E. (2007) The Growth of National Learning Assessments in the World, 1995-2006. Paper for UNESCO. http://campus.ph.fhnw.ch/pub/Main/UniBaselHS09/Benavot_Tanner_2008_growth_assessments.pdf (accessed 6 February 2012).

Bradley, S. & Taylor, J. (2002) The Effect of the Quasi-market on the Efficiency–Equity Trade-off in the Secondary School Sector, *Bulletin of Economic Research*, 54(3) 295-314. http://dx.doi.org/10.1111/1467-8586.00154

Carnoy, M. & McEwan, P. (2000) Does Privatization Improve Education? The Case of Chile's National Voucher Plan, *Educational Evaluation and Policy Analysis*, 22(3), 213-239.

Carnoy, M. & Rhoten, D. (2002) What Does Globalization Mean for Educational Change? A Comparative Approach, *Comparative Education Review*, 46(1), 1-9. http://dx.doi.org/10.1086/324053

Chubb, J. & Moe, T. (1990) *Politics, Markets and America's Schools*. Washington, DC: Brookings Institution.

Cook, T.D. (1974) The Potential Limitations of Secondary Evaluations, in M.W. Apple, M.J. Subkoviak & H.S. Lufler (Eds) *Educational Evaluation: analysis and responsibility*, pp. 155-235. Berkeley, CA: McCutchan.

Crouch, C. (2011) *The Strange Non-death of Neoliberalism*. Cambridge: Polity Press.

Dale, A., Arber, S. & Procter, M. (1988) *Doing Secondary Analysis*. London: Unwin Hyman.

Dohn, N.B. (2007) Knowledge and Skills for PISA – Assessing the Assessment, *Journal of Philosophy of Education*, 41(1), 1-16. http://dx.doi.org/10.1111/j.1467-9752.2007.00542.x

Forsey, M. (2007) *Challenging the System? A Dramatic Tale of Neoliberal Reform in an Australian High School*. Charlotte, NC: Information Age Publishing.

Friedman, M. (1962) *Capitalism and Freedom*. Chicago: University of Chicago Press.

Goldstein, H. (2004) International Comparisons of Student Attainment: some issues arising from the PISA stud, *Assessment in Education*, 11(3), 319-330. http://dx.doi.org/10.1080/0969594042000304618

Gorard, S. (2001) *Quantitative Methods in Educational Research*. London: Continuum.

Gorard, S., See, B.H., Smith, E. & White, P. (2006) *Teacher Supply: the key issues*. London: Continuum.

Gorard, S. & Smith, E. (2004) An International Comparison of Equity in Education Systems, *Comparative Education*, 40(1), 15-28. http://dx.doi.org/10.1080/0305006042000184863

Grant, N. (2009) Foreword, in D. Hill & R. Kumar (Eds) *Neoliberalism and Education and its Consequences*. New York: Routledge.

Green, A. (2006) Education, Globalization and the Nation State, in H. Lauder, P. Brown, J. Dillabough & A. Halsey (Eds) *Education, Globalization and Social Change*, pp. 192-197. Oxford: Oxford University Press.

Grek, S. (2009) Governing by Numbers: the PISA effect in Europe, *Journal of Education Policy*, 24(1), 23-27. http://dx.doi.org/10.1080/02680930802412669

Grubb, W. & Lazerson, M. (2006) The Globalization of Rhetoric and Practice: the education gospel and vocationalism, in H. Lauder, P. Brown, J. Dillabough & A. Halsey (Eds) *Education, Globalization and Social Change*, pp. 295-307. Oxford: Oxford University Press.

Hakim, C. (1982) *Secondary Analysis in Social Research: a guide to data sources and methods with examples*. London: Allen & Unwin.

Hamilton, L., Stecher, B., Vernez, G. & Zimmer, R. (2007) Passing or Failing: a midterm report card for 'No Child Left Behind', *Rand Review*, Fall, 16-25.

Hanushek, E. & Wößmann, L. (2008) *Education Quality and Economic Growth*. Washington, DC: World Bank.

Hargreaves, A. (2003) *Teaching in the Knowledge Society: education in the age of insecurity*. New York: Teachers College.

Hargreaves, A. (2006) Four Ages of Professionalism, in H. Lauder, P. Brown, J. Dillabough, & A. Halsey (Eds) *Education, Globalization and Social Change*, pp. 673-691. Oxford: Oxford University Press.

Harvey, D. (2005) *A Brief History of Neoliberalism*. Oxford: Oxford University Press.

Heyneman, S. (2003) The History and Problems of Education Policy Making at the World Bank 1960-2000, *International Journal of Educational Development*, 23(3), 315-337. http://dx.doi.org/10.1016/S0738-0593(02)00053-6

Hopmann, S.T. & Brinek, G. (2007) Introduction, in S.T. Hopmann, G. Brinek & M. Retzl (Eds) *PISA According to PISA: does PISA keep what it promises?*, pp. 9-19. Vienna: Lit-Verlag.

Hursh, D. (2007) Assessing No Child Left Behind and the Rise of Neoliberal Education Policies, *American Educational Research Journal*, 44(3), 493-518. http://dx.doi.org/10.3102/0002831207306764

Jacob, B. (2005) Accountability, Incentives and Behavior: the impact of high-stakes testing in the Chicago public schools, *Journal of Public Economics*, 89, 761-796. http://dx.doi.org/10.1016/j.jpubeco.2004.08.004

Johnson, B. (2001) Toward a New Classification of Non-experimental Quantitative Research, *Educational Researcher*, 30(2), 3-13. http://dx.doi.org/10.3102/0013189X030002003

Kamens, D. & McNeely, C. (2010) Globalization and the Growth of International Educational Testing and National Assessment, *Comparative Education Review*, 54(1), 5-25. http://dx.doi.org/10.1086/648471

Kumar, R. & Hill, D. (2009) Introduction: neoliberal capitalism and education, in D. Hill & R. Kumar (Eds) *Neoliberalism and Education and its Consequences*, pp. 1-11. New York: Routledge.

Lauer, P.A. (2006) *An Education Research Primer*. San Francisco: Jossey-Bass.

Lingard, B. & Grek, S. (2007) The OECD, Indicators and PISA: an exploration of theoretical perspectives. Paper for the Economic and Social Research Council/European Social Fund. http://www.ces.ed.ac.uk/PDF%20Files/FabQ_WP2.pdf (accessed 12 January 2012).

Marsh, J., Springer, M., McCaffrey, D., Yuan, K., Epstein, S., Koppich, J., Kalra N., DiMartino, M. & Peng, A. (2011) *A Big Apple for Educators: New York City's experiment with schoolwide performance bonuses*. Santa Monica, CA: Rand.

Martens, K. (2007) How to Become an Influential Actor – the 'Comparative Turn' in OECD Education Policy, in K. Martens, A. Rusconi & K. Leuze (Eds) *New Arenas of Education Governance: the impact of international organizations and markets on educational policy making*, pp. 40-56. Basingstoke: Palgrave Macmillan.

Moutsios, S. (2009) International Organisations and Transnational Education Policy, *Compare*, 39(4), 467-478. http://dx.doi.org/10.1080/03057920802156500

Organisation for Economic Co-operation and Development (OECD) (1992) *Education at a Glance: OECD indicators*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD) (2006) *Assessing Scientific, Reading and Mathematical Literacy: a framework for PISA 2006*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD) (2007a) *Executive Summary PISA 2006: science competencies for tomorrow's world*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD) (2007b) *PISA 2006: science competencies for tomorrow's world. Vol. 1: Analysis*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD) (2007c) *PISA 2006. Vol. 2: Data*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD) (2009) *PISA 2006 Technical Report*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD) (2010a) *PISA 2009 Results: vol. I. What Students Know and Can Do: student performance in reading, mathematics and science*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD) (2010b) *PISA 2009 Results: vol. II. Overcoming Social Background: equity in learning opportunities and outcomes*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD) (2010c) *PISA 2009 Results: vol. IV. What Makes a School Successful? Resources, Policies and Practice*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD) (2011) *PISA 2009 Technical Report (preliminary version)*. http://www.oecd.org (accessed 3 January 2012).

Papadopoulos, G. (1994) *Education 1960-1990: the OECD perspective*. Paris: Organisation for Economic Cooperation and Development.

Porter, T. & Webb, M. (2007) The Role of the OECD in the Orchestration of Global Knowledge Networks. Paper for Canadian Political Science Association. http://www.cpsa-acsp.ca/papers-2007/Porter-Webb.pdf (accessed 20 December 2012).

Prais, S.J. (2003) Cautions on OECD's Recent Educational Survey (PISA), *Oxford Review of Education*, 29(2), 139-163. http://dx.doi.org/10.1080/0305498032000080657

Prais, S.J. (2004) Cautions on OECD's Recent Educational Survey (PISA): rejoinder to OECD's response, *Oxford Review of Education*, 30(4), 569-573. http://dx.doi.org/10.1080/0305498042000303017

Ravitch, D. (2010) *The Death and Life of the Great American School System: how testing and choice are undermining education*. New York: Basic Books.

Resnik, J. (2006) International Organisations, the 'Education-Economic Growth' Black Box and the Development of World Education Culture, *Comparative Education Review*, 50(2), 173-196. http://dx.doi.org/10.1086/500692

Rinne, R., Kivirauma, J. & Simola, H. (2002) Shoots of Revisionist Education Policy or Just Slow Readjustment: the Finnish case of educational reconstruction, *Journal of Education Policy*, 17(6), 643-658. http://dx.doi.org/10.1080/0268093022000032292

Rizvi, F. & Lingard, B. (2006) Globalization and the Changing Nature of the OECD's Educational Work, in H. Lauder, P. Brown, J. Dillabough & A. Halsey (Eds) *Education, Globalization and Social Change*, pp. 247-260. Oxford: Oxford University Press.

Robertson, S. (2008) Remaking the World: neoliberalism and the transformation of education and teachers' labour, in M. Compton & L. Weiner (Eds) *The Global Assault on Teaching, Teachers and their Unions*, pp. 11-30. Basingstoke: Palgrave Macmillan.

Robertson, S., Novelli, M., Dale, R., Tikly, L., Dachi, H. & Alphonce, N. (2007) *Globalisation, Education and Development: ideas, actors and dynamics*. London: United Kingdom Department for International Development.

Sahlberg, P. (2004) Teaching and Globalization, *Managing Global Transitions*, 2(1), 65-83.

Sahlberg, P. (2007) Education Policies for Raising Student Learning: the Finnish approach, *Journal of Education Policy*, 22(2), 147-171. http://dx.doi.org/10.1080/02680930601158919

Sandström, M. & Bergström, F. (2002) School Vouchers in Practice: competition will not hurt you! Paper for the Research Institute of Industrial Economics. http://www.ifn.se/eng/publications/wp/2000-2005/2002_1/578_1 (accessed 6 February 2012).

Smith, E. (2007) *Analysing Underachievement in Schools*. London: Continuum.

Steiner-Khamsi, G. (2004) Blazing a Trail for Policy Theory and Practice, in G. Steiner Khamsi (Ed) *The Global Politics of Educational Borrowing and Lending*, pp. 201-220. New York: Teachers College.

Thrupp, M. & Hursh, D. (2006) The Limits of Managerialist School Reform: the case of target-setting in England and the USA, in H. Lauder, P. Brown, J. Dillabough & A. Halsey (Eds) *Education, Globalization and Social Change*, pp. 642-653. Oxford: Oxford University Press.

Turner, R. (2008) *Neo-Liberal Ideology: history, concepts and policies*. Edinburgh: Edinburgh University Press.

Woodhall, M. (1997) Human Capital Concepts, in A. Halsey, H. Lauder, P. Brown & A. Wells (Eds) *Education, Culture, Economy and Society*, pp. 217-223. Oxford: Oxford University Press.

Wuttke, J. (2007) Uncertainties and Bias in PISA, in S.T. Hopmann, G. Brinek & M. Retzl (Eds) *PISA According to PISA: does PISA keep what it promises?*, pp. 241-264. Vienna: Lit-Verlag.

DAN MURPHY is a journalist based in Melbourne, Australia who writes on education and industrial relations. This article is based on research he carried out while studying in the Master's Programme in Education and Globalisation at the University of Oulu, Finland. Currently he is seeking support for an expanded research project involving detailed secondary analysis of PISA results and analysis from 2006 to 2012. *Correspondence*: danmurf71@gmail.com