Chapter 5. Extending Educational Effectiveness: A critical review of research approaches in International Effectiveness Research, and proposals to improve them

David Reynolds
No affiliation, David@DavidReynoldsConsulting.com, +44 7802 309 052

Anthony Kelly
University of Southampton, A.Kelly@Soton.ac.uk, +44 2380 593 351

Alma Harris
University of Swansea, Alma.Harris@Swansea.ac.uk, +44 1792 205 678 ext 3261

Michelle Jones
University of Swansea, Michelle.S.Jones@Swansea.ac.uk, +44 1792 205 678 ext 3261

Donnie Adams
University of Malaya, DonnieAdams@Um.edu.my, +60 379 675 116

Zhenzhen Miao
Jiangxi Normal University, Z.Miao@Jxnu.edu.cn, +86 15895 843 705

Christian Bokhove
University of Southampton, C.Bokhove@Soton.ac.uk, +44 2380 592 415

5.0 Abstract

A review is given of internationally orientated educational effectiveness research over time, looking at the waves of studies from 1960 to 2001, the beginning of the PISA studies. It is concluded that whilst PISA has been a major methodological improvement upon earlier work, there are issues to do with the 'culture fairness' of test items, OECD misrepresenting of its own work, sampling adequacy and manipulation of samples by some governments.

It is argued that educational effectiveness and improvement research can provide useful perspectives for international effectiveness research in terms of its focus upon pedagogy/teaching, its value-added approach, its often longitudinal design, its use of 'supply side' as well as 'demand side' policy variables and its use of 'efficiency' as well as 'effectiveness' measures. It is also argued that the cultural and contextual factors/features of societies would additionally require study.

It is finally argued that high quality international comparative work may help in the generation of more useful educational effectiveness research, particularly in the generation of more valid, sensitive context specific formulations.

5.1 Introduction: The Rise of International Effectiveness Research

In the last twenty years considerable attention has been focussed upon the variation between countries in educational achievement and the apparent relative effectiveness of their educational systems. Partly this is a reflection of a world that is, in many respects, becoming 'smaller' all the time. The spread of mass communications and new information technologies is affording all countries a more international 'reach' in their world views. The revolution afforded by the pervasive spread of information means that ideas now travel 'virally' around the world with great rapidity, making it increasingly possible for educational ideas and processes to move freely too. In education, the process of 'internationalisation' has taken longer to embed than in areas such as the economy, but education now seems an international commodity too, and this Chapter analyses the problems and potential of international studies of educational effectiveness, one of the areas of that commodity.

In the 1990s there were interesting explorations of individual educational policy areas from different societies that achieved some attention, such as the possible relevance to the United Kingdom of Swiss methods of mathematics teaching (Burghes & Blum, 1995), the popularity and usefulness of Japanese 'lesson study' methods of professional development for teachers (Stevenson & Stigler, 1992) and the espousal of 'whole class interactive' teaching methods in mathematics for use in England based upon observations and apparent success in Taiwan (Reynolds & Farrell, 1996; Reynolds et al, 2002).

There had also been attempts historically within certain societies to model their entire educational systems wholesale upon what were seen as successful practices elsewhere, as with the modelling of the Hungarian system of pre-school education upon those of Japan, the modelling on the English Literacy and Numeracy Strategies of the 2000s in some parts of the United States (Barber, 2007), and the modelling of Welsh language learning upon the methods pioneered in Israel in the 1960s (Reynolds, 2008).

There were however, many factors that hindered the internationalisation of education until relatively recently. Comparative education as a source discipline had eroded internationally in quality, influence and significance from the 1990s onwards, particularly in the United Kingdom where the number of University Departments in the area dwindled. Looking specifically at the field of educational effectiveness research (EER) that was probably the most rapidly developing of all educational specialties in and after the 1990s, it is true there was an international organisation (the International Congress for School Effectiveness and Improvement; ICSEI) and increasingly prevalent cross cultural analyses of the factors shown in the *national* research studies in different countries that were associated with effectiveness. However, this was not paralleled by a corresponding growth in the number of *internationally* based research studies - which involved common conceptualisation, operationalisation and measurement. Instead, what existed, simply, were international reviews of national studies in different countries (Reynolds et al, 2014).

In the last two decades this situation has been rapidly changing, largely because of the funding and sponsorship of large-scale international achievement studies by organisations such as the Organisation for Economic Cooperation and Development (OECD), particularly the Programme for International Student Achievement (PISA), but also shown in the World Bank's and IEA's commissioned literature reviews of educational effectiveness and those emanating from the European Economic Community too. Private sector companies, such as Pearson, have also been instrumental in generating influential publications (Barber, 2007; Mourshed, Chijioke & Barber, 2010) that have become a reference point for much of the global debate about international 'best performance'. Interestingly, the historic concern within the Comparative Education community to understand national educational cultures has been replaced by the assumption that 'what works' are organisational arrangements largely independent of cultures.

The global economic and financial crisis and retrenchment of the late 2000s undoubtedly focussed renewed attention on issues of comparative educational performance. In times of greater financial scarcity, the ways in which societies can maximise their economic productivity by utilising their educational systems to generate more 'human capital' assumes greater importance, and the possible blueprints that may exist from more 'effective' educational systems took on greater salience than in times of rapid economic growth like the early 2000s. We now proceed to look at some of the studies that were conducted in the 1960's to 1990's, before looking in some detail at the well-known PISA studies that began later from 2001.

5.2 The First International Studies, 1960-2000

The international studies conducted during these four decades were numerous, as were commentaries, critiques and reviews. The International Association for the Evaluation of Educational Progress (IAEP) published some of these but the greatest number came out of the work of the International Association for the Evaluation of Achievement (IEA).

Briefly, the IEA conducted the First and Second Science Studies (Comber & Keeves, 1973; Rosier & Keeves, 1991; Postlethwaite & Wiley, 1992; Keeves, 1992); the First and Second Mathematics Studies (Husen, 1967; Travers & Westbury, 1989); the Study of Written Composition (Purves, 1992) and the Classroom Environment Study (Anderson et al, 1989). The IAEP conducted two studies of science and mathematics achievement cross-culturally (Keys & Foxman, 1989; Lapointe et al, 1989). The IEA also published the findings of the First Literacy Study (Elley, 1992; Postlethwaite & Ross, 1992). Reviews of the great majority of these studies are in Reynolds & Farrell, (1996), and Reynolds et al. (2002).

The results of the studies differed to an extent, according to the precise outcome measures being utilised whether it was reading/literacy, numeracy or science achievement. What was common though was a 'country against country' league table ranking perspective. However, it remained unclear which cultural, social, economic and educational factors were implicated in the country differences.

There were of course further 'core' problems that placed severe limitations upon the capacity of these international studies to generate valid knowledge about educational effectiveness. Some problems are present in all cross-national research, such as those of accurate translation of test material, of ensuring reliability in the 'meaning' of factors (such as social class or status indicators for example), of problems caused by Southern Hemisphere countries having their school years begin in January and of problems caused because of retrospective, potentially out of date material information being used. In certain curriculum areas, the cross-national validity of the tests utilised gives cause for grave concern and, as an example, the IEA study of written composition failed in its attempt to compare the performances of groups of students in different national systems that used different languages, since the latter study concluded that 'The construct that we call written composition must be seen in in a cultural context and not considered a general cognitive capacity or activity' (Purves, 1992, p. 199).

Even the simple administration of an achievement test in the varying cultural contexts of different countries may pose problems, particularly in countries where the 'test mode' in which closed questions are asked in an examination-style format under time pressure may not approximate to students' general experience of school. By contrast, the use of this assessment mode within societies such as those of the Pacific Rim, where these methods are very frequently experienced, may facilitate country scores.

In addition to these core problems that affect all large-scale international effectiveness research, there were specific problems concerning the IEA and IAEP international effectiveness studies that represented virtually the totality of this international effectiveness research enterprise until the late 1990s.

5.2.1 Methodological Deficiencies

- The basic design of these studies, which were concerned with explaining country against country variation, may have itself been responsible for problems. Generally, a small number of schools each possessing a large number of students were selected, which made it difficult to make valid comparisons between schools once factors such as school type, socio-economic status of students and catchment areas were taken into account.

- Curriculum subjects were studied separately, making an integrated picture of schools and education in different countries difficult.

- There was considerable difficulty in designing tests which sampled the curricula in all countries acceptably, although the Trends in International Mathematics and Science Study (TIMSS) project expended considerable energy to ensure a geographical reach of items, and also published its results 'unadjusted' and 'adjusted' for the curriculum coverage or 'opportunity to learn' of individual countries.

- Cross sectional rather than longitudinal surveys precluded understandings of student progress, instead limiting findings to snapshots of student performance at specified ages.

5.2.2 Sampling Issues

- There were instances of large variations in the response rates that made interpretation of scores difficult in cross country comparisons.

- Sometimes samples of students used were not representative of the country as a whole (e.g. one area of Italy was used as a surrogate for the whole country in one of the IAEP studies).

- Variations between countries in the proportion of their children who could have taken part in the studies made assessment of country differences difficult. Mislevy (1995), notes that, whilst 98 per cent of American children were in the sampling frame and eligible to take part in one study, the restriction of an Israeli sample to Hebrew-speaking public schools generated only 71 per cent of total Israeli children being in eligible schools.

5.2.3 Limited Data and Limited Analyses

- In many studies there was a lack of information upon the non-school areas of children's lives (family and home environment) that might have explained achievement scores. Surrogates for social class utilised, such as 'number of books in the home', were not adequate.

- Outcomes data was collected mostly on the academic outcomes of schooling, yet social outcomes may have been equally interesting and important.

- The factors used to describe schools were overly resource-based (because of the greater perceived chance of obtaining reliability between observers in the former factors across countries, no doubt), in spite of the clearly limited explanatory power of the latter variables. At classroom level, only some studies (including TIMSS and Progress in International Reading Literacy Study; PIRLS) have used any measures of teaching and learning processes, with the use of videotapes of classrooms by the TIMSS project

(Hiebert et al, 2003; Stigler et al, 1999) being particularly interesting, although of course rare.

- Only limited attempts were made to analyse the international samples differentially, for instance by achievement or by social class, with the exception of a limited amount of analysis by gender.

From all these points above, it is clear that the international studies of educational effectiveness of the IAEP and the IEA from the 1960s to the late 1990s necessitate some caution in interpretation. Not all studies possessed the same design, analysis and methodological problems, and no studies possessed all the design, analysis and methodological problems in total. But enough studies possessed sufficient problems to make firm and generalisable conclusions difficult and problematic.

The attention given to these international achievement surveys was by the 1990s considerable, although most accounts do not grant them the same importance globally as the more recent PISA studies (e.g. Waldow, Takayama & Sung, 2014). There were, of course, also a limited number of critiques of the individual studies, and of the paradigm within which they were constructed (Alexander, 2000, 2012; Reynolds et al, 1994). Additionally, much of the discussion of the findings of these studies and their merits and de-merits remained within the academic community, rather than spreading widely into the news media as have the results of PISA. The globalisation phenomenon, facilitated by the spread of information technology (IT), had not yet fully emerged in the 1990s to spread interest in what certain countries were doing educationally across all countries. And the financial crisis, as well as its effects in multiplying the pressures upon countries and politicians to pay enhanced attention to their economic and educational systems, did not influence policy until the late 2000's, perhaps with the exception of the global interest in the 1999 TIMSS video study.

5.3 The PISA International Achievement Studies, 2001 Onwards

PISA has spawned an extensive literature, including commentaries that emanate from a politically and socially critical perspective (e.g. Bulle, 2011; Dobbins & Martens, 2012; Eivers, 2010; Fischbach et al., 2015; Gaber et al., 2012; Grek, 2009; Hanberger, 2014; Kankaras & Moors, 2013; Lewis, 2014; Morgan and Shahjahan, 2014; Ozga, 2012; Sellar, & Lingard, 2013a;2013b; Waldow, Takyama & Sun, 2014).

The main points made within the group of studies cited above are that:

- PISA studies have encouraged a convergence of policy intention and borrowing in educational policies and practices across the world;

- PISA has encouraged a positioning of international organisations – e.g. OECD, World Bank – in international educational matters that could be viewed as problematic because these organisations possess their own agendas;

- PISA has encouraged a 'one size fits all' response by policymakers that does not attach enough salience to the importance of local and national cultures;

- PISA is part of the phenomenon of the globalisation of the world's social and political structures, which has the potential to limit national, country-level influences.

In addition to these important perspectives, severe doubts have been expressed about PISA in terms of its methodological adequacy.

The core purpose of PISA has been stated by the OECD (2009) as follows:

> 'Are students well prepared for future challenges? Can they analyse, reason and communicate efficiently? Do they have the capacity to continue learning throughout life? The OECD Programme for International Student Achievement answers these questions and more, through its surveys of 15 year olds in the principal industrialised nations'. (p.1).

In practice, PISA seeks to achieve this through IT-based and 'paper and pencil' tests that are given to students aged 15 in different countries, in three achievement areas: mathematics, reading and science, with an emphasis not upon measuring student factual 'knowledge', but more upon measuring the capacity of students to apply that knowledge in real world situations, the so called 'skills based' approach. Using Item Response Theory (IRT) and based upon the assumption that a latent 'trait' determines all responses to test items, comparisons of students who have taken different test items can be made based upon the statistical equivalence of the items, in terms of the data they produce. Additionally, surveys are conducted on students, Headteachers/Principals and parents, focussing on their attitudes and particularly on student and parental backgrounds/perceptions. The PISA surveys were published in 2001, 2004, 2007, 2010, 2013 and 2106. An additional outcome – problem solving – has been added to the existing three skill areas over time, and there has been a movement towards the increased measurement of 'metacognitive' skills shown first in the 2013 PISA use of specific items in this area in the reading test, and in the increased proportion of metacognitive items across the skill areas in the 2015 testing. Something called 'Global Competencies' is be measured from 2018 testing.

Unquestionably, PISA represents a major improvement upon the methodology and analyses that were characteristic of the previous group of studies from the late 1960s to the late 1990s. Other studies, like TIMSS and PIRLS, have also shown methodological advances over time. By focussing on the 'skills' to apply knowledge in real world settings rather than on

the knowledge itself, PISA has theoretically made it easier to generate cross-cultural assessments, since there is likely to be smaller variation in the 'skills' aimed at in different societies than in the actual 'knowledge bases' that are taught to generate those skills, though this argument itself has been the focus of some criticism as the distinction between knowledge and skills is seen as artificial by some.

Also, PISA has paid significant attention to the rigour of the sampling process, especially to ensuring uniformly high response rates across countries, although securing such high response rates has frequently required the use of replacement schools, and concerns have been raised about PISA's method of calculating its response rate (Murphy, 2010). Attempts have also been made to ensure that various countries – like China – enter more representative portions of their national populations into the sampling frame, with the exclusion from the main sample of any countries having unsuitably low response rates (as in the case of England in the 2004 PISA survey). In addition, criticism has been levelled about gender imbalances in some country samples and the varying levels of exclusions due to intellectual impairment or special educational needs (Wuttke, 2007).

PISA has also been particularly committed to its attempts to measure the effects of the 'macro' level of national level policies related to the governance, funding, organisation and management of education within different jurisdictions, a marked contrast to the earlier IEA studies that collected very little data in these areas.

PISA has also opened up – in the publicity it has had – an important debate about educational excellence and equity that has been generated by participation of many different constituencies and groups around the globe.

Finally, the historic drift of the design of the PISA studies represents a sensible reflection of how international emphases and knowledge in areas associated with educational effectiveness have been changing over time. More socio-economic background data on students is being collected in the more recent PISA studies, reflecting enhanced interest in the interactions between schools *and* their communities, although there remain concerns that data on structural inequality in different countries is still not fully reflected in the PISA data collection process and in its analysis (Eivers, 2010). The data that have been collected on the levels 'above' that of the school, such as the 'meso' level of the District/Local Authority and the 'macro' level of national educational policies on such areas as accountability, assessment, governance and funding, reflect the increasing attention that is being given to educational policy issues within educational research. The movement towards the measurement of metacognitive skills is also sensible, given their increased salience in the recent international assessment literature (Muijs et al, 2014).

In many important respects, however, PISA may not have succeeded completely in addressing all the concerns of the critics of the earlier studies in important areas of research design, research methodology and data analysis. There are also additional specific issues to do with PISA's specific use of its data and its data reporting.

A particular focus recently has been upon five 'core' issues where PISA appears vulnerable to criticism:

- It is crucial that tests used in work like PISA should be 'culturally fair', with students of the same level of achievement receiving the same assessment test scores independently of the country they come from. Early suggestions from small scale studies were that measurement equivalence may have been limited (Allerup, 2007; Goldstein, 2008; Grisay & Monsuer, 2007; Yildirim, 2006). However, the recent study of Kankaras & Moors (2013), based upon 2009 OECD data from 64 countries and 475,460 15-year-olds, found that "equivalence occurred in a majority of test questions in all three scales researched and, is on average, of moderate size" (p.1).

Given the fact that the PISA tests are translated and administered in a large number of countries and in cultures that are diverse linguistically, socially and economically, it could be expected that some differences in interpretation and understanding might occur. However, the extent and degree of the inequivalence across datasets, as stated by Gorur and Wu (2014), suggests that it actually "impairs the validity of country comparisons" (p. 17).

- There is evidence that the OECD's written reports on their data may not be totally in accordance with the findings of their own data - in other words, that there may be mis-reporting or instances of cherry-picking some findings over others. Whilst considerable attention has been historically drawn in PISA publications to positive 'key findings' for 'demand side' national policies involving school autonomy, competition and accountability, secondary analysis of the 2007 PISA data showed that other consequential and related accountability policies such as using student achievement data to evaluate teachers and to allocate resources were associated with *worse* student performance (Murphy, 2014). PISA applied more consistent country correlation methods to the 2010 study than in previous years, but negative correlations involving achievement and some accountability policies – for example, the use of student achievement data to allocate resources -- were left unremarked-upon in tables in an annex to the relevant report. Secondary analysis showed this policy to again be associated with *worse* student performance on all measures of performance across all countries in the study (Murphy, 2014), and on mathematics within the OECD countries.

- There have also been concerns about sampling issues, such as variation in the participation of special schools in some societies and not others, (Bracey, 2004; 2009), and the possible effects of differential country response rates being determined by Principal/Headteacher enthusiasm for the PISA testing process (Meyer & Schiller, 2013).

Some other design and methodological limitations related to cultural differences include the extent to which students from different countries take 'non-PISA' tests and how far they take 'tests' seriously. Sjoberg (2012) argues that students in different countries and cultures understand tests differently and place different emphases upon their importance, such differences being a by-product of deep historical cultural influences and differences. He also argues that the desire to do well in 'the test' and persisting to complete all of it is uneven across cultures and could be a variable that explains differences in outcomes and performance.

- Most importantly, it still seems possible for governments to 'game' PISA, as with the early use of 'Shanghai' as a surrogate for all China when in fact it is a highly atypical region, both educationally and socio- economically. In Shanghai, it has been argued, that many migrant children have not been systematically involved with PISA testing, and that over half of all fifteen year olds may have been excluded from testing because of the effect of the 'hukou' identity card or passport system (Loveless, 2013; 2014). The 'hukou' controls access to Municipal services and those Chinese with rural 'hukou' may be discriminated against in the urban environment, for example by being banned from public schools and put into migrant schools. Also, there are a proportion of migrant children who are left behind in villages as their parents migrate to Shanghai, and a further proportion of migrant children leave school before the age of fifteen when children are meant to take the PISA tests. Both these factors may generate an unrepresentative sample of children in the Shanghai PISA sample. A larger sample of Provinces has been used in 2015 and 2018 testing, however.

- There are also methodological issues of concern, some of which relate directly to PISA's overarching goals. The OECD states that tests are designed to assess to what extent students at the end of compulsory education "can apply their knowledge to real-life situations and be equipped for full participation in society" (OECD, 2015). This is apparent in the nature of the content where 'real life' is demonstrated by the heavy use of contexts in, for example, maths and science assessment items. This might be problematic for two reasons. Firstly, contexts are often culturally rooted. Some contexts might be more appropriate for certain countries; for example, assessment items about bicycles and sailing ships might be less appropriate for some countries. Secondly, the context often adds additional reading demands to the assessment items. A valid question to ask, then, is whether test items might measure more than just maths and science constructs, which might weaken the validity of the measurements by creating additional variance irrelevant to the intended construct (Messick, 1998; Eivers, 2010).

The issue of reading load and difficulty has also been raised by Bodin (2007), who states that it is unclear whether difficulties in PISA's mathematics items are caused by the underlying text or the mathematical problem's degree of difficulty. A 'relationship' report from the IEA combined data from their maths and science test (TIMSS) and literacy test (PIRLS) and indicated that reading difficulty was associated with achievement (Mullis, Martin & Foy, 2013). Results varied from country to country and even between mathematics and science within countries, yet there was overall support for the idea that higher reading demands can make the fourth grade TIMSS items more challenging for weaker readers. It is possible that this point applies to PISA, too, and indeed Ruddock, Clausen-May, Purple & Ager (2006), noted that "the higher reading demand of questions in PISA is often accompanied by a relatively lower demand in the mathematics or science required" (p. 123).

Methodological challenges are not only apparent in PISA's assessment items, but also in several of the assumptions underpinning the statistical methods used. One criticism concerns the use of the Rasch model (Kreiner & Christensen, 2014). The OECD clearly explains its methods in the PISA technical manual, but the underlying assumption of it is that all the questions used in the study would have to function in exactly the same way (be equally difficult) in all participating countries. Given the cultural differences, but also differing curricula, this seems an unlikely assumption. If indeed the approach includes removing outlier items because of a 'unidimensional' requirement, this might create cultural bias.

5.4 PISA: A Perspective from Educational Effectiveness and Improvement Research (EEIR)

In addition to the range of recent methodological concerns and criticisms we have outlined above, our extensive historic bodies of knowledge, methodology and agreed approaches within EEIR suggest a further set of limitations that must be placed upon the utility, reliability and validity of the PISA studies. We outline these now, before concluding with a plea for, and some suggestions about, the generation of an enhanced focus upon differently conceptualised international effectiveness studies to be undertaken within the general field of EEIR.

5.4.1 The Absence of Teaching/Pedagogical Focus. All existing reviews of research conducted within EEIR have argued for the primacy of teacher effects and for these effects to be bigger than the 'levels' of the school and of the District/Local Authority. However, PISA collects no data upon the methods of teaching used in different countries, focussing upon the organisational arrangements of classrooms within schools more than on the actual behaviours of teachers and support personnel that have been shown to be highly important in the five decades of research on what can be called 'teacher effectiveness' (Brophy, 1976; Brophy & Good, 1986; Muijs & Reynolds, 2011).

Clearly, this absence of classroom data is because individual teachers would find it hard to rate and describe their own teaching in ways that could command cross cultural validity, and alternatively sending researchers to observe the classrooms of teachers in different countries would be hugely time intensive and expensive.

While these are sound reasons, it remains the case that the absence of any major focus upon pedagogy may well function both to severely limit the capacity of PISA to understand the causes and nature of country differences in educational effectiveness and also to imperil the prospects of success of any policies that may be tried out in different countries based upon PISA findings, given that it is likely to be 'teaching,' the 'alterable variable' with the largest likely effects, that countries may well wish to influence and improve. It may also be that the continued concern with the managerial arrangements of schools, Districts/Local Authorities and national policymaking/policies promoted by those who draw upon PISA to offer policy guidance, rather than with pedagogical practices, may not resonate with practitioners who are more interested in the classroom itself rather than the organisational layers that sit above it. However, the linkage between PISA and teacher behaviours being trialled in the Teaching and Learning International Survey (TALIS) suggests some progress in this area.

5.4.2 The Limited Use of a 'Value Added' Approach. EEIR has, over time, successfully established its bodies of knowledge about 'what worked' in generating its described 'effective schools' and 'effective teaching practices' (see reviews in Reynolds et al, 2014, for example), and generally enjoyed much practitioner and policymaker enthusiasm for the knowledge bases as they began to appear after the 1990s. Over those years, it became axiomatic that in order for knowledge to be reliable and accurate concerning the nature of the important school and classroom effectiveness factors, the 'raw' achievement results of schools should be made reflective of the variation in the nature of the intakes that went into the particular schools and classrooms, in order for better estimates of the 'value added' by particular educational settings to be available to correlate against their educational practices. This is what generated the effectiveness knowledge base. Using 'raw', non-value added measures of achievement that do not distinguish between the contribution of educational factors and the contribution of non-educational factors like socio economic status, parental attitudes, and cultural factors linked to differences between schools in their intake characteristics, risks invalid conclusions.

This is what PISA has done. In the earlier PISA studies there were limited measures of non-educational factors employed, so there was no systematic attempt to control out the major differences in social, cultural, environmental and economic factors in the wide range of countries utilised as the sample. Effectively, by default, the assumption was being made that it was educational factors solely that were involved in determining country differences.

More recently, by the time of the 2010 study, PISA was in fact also using 'national income per head of employed population' to provide a measure of the quality of what the educational systems of different countries were receiving as intakes, but only 6% of the differences in average student performance were due to GDP per capita (Reynolds et al, 2015), suggesting the need for other factors to be used. A wider range of socio-economic background and attitudinal data on students and parents were used in and after the 2013 studies, but the continued absence of any student achievement or student ability measure as controls means that even this wider range of social background factors were probably not functioning adequately to control out non-educational factors and influences. Although more socio-economic background data on students has been collected in recent years, the attempt has not been made to use these to generate 'value added' measures of relative country performance to supplement the much publicised and prevalent raw 'league tables'.

It is interesting that the PISA 2013 and 2015 studies do indeed show considerable recognition of the very important role played by non-educational factors in the determination of achievement outcomes, evidenced in the large number of analyses presented as to how all countries perform in the educational achievement of their lower socio-economic status groups of students. This attention given to 'equity' of performance within different countries, rather than merely 'excellence', is much to be welcomed and parallels the emphasis upon the differential effectiveness of schools shown within the EEIR community, particularly in research from Continental Europe in the last fifteen years (Reynolds et al. 2014).

But if the effect of home social background and other non-school factors upon children's educational prospects *within* all different PISA societies deserves attention, these should surely also deserve to be used as factors in the analysis of differences *between* societies, by making allowance for background effects upon country achievement scores. Educational effectiveness research made its rapid progress in understanding how schools had their effects only *after* it had adopted 'value added' perspectives, particularly those involving multilevel analyses (Goldstein, 2003) and reflecting other methodological advances in the field (Creemers, Kyriakides & Sammons, 2010). It is suggested that 'raw' achievement data is useful given that countries can see in *absolute* terms how they are doing with the development of their human capital, but that 'value added' or 'relative' data may be very useful too.

5.4.3 The Absence of a Longitudinal Research Design. PISA uses a cross sectional research design, whereas in contrast, EEIR now regards it as best practice and axiomatic to use longitudinal research designs. No doubt reasons of cost may be the explanation for use of the current PISA research design, given that it involves one testing point only. Also, to wait for differences in the gains of a cohort passing through schools in different countries to appear would certainly be half a year, and even then these 'gain' differences *over* a short

time would be nowhere near as large as differences between countries *at a point in time* that have evolved, making 'at a time point' the preferred option.

However, longitudinal studies following students for a year have been profitably used in exploratory work in this area (Reynolds et al, 2002; 2014), and a longitudinal research design would be helpful for PISA in two ways. Firstly, it would permit a more valid exploration of the effects of non-educational determinants of achievement – even if (as at present) only socio-economic background factors were measured rather than also using a prior achievement measure, the moderate correlations likely between all of these factors and achievement means that the levels of achievement at the first stage of the two stage testing involved in a cohort design ('pre' and 'post') would reflect the influence of these socio-economic factors. The gain over time in different countries that remained after 'stripping out' the start scores would be highly likely to reflect educational influences, thus generating more valid, 'true' educational effects.

Secondly, the study of the same children over a given time period would be likely to increase our understanding of the complex interactions between schools, educational systems and their children. Following the same children over time, with repeated visits made necessary by the need to do 'pre', and 'post' testing, does not necessarily improve our understanding of educational processes – as 'one off' events, how can they? But this approach does make more possible the collection of longitudinal data on student school experiences that are likely to give greater understanding of educational processes, and explain more variance. Interestingly, the effects of educational factors in the cohort studies following the same children over a longer period of time are typically much higher than those from cross-sectional work, or from those longitudinal studies undertaken over a short period of time (e.g. Guldemond & Bosker, 2009).

5.4.4 The Use of Educational Policy/Educational Process Factors of Limited Explanatory Power. It is important to note firstly that 'supply side' policies – concerned with teacher professional development, or national level programmes to build capacity for example -- are utilised in PISA much less than those related to the 'demand side'. All things being equal, 'demand side' effects are likely to feature more strongly than the 'supply side' in the explanations of the success/failure of educational policies, then.

More importantly, this effort put into the data collection in the 'demand side' 'macro' policy areas may not be particularly useful in explaining variance. For example, the United States 'No Child Left Behind' Act (NCLB) of 2001 requires States to have accountability systems which typically involve State-wide testing for all children in grades 3-8, the disaggregated reporting of data on student performance and the employment of sanctions when student performance is poor. Hanusheck & Raymond (2005), show an effect of only 0.2 of a standard deviation (using individual State data), on test scores. Dee & Jacob (2009;

2011), report a 0.5 student standard deviation impact of NCLB on student Maths scores, but no impact upon Reading scores. The Burgess, Wilson & Worth (2010) report on the effect of the national regime in Wales that abolished the publication and consequential use of the individual school national performance tables finds, after stripping out socio-economic factors by matching schools in 'experimental' Wales with 'control' England, that this is equivalent to a 0.23 of a (school level) standard deviation negative effect. All these studies suggest low effects for any of the 'demand side' policy levers.

The enthusiasm shown within PISA for collection of large amounts of data upon a limited range of 'demand side' policies may be understandable, given that the OECD wishes to influence the practices of policymakers, but it may be that some of the 'supply side' factors should interest them too to them rather than being consigned to the TALIS programme alone.

5.4.5 The Absence of an Efficiency Perspective. PISA has focussed primarily upon an 'effectiveness' perspective related to attempting to explain and understand national differences in the output of educational achievement. In doing this it has paralleled the EEIR field closely, which itself has focussed more on 'effectiveness' than 'efficiency'.

The 'efficiency' of countries, in the sense of the scale of the material 'inputs' that are necessary to generate the 'outputs' of the effectiveness levels shown in different societies, have so far received little attention in PISA data collection, with the exception of the expenditure levels of different societies being shown as unrelated to overall country variation in achievement test scores. Such a finding is not surprising, of course. About 80 to 90 per cent of the variation in 'per student expenditure' is due to variation in the pay of teachers, which in PISA is then expressed on a linear scale. But since the overall individual country level of national income is closely related to national individual country teacher pay, this finding only tells us that national income per head is not related to student achievement. It does not tell us that other 'efficiency' measures may be unimportant. Indeed, more recently a study by Dolton et al. (2015), did relate PISA scores to a range of financial inputs, finding two -- teacher salaries and class size -- to be significant, creating an 'efficiency index' that is essentially a measure of how highly a country's pupils score on PISA given how much (or little) a country spends on its teachers. This method provides a useful starting point for looking at the question of efficiency. However, it also further illustrates some of the limitations of PISA data (i.e. the difficulty of causal attribution from cross-sectional datasets) and the misuse of the data to which PISA has been prone, through (again) creating a league table notwithstanding the measurement error and the overlapping confidence intervals involved, and through simplistic policy advice. In general, of course, efficiency is important, and therefore both the use of production functions taken from economics to study efficiency, and the use of cost-benefit analysis when looking at

particular interventions or policy changes, would be beneficial to the study of education as a whole.

One 'efficiency' measure that was also strangely neglected in PISA is 'time', in terms of the 'inputs' of time that students in different countries are exposed to in their instructional activities. Time is an international measure that has a common metric and means the same in all countries. An hour of time is exactly the *same* in Oman as it is in Shanghai (whereas an hour of teacher's *pay* is highly variable depending on the country setting). Exposure to this thing called 'time' is highly variable cross culturally – there is a range in the days of schooling per student per year in different societies of from 230+ at the top to perhaps 160 days at the bottom, with some Pacific Rim societies scoring particularly highly on this. Is it possible that this efficiency measure may be related to country scores?

If one were to add an 'hours in school' to a 'days in school' measure, then country differences may widen. In addition to 'time in school', there also is some scale and impact of private supplementary tutoring, often referred to as 'shadow education'. Shadow education (Bray, 2014) has expanded significantly worldwide and is now recognised to potentially have considerable significance. Thus measures of the duration and intensity of schooling and also additional learning outside school via tutors/private tutoring schools need to be recognised as of potential influence.

Whether one stays with financial factors as 'inputs', or adopts additional non-financial ones like time, an 'efficiency' perspective may be useful for PISA, with its policymaking evidence, to employ.

5.4.6 The Absence of National Cultures and Context in the Analysis of Effectiveness. The central assumption underlying PISA is that national educational structures and policies – and by default only these things -- explain global variation in students' academic performance (Feniger and Lefstein, 2014). In other words, there is an underlying assumption that cultural factors or features play little or no part in explaining differences in relative country performance.

Explanations of success in comparative work – as in any educational research work -- are dependent on a number of inter-related factors, but studies such as PISA may produce analyses that are not sensitive to the *relationality* of the phenomena being studied. Consequently, many of the factors that might affect educational performance, particularly those that are culturally defined or contextually shaped, are not included or captured in existing PISA analyses. This leads to what Gorur and Wu (2014) have nicely termed the 'problem of the unmeasured' and the fact that drawing any meaningful parallels or conclusions about cross-national performance from the existing PISA studies will be difficult, if not impossible. Pereyra et al (2011, p.261) aptly argue that 'PISA is a brilliant big-social

science mapping of outcomes but in no anthropological, historical or cultural sense is it comparative work'.

There are a number of specific issues that result from the absence of cultural and social contexts as factors within PISA:

- Unmeasured cultural/social factors may have effects on country achievements that are at the moment explained by default as due to educational factors. Examples of this, from research outside PISA, involve the stress on ambition related to the simultaneous 'internal/external' loci of control on Asian children (Reynolds et al, 2002), the value given to education through positive parental perceptions of literacy as a goal in Finland (Sahlberg, 2011), or the historical enthusiasm -- verging on idolatry -- for education in the Welsh society of the mid-20th century, formed by religion, socialist policies and the effects of a self-educated working class (Reynolds, 2008).

- Without an understanding of context we cannot know 'how' any possible educational policy factors have their effects, and causal inferences and attributions cannot be made. At present, romanticised accounts of how policies have their apparently dramatic effects dominate discussions (e.g. Mourshed, Chijhoke & Barber, 2010), as Zhao (2014) notes, and there is no explanation of the possibility that sets of educational policies and educational factors are *differentially* effective across societies in accordance with country cultures and social structures, in other words that there may be an interaction between societies and their educational systems.

- PISA reflects the view that the *same* educational/school/policy factors are effective everywhere independent of context, but one pilot study (Reynolds et al., 2002) found that the same effectiveness factors only 'travelled' across their sample of eight countries at the level of the classroom, where the effects of detailed factors such as structured teaching, high expectations and the other factors from the teacher effectiveness literature all appeared important in all the different societies studied. By contrast, details and operationalised characteristics of the school level factors and educational policy factors associated with effectiveness were very *differen*t in different societies, raising the possibility that high levels of country achievement may be generated by factors that are *different* at the higher levels of educational systems. Do different countries need different systems – in accordance with their cultures – to generate teaching and classrooms which, because of the nature of children's physiology internationally, should be the same if they are to be effective? We do not know.

The analyses of PISA pay little attention to cultural and contextual factors that may partly explain the differences in the performance of education systems (Harris & Jones, 2015). As Feniger and Lefstein (2014) underline, "we need a much better understanding of comparative cultural contexts" to explain relative educational performance and outcomes.

Without such understanding, the educational policies, strategies and interventions associated with, and indeed endorsed by PISA and the OECD, will continue to be founded on scientific sand.

5.5 Conclusions: The Potential Value of Improved International Effectiveness Research

We have seen so far that international surveys of educational achievement have been receiving more and more attention in the last decade. One early analysis (Dominguez, Vieira & Vidal, 2012) found 322 publications from 2002-2010 on PISA in the three main scientific databases for the social sciences: ERIC, EBSCOhost and the ISI Web of Science. This number would miss those that have appeared since, and the considerable number of articles in the 'grey' area of magazines, web sites, think tank publications and the like. Additionally, the volume of material about PISA in newspapers and in other media is extensive. Studies by the IEA, that have continued to be published, also receive extensive coverage. A more recent review moves discussion into a more considered and comprehensive direction (Hopfenbeck et al., 2016).

What *has* been pursued however, as comparative work by EEIR, are mostly studies that are done in multiple countries, collecting data on a defined issue, or problem, or factor, which *purport* to be cross cultural or comparative work but which lack across nation methodological communality in conceptualisation, operationalisation and measurement (that PISA does show considerably, of course). We have in EEIR also just lots of national reviews, not a comparative knowledge base.

It would be wrong to omit to mention here the efforts of those who have attempted to do cross cultural work from a 'school improvement' perspective, particularly those concerned to generate strategies that may be useful for system reform internationally (e.g. Hopkins, 2007; Barber, 2009; Hargreaves & Shirley, 2009; Fullan, 2009; Harris and Chrispeels, 2007). A number of interesting country system-level studies have been published, such as those of Whelan (2009). Further, improvement researchers have generated some analysis of context specificity whereby 'what works' for improvement is seen as variable at different growth states for schools (Hopkins, Harris & Jackson, 1997; Hopkins, 2013; Hopkins & Reynolds, 2001; Harris & Jones, 2015).

Much of this work is also sensitive to local country context, and attempts to avoid a 'list of ingredients' approach, preferring a 'recipe' based upon a different 'mix' of approaches to system level reform in different countries (Hopkins et al, 2014). However, it is sometimes unclear in many school improvement studies what data exist to support their formulations, how they have been analysed, how intensive has been the immersion of the individual 'improvers' in their different societies and whether the understandings of the highly complex interactions between systems, countries and cultures are quite advanced enough.

Nowadays, therefore, the pressures on us in the EEIR field to create better research and better explanations in the field of cross cultural/comparative work are immense, given PISA's partial contribution as outlined above. Educational knowledge is travelling now with same rapidity as other bodies of knowledge – in medicine, industry and commerce – because of the effects of the information society created by the pervasiveness of the internet. There is much evidence that the world benefits by the spread of good practice in non-educational fields – by spreading best practice in automobile engineering, by medical best practice helping to deal with infections that stalk the planet, and by management systems such as 'just in time' or 'big data' or 'simultaneous top down and bottom up'. However, we now need to ensure that 'what travels' in the area of education is also valid and reliable because of the considerable practical and intellectual benefits of comparative work in our field:

- There is a real danger of educational damage when the interpretations of big data sets like PISA are associated with the attempted transplant of factors from country to country independent of the cultural contexts of these countries.

- Culturally sensitive explanations of 'what works' and 'why' in international studies can improve the quality of educational discourse more generally, generating more nuanced, complex sets of understandings that have greater explanatory power by being generated in multiple different national settings.

- Internationally based studies tap the full range of variation in school and classroom effects. *Within* any country, the range of school and teaching factors in their 'quality' and their 'quantity' is likely to be much smaller than *between* countries. Put simply, therefore, international studies are likely to show *greater* educational effects than within nation ones which have settled into a range of perhaps 10 to 20 per cent in variance explained (Chapman et al., 2015). The true power of school and classroom is, if this reasoning is correct, only likely to be shown by authentic international comparative work.

- Internationally comparative work helps in the generation of theory, where we still have an absence of theorising of even a middle-range variety, although some recent attempts may suggest this to be changing (Creemers & Kyriakides, 2008). Why is it that 'assertive Principal leadership' does not predict school effectiveness status in the Netherlands but does in Anglo Saxon societies (Bosker & Scheerens, 1994)? Why do some of our factors travel across socio economic contexts within countries better than others (Teddlie & Stringfield, 1993)? Answering these questions forces us to develop a deeper analysis, more complex explanations and more multi layered interactions between 'levels' of education which have theoretical potential;

- If comparative work were to show that 'what works' varies within different contexts, it would compel us to generate more sensitive, contextually specific explanation in our field. In the early days of EEIR, we largely researched in low socio-economic status communities, rendering ourselves unable to see if there was within-nation contextual variation in effective educational process factors. Later work, particularly in the United States, found these differences, with the particularly interesting finding of effective schools in poorer Californian communities actively pursuing policies to dis-involve their parents (Hallinger & Murphy, 1996; Teddlie & Stringfield, 1993). The District/Local Authority context-specific policies necessary to improve in socially challenged communities were a focus also in the UK (Harris et al., 2006; Muijs et al., 2004; Reynolds et al., 2001). This tradition was largely eroded by the 2000s, perhaps due to the across-context 'lists' (e.g. Reynolds et al., 1996), that were a feature of the 1990s and which were an attempt to extract 'what works' from the context-bound early studies, and which also sometimes arose out of government sponsored and funded projects where researchers might have seen any context specific formulations as 'inconvenient' to conventional 'one size fits all' educational policies.

For the policymaker/political sponsors of EEIR, it may also be that the absence of context-specific effectiveness research also reflected the desire to ape medical research by generating universal findings that could be applied in all contexts, as in Slavin's (1996) use of the phrase 'wherever and whenever we choose'. This is of course to fundamentally misunderstand medical research and practice, since the latter involves 'universal' treatments (e.g. statins) but applied in highly context-specific fashions to individual patients (with variation in the type of drug used, the dosage, the length of use, the possible combination with other drugs, and the sequence of combination if it was used). This context specificity in medicine has been ill understood. Another factor limiting context-specific formulations may have been the popularity of meta-analyses which avoided splitting up samples by context in the interests of maintaining high sample size (e.g. Hattie, 2009).

To conclude on a positive note, there is of course much that the EEIR community can do to generate high quality research into International Educational Effectiveness, and the following things seem to be particularly useful:

- The OECD has made available individual student level PISA data to facilitate research;

- EEIR researchers have been involved in conducting PISA studies, providing advice on issues such as questionnaire design and use, and providing advice on methodological issues;
- EEIR shows interest in interacting with the existing international comparative work on the effectiveness of international systems (Kyriakides et al., 2018). Existing studies like PISA and TIMSS use a repeated series design approach, so it is possible to add

educational factors like national policies in educational areas and changes in these policy areas to the student achievement data, to test the extent to which changes in these factors are associated with the effectiveness status of different country systems;

- EEIR shows recent interest (Chapman et al., 2015) in conceptualising and measuring social, in addition to academic, outcomes and there are increasing hints that these may be highly relevant to the apparent economic success of "high performing" societies;

- The OECD is gathering 'teacher' and 'teacher behaviour' data in the current TALIS, an area in which EEIR researchers have become more interested in the last decade;

- Small scale studies that use routinely available existing international social and economic data to consider the contextual effects upon educational achievement of different countries are powerful and profitable (Kelly, 2018). In this work, contextual factors such as targeted expenditure on education relate to overall PISA country scores, more than does overall expenditure. Likewise, mean PISA scores and 'resilience' in children (being in the bottom quarter socio economically but in the top quarter on achievement) are closely related;

- The picture of high achieving countries that has begun to emerge suggests more complex explanations than early formulations, focusing on cultural factors and pedagogical practice (Deng and Gopinathan, 2016). Disentangling these influences would seem entirely appropriate for researchers and practitioners in the EEIR paradigm.

- Methodological advances in the EER field (Creemers, Kyriakides & Sammons, 2010) including interest in mixed methods research may provide a fruitful way forward to explore in more detail within and between country school and classroom differences in student outcomes of various kinds (not just academic) and exploring the role of differences in culture and educational systems and their associations with variations in outcomes in new and creative ways (Sammons, 2010). It could also provide a richer picture of *within as well as between* system variation including country, regional and contextual perspectives, and within school variation too.

We would suggest with others (Rutkowski & Rutkowski, 2016) that it is time for EEIR and those responsible for international surveys and studies to suspend any possible turf wars and work collaboratively to deliver the promise of International Effectiveness Research.

5.6 References

Alexander, R. (2010). *Culture and Pedagogy: International comparisons in primary education.* Oxford: Basil Blackwell.

Alexander, R. (2012). Moral panic, miracle cures and educational policy: what can we really learn from international comparison? *Scottish Educational Review, 44* (1), 4-21.

Allerup, P. (2007). *Identification of group differences using PISA scales – Considering effects of inhomogeneous items. In S. Hopmann & G. Brinek (Eds.), PISA according to PISA*. Wien, Austria: Lit-Verlag, University of Vienna.

Anderson, L. W., Ryan, D. W. & Shapiro, B. J. (1989). *The IEA Classroom Environment Study.* Oxford: Pergamon Press.

Barber, M. (2007). *Instruction to deliver: Tony Blair, the public services and the challenge of delivery.* London: Politico's.

Barber, M. (2009). From system effectiveness to system improvement. In A. Hargreaves & M. Fullan (Eds.). *Change wars.* Bloomington, IN: Solution Tree.

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

Bosker, R. & Scheerens, J. (1994). Alternative models of school effectiveness put to the test. *International Journal of Educational Research, 21*, 159-180.

Bray, M. & Kobakhidze, M. M. (2014). Measurement Issues in Research on Shadow Education: Challenges and Pitfalls Encountered in TIMSS and PISA, *Comparative Education Review, 58*(4), 590-620.

Brophy, J. (1979). Teacher behaviour and its effects. *Journal of Educational Psychology, 71* (6), 733-750.

Brophy, J. & Good, T. L. (1986). Teacher behaviour and student achievement. In M. C. Wittrock (Ed.). *Handbook of research on teaching,* (3rd ed). New York, NY: Macmillan.

Bulle, N. (2011). Comparing OECD educational models through the prism of PISA. *Comparative Education, 47* (4), 503-521.

Burghes, D. & Blum, W. (1995). *The Exeter Kassel Comparative Project: A Review of Year 1 and Year 2 Results in Gatsby Foundation, Proceedings of a Seminar on Mathematics Education.* London: Gatsby.

Burges, S., Wilson, D. & Worth, J. (2010). *A natural experiment in school accountability: the impact of school performance information on pupil progress and sorting.* CMPO Working Paper Series No. 10/246. Bristol: CMPO.

Chapman, C., Muijs, D., Reynolds, D., Sammons, P. & Teddlie, C. (2015). *The International Handbook of Educational Effectiveness and Improvement.* London: Routledge.

Close, S., & Shiel, G. (2009). Gender and PISA Mathematics: Irish Results in Context. *European Educational Research Journal, 8* (1), 20-33.

Comber, L. C. & Keeves, P. (1973). *Science Education in Nineteen Counties.* London: John Wiley

Creemers, B. P. M. & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. London: Routledge.

Creemers, B., Kyriakides, L., & P. Sammons, P. (Eds.) (2010) *Methodological Advances in Educational Effectiveness Research,* London: Routledge Taylor Francis.

Dee, T. & Jacob, B. (2011). The Impact of No Child Left Behind On Student Achievement. *Journal of Policy Analysis and Management, 30* (3), 418-446.

Deng, Z. and Gopinathan, S. (2016). PISA and high performing education systems: explaining Singapore's educational success, *Comparative Education, 52*(4), 449-472.

Dobbins, M., & Martens, K. (2012). Towards an education approach à la finlandaise? French education policy after PISA. *Journal of Education Policy, 27* (1), 23-43.

Dolton, P., Gutierrez, O. M. & Still, A. (2015). *Educational efficiency: value for money in public spending on schools.* Paper No. CEPCP 441.

Eivers, E. (2010). PISA: Issues in Implementation and Interpretation. *The Irish Journal of Education/Iris Eireannach an Oideachais*, 94-118.

Elley, W. B. (1992). *How in the World Do Students Read?* Newark, DE: IEA.

Feniger, Y., & Lefstein, A. (2014). How not to reason with PISA data: an ironic investigation. *Journal of Education Policy, 29*(6), 845-855.

Fischbach, A., Keller, U., Preckel, F., & Brunner, M. (2013). PISA proficiency scores predict educational outcomes. *Learning and Individual Differences, 24*, 63-72.

Fullan, M. (2009). Large scale reform comes of age. *Journal of Educational Change, 10*, 101-113.

Gaber, S., Cankar, G., Umek, L. M., & Tašner, V. (2012). The danger of inadequate conceptualisation in PISA for education policy. *Compare: A Journal of Comparative and International Education, 42*(4), 647-663.

Goldstein, H. (2008). Comment peut-on utiliser les etudes comparatives international pour doter les politiques educatives d'information fiables?/ How can international comparative studies be used to provide reliable educational information policies? *Revue Francaise de Pedagogie/French Review of Pedagogy*. *164*, 69-76.

Goldstein, H. (2003). *Multilevel models in educational and social research (3rd ed.).* London: Edward Arnold.

Gorur, R., & Wu, M. (2014). Leaning too far? PISA, policy and Australia's 'top five' ambitions. *Discourse: Studies in the Cultural Politics of Education, 36*(5), 647-664.

Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy, 24*(1), 23-37.

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

Guldemond, H. & Bosker, R. (2009). School effects on student progress – a dynamic perspective. *School Effectiveness and School Improvement, 20*(2), 255-268.

Hallinger, P. & Murphy, J. (1986). The social context of effective schools. *American Journal of Education. 94*, 328-355.

Hanberger, A. (2014). What PISA intends to and can possibly achieve: A critical programme theory analysis*. European Educational Research Journal, 13*(2), 167-180.

Hanusheck, E. & Raymond, M. (2005). Does school accountability lead to improved school performance? *Journal of Policy Analysis and Management, 24*(2), 297-327.

Hargreaves, A. & Shirley, D. (2009). *The fourth way.* Thousand Oaks, CA: Corwin.

Harris, A., Chapman, C., Muijs, D., Russ, J. & Stoll, L. (2006). Improving schools in challenging contexts: Exploring the possible. *School Effectiveness and School Improvement, 17,* 409-424.

Harris, A. & Chrispeels, J. (Eds.). (2008). *International perspectives on school improvement.* London: Routledge Falmer.

Harris, A, & Jones, M. (Eds.). (2015). *Leading Futures.* Singapore: Sage.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London: Routledge.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., ….(2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study.* Washington: National Centre for Educational Statistics.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., and Baird, J. A. (2016) Lessons learnt from PISA: A systematic review of peer reviewed articles on the programme for international student assessment, *The Scandinavian Journal of Educational Research, 62*(3), 333-353.

Hopkins, D. (2007). *Every school a great school.* Maidenhead: Open University Press.

Hopkins, D. (2013). *Exploding the myths of school reform.* Maidenhead: Open University Press.

Hopkins, D., Harris, A. & Jackson, D. (1997). Understanding the school's capacity for development: Growth states and strategies. *School Leadership and Management. 17*, 401-412.

Hopkins, D. & Reynolds, D. (2001). The past, present and future of school improvement: Towards the third age. *British Educational Research Journal. 27*, 459-475.

Hopkins, D., Stringfield, S., Harris, A., Stoll, L. & Mackay, T. (2014). School and system improvement: a narrative state-of-the-art review, *School Effectiveness and School Improvement. 25*(2), 257-281.

Husen, T. (Ed.). (1967). *International Study of Achievements in Mathematics, Volumes One and Two.* Stockholm: Almquist and Wiksell.

Kankaraš, M., & Moors, G. (2013). Analysis of Cross-Cultural Comparability of PISA 2009 Scores. *Journal of Cross-Cultural Psychology, 45*(3), 381-399.

Keeves, J. P. (1992). *The IEA Study of Science 111: Changes in Science Education and Achievement, 1970 to 1984.* Oxford: Pergamon Press.

Kelly, A. (2018) *PISA – Explaining new contextual factors at the level of the nation state. Unpublished paper.* Southampton: University of Southampton.

Kyriakides, L., Giorgiou, M. P., Creemers, B. P. M., Panayiotou, A. Reynolds, D (2018) 'The Impact of National Policies on student achievement: a European Study, *School Effectiveness and School Improvement, 29*(2), 171-203.

Keys, W. & Foxman, D. (1989). *A World of Differences (A United Kingdom Perspective on an International Assessment of Mathematics and Science).* Slough: National Foundation for Educational Research.

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79* (2), 210-231.

Lapointe, A. E., Mead, N. & Phillips, G. (1089). *A World of Difference: An International Assessment of Mathematics and Science.* New Jersey: Educational Testing Services.

Lewis, S. (2014). The OECD, PISA and educational governance: a call to critical engagement. *Discourse: Studies in the Cultural Politics of Education, 35* (2), 317-327.

Loveless, T., (2013). *Attention OECD-PISA: Your Silence On China Is Wrong.* Washington D. C.: Brown Center on Education Policy At Brookings.

Loveless, T., (2014). *PISA's China Problem Continues: A Response to Schleicher, Zhang & Tucker.* Washington D.C.: Brown Center on Education Policy At Brookings.

Messick, S. (1998). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed). New York: Macmillan.

Meyer, H. D. & Schiller, K. (2013). Non-educational influences on PISA outcomes. In H. D. Meyer & A. Benavot (Eds.), *PISA, Power and Policy: the Emergence of Global Educational Governance*, Oxford: Symposium Books.

Morgan, C., & Shahjahan, R. A. (2014). The legitimation of OECD's global educational governance: examining PISA and AHELO test production. *Comparative Education*, 50(2), 193 – 205.

Mourshed, M, Chijioke, C. & Barber, M. (2010). *How the world's most improved school systems keep getting better.* London: McKinsey.

Muijs, D. & Reynolds, D. (2011). *Effective teaching. Evidence and practice*. London: Sage.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H. & Earl, L. (2014). State of the art – teacher effectiveness and professional learning, in *School Effectiveness and School Improvement*, *25*(2), 231-256.

Muijs, D., Harris, A., Chapman, C., Stoll, L. & Russ, J. (2004). Improving schools in socio-economically disadvantaged areas: A review of research evidence. *School Effectiveness and School Improvement, 15*, 149-175.

Mullis, I.V.S., Martin, M.O., & Foy, P. (2013). The Impact of Reading Ability on TIMSS Mathematics and Science Achievement at the Fourth Grade: An Analysis by Item Reading Demands. In M. O. Martin, & I.V.S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade — Implications for early learning* (pp.67–108). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Murphy, D. (2014). Issues with PISA's Use of its Data in the Context of International Education Policy Convergence. *Policy Futures in Education, 12* (7), 893-916.

Murphy, S. (2010). The pull of PISA: Uncertainty, influence, and ignorance. *Inter-American Journal of Education for Democracy, 3* (1), 27-44.

OECD, (2009). *PISA 2009 Assessment Framework: Key Competencies in Reading, Maths and Science.* Paris: OECD.

OECD.(2015). *PISA frequently asked questions.* Retrieved from http://www.oecd.org/pisa/aboutpisa/pisafaq.htm

Pereyra, M. A., Kotthoff, H. G., & Cowen, R. (2011). *PISA under examination*. SensePublishers.

Postlethwaite, T. N. & Ross, K. (1992). *Effective Schools in Reading: Implications for Educational Planners.* Newark, DE: IEA.

Postlethwaite, T. N. & Wiley, D. E. (1992). *The IEA Study of Science 11, Science Achievement in Twenty Three Countries.* Oxford: Pergamon Press.

Purves, A. C. (1992). *The IEA Study of Written Composition 11: Education and Performance in Fourteen Countries.* Oxford: Pergamon Press.

Rautalin, M., & Alasuutari, P. (2009). The uses of the national PISA results by Finnish officials in central government. *Journal of Education Policy, 24* (5), 539-556.

Reynolds, D. (2008). New Labour, education and Wales: The devolution decade. *Oxford Review of Education, 34,* 753-765.

Reynolds, D., Creemers, B. P. M., Stringfield, S., Teddlie, C., Schaffer, E. & Nesselrodt, P. S. (1994). *Advances in school effectiveness research and practice.* Oxford: Pergamon Press.

Reynolds, D. & Farrell, S. (1996). *Worlds Apart? A Review of International Surveys of Educational Achievement Involving England.* London: HMSO for OFSTED.

Reynolds, D., Bollen, R., Creemers, B. P. M., Hopkins, D., Stoll, L. & Lagerweij, N. (1996). *Making good schools: Linking school effectiveness and school improvement.* London: Routledge.

Reynolds, D., Creemers, B. P. M., Stringfield, S., Teddlie, C. & Schaffer, E. (2002). *World class schools: International perspectives in school effectiveness.* London: Routledge Falmer.

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C. & Stringfield, S. (2014). Educational effectiveness research (EER): a state of the art review, *School Effectiveness and School Improvement. 25* (2), 197-230,

Reynolds, D., Caldwell, B., Cruz, R. M., Miao, Z., Murillo, J., Mugendawata, H., Mayol, B. D. L. E., Medina, C. P. & Ramon, M. R. R. (2015). Comparative Educational Research in C. Chapman, D. Muijs, D. Reynolds, P. Sammons & C. Teddlie (2015). *The Routledge*

International Handbook of Educational Effectiveness and Improvement. London: Routledge.

Rindermann, H. (2007). The g-Factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality, 21*(5), 667-706.

Rindermannm H, & Baumeister, A.E.E. (2015). Validating the interpretations of PISA and TIMSS tasks: A rating study. *International Journal of Testing, 15* (1), 1-22.

Rosier, M. J. & Keeves, J. P. (1991). *The IEA Study of Science 1, Science Education and Curricula in Twenty Three Countries.* Oxford: Pergamon Press.

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

Rutkowski, L. & Rutkowski, D. (2016) A Call for a More Measured Approach to Reporting and Interpreting PISA Results*. Educational Researcher, 45*(4), 252-257.

Sahlberg, P. (2011). *Finnish Lessons.* New York, NY: Teachers College Press.

Sammons, P. (2010). The contribution of mixed methods to recent research on educational effectiveness. In Tashakkori, A., & Teddlie, C. *SAGE handbook of mixed methods in social & behavioral research* (pp. 697-724). Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781506335193

Sellar, S., & Lingard, B. (2013a). Looking East: Shanghai, PISA 2009 and the reconstitution of reference societies in the global education policy field. *Comparative Education, 49* (4), 464-485.

Sellar, S., & Lingard, B. (2013b). The OECD and global governance in education. *Journal of Education Policy, 28* (5), 710-725.

Sjøberg, S. (2012). PISA: Politics, fundamental problems and intriguing results. *Recherches en Education*, 14, 1-21.

Slavin, R. E. (1996). *Education for all.* Lisse: Swets & Zeitlinger.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S. & Serrano, A. (1999). *The TIMSS Videotape Classroom Study: Methods and findings from an exploratory research project on*

eighth grade mathematics instruction in Germany, Japan and the United States. Washington, DC: National Center for Education Statistics.

Teddlie, C. & Stringfield, S. (1993). *Schools make a difference: Lessons learned from a 10-year study of school effects.* New York: Teachers College Press.

Teddlie, C. & Reynolds, D. (2000). *The international handbook of school effectiveness research.* London: Falmer Press.

Tienken, C. H., & Mullen, C. A. (2014). The Curious Case of International Student Assessment: Rankings and Realities in the Innovation Economy. *Building Cultural Community Through Global Educational Leadership*, 146-164.

Travers, K. J. & Westbury, I. (1989). *The IEA Study of Mathematics 1: Analysis of Mathematics Curricula.* Oxford: Pergamon Press.

Waldow, F., Takayama, K., & Sung, Y. K. (2014). Rethinking the pattern of external policy referencing: media discourses over the 'Asian Tigers'' PISA success in Australia, Germany and South Korea. *Comparative Education, 50*(3), 302-321.

Whelan, F. (2009). *Lessons learned: How good policies produce better schools.* London: Author.

Wuttke, J. (2007). Uncertainty and Bias in PISA. In S. T. Hopmann, G. Brinek & M. Retzl (Eds.), *PISA According to PISA: Does PISA Keep What It Promises?* pp. 241-263. Berlin: Lit Verlag.

Yildirim, H. H. (2006). *The differential item functioning (DIF) analysis of mathematical items in the international assessment programs.* (Unpublished doctoral thesis). Middle East Technical University, Ankara, Turkey.

Zhao, Y. (2014). *Who's afraid of the big bad dragon: China the best and worst education system in the world.* San Francisco, CA: Jossey Bass.