Aust. Educ. Res.
DOI 10.1007/s13384-015-0183-2

# PISA Data: Raising concerns with its use in policy settings

Shelley Gillis[1] · John Polesel[1] · Margaret Wu[2]

Received: 22 December 2014 / Accepted: 26 June 2015

**Abstract** This article considers the role played by policy makers, government organisations, and research institutes (sometimes labelled "think tanks") in the analysis, use and reporting of PISA data for the purposes of policy advice and advocacy. It draws on the ideas of Rizvi and Lingard (Globalizing Education Policy, 2010), Bogdandy and Goldmann (Governance by Indicators/ Global Power through Quantification and Rankings, 2012) and others to explore the ways in which such "agents of change" can interpret, manipulate and disseminate the results of data arising from large scale assessment survey programs such as PISA to influence and determine political and/or educational research agendas. This article illustrates this issue by highlighting the uncertainty surrounding the PISA data that have been used by a number of prominent, high profile agents of change to defend policy directions and advice. The final section of this paper highlights the need for policy makers and their advisors to become better informed of the technical limitations of using international achievement data if such data are to be used to inform policy development and educational reforms.

**Keywords** PISA · Country rankings · Causal inferences · Think tanks · Data limitations

✉ Shelley Gillis
sgillis@unimelb.edu.au

John Polesel
jpolesel@unimelb.edu.au

Margaret Wu
wu@edmeasurement.com.au

[1] The University of Melbourne, Parkville, Australia

[2] National Taiwan Normal University, Taipei, Taiwan

Published online: 04 September 2015

This article considers the role played by policy makers, government organisations and research institutes (sometimes called "think tanks") in analysing, using and reporting the Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA) data for the purposes of policy advice and advocacy. It explores the ways in which think tanks have flourished in a climate characterised by a shift from government to governance and how they interpret, manipulate and disseminate the results of data arising from large scale assessment survey programs such as PISA to influence and determine political, educational and research agendas. It also highlights problems in the way PISA data have been used by some prominent, high profile agents of change to defend policy directions and advice. Finally, it highlights the need for policy makers and their advisors to become better informed of the technical limitations of using international achievement data if such data are to be used to inform policy development and educational reforms.

PISA was designed to assist governments to monitor the outcomes of education systems on a regular basis using an internationally accepted common framework for assessing and reporting student achievement. It is a sample-based study of young adults, at age 15, covering the domains of reading literacy, mathematical literacy, scientific literacy and problem solving. Since its introduction in 2000, the publication of PISA results has attracted wide media coverage both in Australia and internationally. At the same time, it has led to increasing pressure within participating countries for government agencies to publicly defend their country's performance.

Bogdandy and Goldmann (2012) argue that the PISA program itself constitutes a form of control which is best described as governance rather than government, in which data, analysed and interpreted by agents of change, are used to influence policy making—described as a form of governance by information or by indicators. They suggest that the analysis and publication of PISA results have an effect on government policies, especially with respect to the introduction of "standards" in education. In this sense, the OECD itself plays the role of a think-tank. It is this status in itself which may lead policy makers and high profile commentators to attempt to identify features within high performing educational systems and use these inappropriately to make causal inferences. We will argue that PISA can only tell us how countries have performed on a single assessment task (in terms of strengths and weaknesses) but cannot provide any information on why or how to improve performance. Yet attempts are made to use PISA data to make such inferences, which we will illustrate in this paper are often not justifiable and are often undertaken to support a political position or justify an existing reform.

Other studies have argued that the policy impact of PISA is also magnified by the peer effect which arises from comparisons between countries and in particular by their standing in league tables (Figazzolo 2009). Grek (2009), however, notes that this impact may be differentiated by the extent to which different nations have already embraced certain types of reform, with some nations such as the United States and the United Kingdom having already moved significantly in the direction of marketisation and increased accountability measures. By way of contrast, the impact of poor PISA results in Germany after the 2000 program resulted in questioning of the secondary education structures and in particular the impact of early tracking, with, admittedly, little change in actual policies.

Ball and Exley's recent research (2010), conducted using a methodology called "network ethnography" and based on internet searches and links as well as interviews with policy actors, has further contributed to our understanding of the manner in which this type of governance operates and in particular its relationship to government. Ball (1990) argues that policies are formed by groups of actors who come together to form Networks. These actors may include non-government organisations such as Teach for xxxx (e.g. Australia, America, etc.), private trusts and charities, major corporations such as Boston Consulting, McKinsey, Pearson and KPMG, and research and policy institutes. Ball argues that these networks are often made up of individuals who have travelled and studied overseas and have close personal, social and professional links. The networks are therefore built on trust and personal relationships.

What is important is that these networks, because of their international linkages and commonality of shared experiences and views, come together to facilitate "Global Forms", that is common approaches across different and sometimes widely differing national contexts. These forms typically include initiatives such as public private partnerships (PPPs), the implementation of standardised and high stakes forms of assessment, impact measurement, data-driven learning, Charter Schools/ Academies, blended learning, entrepreneurship, leadership, vouchers and phonics, amongst others. Typically, these reflect market-based reforms, strongly influenced by neoliberal economic theories and what has been described as the neoliberal imaginary (Rizvi and Lingard 2010).

Ball, as well as others (e.g. Rizvi and Lingard 2010) have argued that this leads to a very different role for national (and state or provincial) governments, with the state becoming a contractor or facilitator of services, rather than the leader or designer of policy—i.e. governance, not government. This changes the role of the state, but perhaps just as crucially it leads to local policy actors promoting and advocating for "Global Forms" of policy and acting globally in their own right, using data like that generated by PISA. Rizvi and Lingard (2010) argue that in this context the global and local binary becomes redundant, leading to what Robertson (1995) describes as "glocalisation", a hybrid of the global and local.

The real problem with this is that it may lead to decisions being made and Global Forms selected and promoted without accompanying credible research evidence or with evidence which is poorly understood or deliberately misapplied. Just as worryingly, it may lead to a decontextualized application of solutions or initiatives which are inappropriate or ill-suited to the local national contexts.

Important amongst the many policy actors are think tanks or research and policy institutes which act as agents of change. These organisations may be funded by governments, by business, or from consulting and research work but are in a position to influence policy. Rizvi and Lingard (2010) argue that organisations such as these influence policy goals and accountability regimes established by government with the reported data being used for "steering at a distance via performance measures (p. 119)".

Using the PISA program as an example, this article highlights the uncertainty surrounding the use of large scale assessment surveys which are not always taken into consideration (either by accident or by design) when used by prominent, high profile agents of change for policy reform purposes. To illustrate such a concern, we focus on two major limitations of using student achievement survey data for policy reform purposes. The first is associated with attempts to make direct causal links between student achievement and factors impacting on achievement. The second is associated with the need to be more cautious with interpreting and using country rankings given the technical challenges and limitations associated with measuring and monitoring achievement trends over time.

## Caution with making claims of causation

One of the major problems with using PISA data to inform policy reform is that often links are unjustifiably made between student achievement and factors impacting on achievement. Surveys such as PISA assess the level of student performance and at the same time collect student and school background information. While such a survey can provide information about how a country has performed and can highlight strengths and weaknesses of a particular educational system, it has limited capacity to identify how student performance could be improved (Buckingham 2012). It may provide some hypotheses about factors leading to educational success, but any links between student background information and student achievement are only conjecture. Statistics alone cannot prove causal relationships (Wu 2014). In fact, the OECD, in a recent report, stated the following:

> While PISA cannot identify cause-and-effect relationships between inputs, processes and educational outcomes, it can highlight key features in which education systems are similar and different, sharing those findings with educators, policy makers and the general public.
>
> (Vol. V, p 18, OECD 2010a).

While PISA can report similarities between education systems with high achieving results, it is not possible to directly link these similarities to factors for higher educational performance. To begin with, PISA has not collected information on all factors relating to educational success. For example, private investments in education such as parental support and after-school studies have not been formally included in PISA data. Further, while student and school background variables may correlate with achievement, one cannot draw any definitive conclusion as to which factors actually relate directly to academic performance. This is because there are often mediating variables for the positive correlations between any two variables. For example, students from homes with dishwashers have been found to have higher average educational performance than students from homes without dishwashers. This is not because dishwashers can lead to success in education, but may be because they reflect the presence of higher or lower income families. Furthermore, higher income in itself may not contribute directly to educational success either.

Higher income may lead to better educational resources available to children. Parents in higher income families may also have high expectations and aspirations for children to do well in education. So there are many levels of mediating variables between student background variables and achievement, and it is very difficult to positively prove any direct link between academic performance and student background characteristics.

Yet, as a result of a global push toward greater transparency and accountability of educational systems, there is increasing political pressure for government organisations to defend PISA results and look to change agents to identify causal connections and to provide guidance on future directions and/or solutions (Buckingham 2012; Zyngier 2014). A recent example of this is a report prepared by the Grattan Institute, which describes itself as an independent think tank which contributes to public policy in Australia. A recent, high profile publication of the Grattan Institute, *Catching up: Learning from the best school systems* (Jensen et al. 2012) presents analyses which raise concerns about the drawing of causal inferences based on observed correlations (e.g., Gorur and Wu 2014). Researchers from the Grattan Institute visited Hong Kong, Shanghai, Korea and Singapore to learn from these four top performing countries. They met government officials, principals, teachers and researchers and collected documentation at the system levels. The researchers examined various educational policies in these four countries and claimed that "*The full report provides … information on the design and implementation of the programs that underpin success*" (Jensen 2012, p. 5). This indicates that an inference had been already made for factors of success. Yet there was little, if any, attempt in the report to provide empirical evidence to support the claim that the education policies reported actually were the levers for the high educational performance in these four countries. As such, the claim contains an element of speculation which should not be so unequivocally stated without strong support.

One key element excluded from the Grattan Institute study and analysis is consideration of the impact of the home environment in terms of parental expectations and cultural influences. When one examines the educational systems and societal values in the top performing Asian countries in PISA, it becomes clear that there is a wide range of influences on academic achievement that are external to educational structures and policies (e.g., the wider social conditions and cultural history of the nation) that are often ignored by the media and by politicians when making PISA country comparisons (Feniger and Lefstein 2014; Alexander 2010). For example, it has been well documented that the top performing Asian countries tend to have a well-entrenched examination/testing culture within their schools, high student participation rates at supplementary education classes offered outside the formal system of education (referred to as 'cram schools' in Asia and 'coaching colleges' in Australia) and strong parental pressures and expectations for student success (e.g., Liu 2013; Feniger and Lefstein 2014; Buckingham 2012). Yet, such factors are ignored when 'top performing' systems and policies are reviewed for comparative purposes.

It has also been well documented that Asian parents hold higher education expectations for their children than white Anglo-Saxon parents (e.g., Goyette and Xie 1999; Hao and Bonstead-Bruns 1998). For example, in Taiwan where there is an ancient Chinese saying "*all pursuits except studying are of little value*", education is highly valued and judged as a strong indicator of success among society, more so than wealth or occupational status (Liu 2013, p. 490). Given such values, the majority of Taiwanese students attend some form of supplementary education classes in the evening and/or on weekends in an effort to improve their grades and test scores so that they have a better chance of success (Liu 2012, 2013; Wei and Eisenhart 2011). Not only is attendance at evening and weekend supplementary classes highly valued by parents and students in Taiwan, but Liu's (2012) investigation of nearly 10,000 Taiwanese students found that attendance at such classes had a clear positive influence on students' academic performance, as measured by test scores on a general analytical ability and math performance tests. Yet such cultural differences have not been considered in the PISA analyses nor do they tend to be adequately addressed by the media, politicians or influential commentators when there is a temptation for international comparisons and the manipulation of PISA data to occur for political or partisan purposes (Alexander 2010).

To illustrate the importance of this omission, the third author of this paper provides the following account of her normal school day at the age of 14 in Taiwan:

> I would arrive at school at 7:30 in the morning. After an 8-h day at school, my school friends and I would go to a large study room provided by a local church to do homework. I would arrive at home at about 6:30 pm. After dinner, I would do more homework until 9 pm. I would go to sleep for 3 h and ask my parents to call me at 12 midnight when my parents went to sleep. I would continue to study until 2–3 am, and get up at 6 am for school again. Most of my school friends did the same. The idea of having a sleep at 9 pm was to be more refreshed when studying again after midnight.

This pattern of a high school student's life in Taiwan has not changed in recent years (Wei and Eisenhart 2011). Hwang (2015) argues that "the study hard phenomenon is still very much alive for secondary students in Taiwan today"….." if they go to bed by midnight, they consider themselves lucky" (p. 129). Attendance at supplementary education classes has become so popular in Taiwan that by the time students enter college, the majority have attended numerous years of after school tutoring in the evenings and on weekends (Hwang 2015; Liu 2012, 2013; Wei and Eisenhart 2011). It should be noted that this phenomenon is not unique to Taiwan. Other high PISA performing East Asian countries such as Japan and Korea have strong parental pressure to study hard to succeed, including attendance at supplementary educational classes outside of school hours (Hwang 2015).

The Grattan Institute report also made an interesting observation that "Hong Kong acknowledges that its move away from a strict examination focus has not yet persuaded most parents" (Jensen et al. 2012, p. 2). The same issue has been raised in Taiwan in recent months. Following the announcement from the Ministry of Education of moving to a 12-year compulsory education to replace the 9-year compulsory education in Taiwan in 2014, the high-stake examination for junior high school students to be admitted to senior high schools has been abolished (see Ministry of Education 2013). The removal of the senior high school selection test has met with strong opposition from parents who demand the government bring back the examination (see Lee 2014), even though such examinations may lead to hardship and psychological stress for students (Hsin and Xie 2014). The main reason for the strong objection from parents in relation to abolishing the examination is the fear that their children would not continue to exert high levels of academic effort through extended periods of study if there were no motivation to prepare for such an examination. Such concerns arise from the strong cultural belief that effort and achievement are strongly connected (Hsin and Xie 2014; Chen and Stevenson 1995). The desire to do well academically is so high that parents prefer a rigid examination system that drives young people to study hard, which in turn is thought to lead to greater academic success. This mindset is culturally different to that of parents in western countries where academic success is typically thought to be largely dependent upon intelligence and innate ability and where test scores alone are not adequate measures of success (Wei and Eisenhart 2011).

It should be acknowledged however that not all high profile commentators and think tanks discount the importance of cultural differences when interpreting PISA data and country performance. For example, a recent study produced by the Centre for Independent Studies (i.e., a high profile think tank based in Australia), has urged policy makers to be more cautious when attempting to draw inferences between any one feature of a country's education system and its success on PISA with another:

> The top ranking PISA and TIMSS countries are often very different to Australia socially, culturally, demographically, geographically and linguistically. These features influence education policy and performance. Student performance in PISA and TIMMS must be viewed in this context.
>
> (Buckingham 2012, p. 1).

By way of contrast, and given the well documented phenomenon that the educational aspirations and expectations of East Asian parents generally far exceed that of their western counterparts, it is surprising that the Grattan Institute report discounted cultural differences and parental expectations as a possible reason for the top performing Asian countries, as evident in the following statement:

> Nor is success culturally determined, a product of Confucianism, rote learning or Tiger Mothers.
>
> (Jensen et al. 2012, p. 12).

It appears that many policy-makers may also be ignoring the OECD's caution about drawing inferences from PISA results at the government policy level, as examples of the misuse of PISA results also exist in this sphere. For instance, in recent years the New Zealand government has announced a number of education policies placing the onus on teachers for students' academic performance, while quoting PISA findings as the basis for these policies (Thrupp 2014). In 2012 the New Zealand government proposed to increase class sizes to free up funds for teacher professional development. The proposal cited a number of research studies including PISA. In fact, the 2012 OECD report made the following claim:

> …many successful school systems share some common features: … spending in education that prioritises teachers' salaries over smaller classes
>
> (Vol. IV, p. 29, OECD 2010b).

Yet Wu (2014) has found that there was no statistical evidence from the OECD PISA data to support such a claim. That is, she found that the performances of countries with large class sizes and high teachers' salaries were not statistically higher than the performance of countries with small classes.

In Australia, high profile commentators such as Ashenden (2013) and Jensen (2010) have also pointed to PISA's high performing education systems such as Shanghai, Hong Kong, South Korea, Taiwan and Singapore, in which large class sizes are the norm, as evidence that reducing class sizes has no impact on student outcomes. Such claims have been picked up in Australia and internationally by the media and education politicians (e.g., The Economist, Feb 18th 2012; Pyne 2012). Yet Zyngier (2014) and Tran (2012) argue that such claims fail to acknowledge that students from Confucian heritage cultures tend to be passive learners who have been socialized in ways that make them amenable to large classes using whole-class teaching methods. Therefore, the proposal to increase class sizes in New Zealand is surprising given that a large class size is not compatible with the approach of modern western education which emphasizes individualized learning and creativity. This proposal was later withdrawn because of a backlash from the public—mostly parents and teachers. The New Zealand prime minister was quoted as saying the following:

> We were effectively saying to the sector 'here is quite a lot of cash to fund [teacher] development [funded] by making what we think—I still think—is a very modest alteration of class sizes'. But what is clear is that parents don't see it as modest and in the end perception is reality.
>
> (cited in Watkins et al. 2012).

In addition to the class size issue, the New Zealand Minister of Education claimed that PISA results indicated that only 18 % of the student achievement variance was attributable to students' socio-economic status (Thrupp 2014). The implication of this claim is that students' academic success rests mostly on the effectiveness of the teachers. This claim is inconsistent with the findings from most studies on the impact of SES on student achievement. For example, in the literature, it has generally been found that teacher effect accounts for only around 10 % of the student performance variance (e.g., Leigh 2010; Byrne et al. 2009). Thrupp (2014) reported that further investigation revealed that the 18 % claim was based on a narrow definition of SES defined in PISA. If a wider definition of SES is used, about 78 % of New Zealand students' achievement variance can be explained by socio-economic conditions.

The above examples highlight a lesson one can learn. That is, owing to the complexity of the relationship between educational achievement and multiple factors, results can be cherry-picked to suit one's purpose. Findings from one study can often be inconsistent with findings from other studies, yet individuals, organisations and think tanks can look for evidence (either consciously or subconsciously) to support their position or to justify policies and directions to which a government has already committed itself (Tuchman 1984). If findings from large scale assessment programs such as PISA are to be used to inform education policy development, it is essential for users (whether that be politicians, commentators, think tanks or the media) to be statistically literate and cross-check a variety of sources and research methodologies. It is also vital that such users consider non-statistical aspects of education such as those aspects valued by a particular culture, and to recognize that the OECD views are not always appropriate for the local context. In fact, a great deal of common-sense-making is essential in turning the results from international assessments into local policies.

In Australia, while education policies have not been greatly affected by international assessment results in the shorter term, there have been cases where PISA results have been over-emphasised and over-interpreted. In August 2012, Julia Gillard, then Prime Minister of Australia, declared that Australia would strive to be ranked in the 'top five' in international education assessments by 2025. This declaration has been incorporated into the *Australian Education Act 2013*. However, there are some difficulties in achieving such a target in practical terms. Given that rankings of countries are about relative performances, even if Australia strives to improve its education, it would only improve its ranking if other countries did not improve or if Australia improved more than other countries did. Furthermore, one needs to question whether rankings in PISA indeed reflect success in education in a broader context (e.g., Gorur and Wu 2014). It is somewhat limited to set a major goal for Australia in education based on the rather narrow basis of performance in international rankings in large-scale assessments.

The final section of this paper explores country rankings in more detail and explains why sometimes such rankings can be affected by technical issues and other problems that are not always acknowledged by the media, policy makers, think tanks and high profile commentators.

## Caution with drawing inferences from country rankings

In this section we look at some technical issues relating to measuring change over time (trends) in international surveys. We first show a graph of country mean score change in PISA reading between 2000 and 2009 (Fig. 1: PISA reading trends between 2000 and 2009).

In Fig. 1, the countries with positive reading change scores (improvement from 2000 to 2009) are at the left side of the graph, while countries with negative change scores (decline) are at the right side of the graph. The darker bars indicate that the change in score is statistically significant while the lighter bars indicate that the change in score is not statistically significant. It can be seen that Ireland has the largest decline with 31 PISA points from 2000 to 2009, and that this drop is statistically significant. It can be seen that all English speaking countries are on the right-side of the graph, showing a decline in reading scores, although some of these are not statistically significant. The countries on the extreme left side of the graph (showing large improvement) are mostly lower performing countries. In fact, the

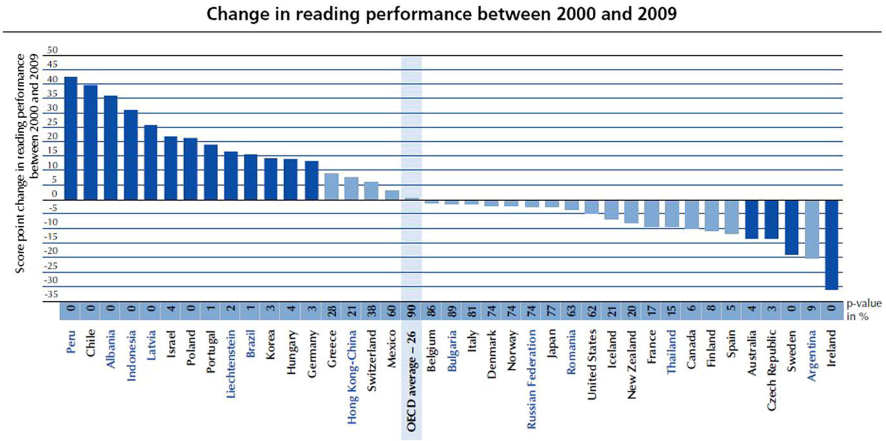

**Fig. 1** PISA reading trends between 2000 and 2009

correlation between country mean score in 2000 and country change score (2009–2000) is −0.6, showing that the higher the country mean score in 2000, the larger the decline is likely to be in 2009. One way to look at this result is that it appears that countries in the west (English speaking countries and western/northern European countries) generally performed higher in 2000 than their relative performance in 2009. This phenomenon of similarities in trends between groups of countries is extremely unlikely to reflect a real change in country mean scores. It is more likely to be caused by some systematic issues in measuring trends across different countries. We discuss some of these issues below.

## Measurement issues

As illustrated earlier in this paper, one major difficulty in conducting international surveys is that there are cultural and language differences between countries. Even though the translations of test items are double-checked, two languages are never fully equivalent. Further, owing to different curricula and life experiences, the relative item difficulties are often different across countries. For example, students in one country may find a test item on high-speed trains easier than an item about snow, while students in another country may find the item about snow easier than the item on high-speed trains simply because they live in a country that has snow but has no high-speed trains. In educational measurement, the term *differential item functioning* (DIF) or item bias is used to indicate that the relative item difficulties are not the same across multiple groups of students. Many research studies have identified DIF items in international student assessments (e.g., Grisay et al. 2009; Wu 2009). The existence of DIF items threatens the reliable measurement of student performance, and even more so, it threatens the measurement of trends over time.

To measure trends in large-scale assessments over time, typically a set of test items are selected as “link items” across two assessment cycles (e.g., 2000 and 2003). Based on students’ performances on these link items, the tests from two assessment cycles can be placed on the same measurement scale so results can be compared to provide a measure of change over time. Because there are DIF items across countries, the selection of the link items has an impact on a country’s results. For example, if most link items happen to be biased against country A (i.e. students in country A find these link items more difficult than other items, but students in other countries do not), then country A’s mean score is likely to be lower compared to their mean score should another set of link items be chosen. Monseur and Berezner (2007) showed that Japan’s mean score would increase by 10 score points in PISA 2003 if one particular reading passage was removed from the set of eight linking reading passages. A difference of 10 score points on the PISA scale is about 2–3 months of growth in achievement. Such a large difference would change the rankings of the countries. It just happened by chance that the PISA reading link items between 2000 and 2003 had many items biased against Japan.

Urbach (2013) explored the use of two different methods in measuring trends. He compared the results on trends by using international item difficulty parameters and using Australian item difficulty parameters in scaling the student item response data. If there were no DIF items, the two methods should produce the same trend results. Since there were DIF items, the international item difficulty parameters were not always the same as country specific item difficulty parameters, so different trends results were obtained. Urbach noted the following:

> Published Australian Reading distributions reported a decline over the first three cycles in the performance of Australian students located at the top end of the distribution. Using Australian data only, a decline between the first two PISA cycles was found, but remarkably in the bottom 15 % of the distribution only. Between cycles 2003 and 2006 an almost constant decline across the whole proficiency distribution was found and not a decline that was limited to the top end of the distribution, as published by the media.
>
> (Urbach 2013, p. 165).

What Urbach has found is that the trend results are substantially different when Australian data are used for the estimation of item difficulties instead of using international item parameters. So, clearly, differential item functioning (DIF) is present in the data and has an impact on trend results. Should the published results be used for policy reform, the reform could be on the wrong track if the focus is on raising student performance at the top end of the ability distribution. We note that Urbach carried out this research as a staff member of the Australian Council for Educational Research (ACER) which was also the institution responsible for producing and publishing the international results. What this shows is that there is an on-going exploration to improve the methodologies of large-scale assessment among the psychometric community both nationally and internationally, and there is not one single unequivocally best method. Consequently policy makers need to be aware that any published results are always subject to the limitations in survey methodology and data. Any interpretations of assessment results need to be made in that light. Hence, those who analyse, digest and disseminate the data arising from programs such as PISA should clearly communicate the limitations of such data in an accessible, non-statistical format to avoid potential misuse and/or misinterpretation by end users, including the media and policy makers.

## Contextual issues

In this section we use Ireland's PISA trend data to highlight the danger associated with interpreting country level trends in the absence of a thorough understanding of the local political, historical, cultural, economic, demographic and social conditions of that particular country. As can be seen in Fig. 1, Ireland had a decline of PISA reading scores in 2009. Cosgrove and Cartwright (2014) noted that the PISA 2009 results were unexpected given the absence of other evidence of a decline in educational standards in Ireland. They carried out an in-depth study of the factors that might explain a 30-point drop. Cosgrove and Cartwright reported a number of changes in the implementation of PISA 2009 in Ireland. The first was the introduction of a prize draw to incentivize student participation, an incentive that could increase participation and have an unexpected impact on the composition of the achieved sample. Second, test sessions in three-quarters of the schools were administered by teachers at their own schools instead of external administrators. Third, the sampling process was changed owing to the simultaneous participation of Ireland in two international assessments in 2009 (PISA and the Civics study).

In addition to the implementation changes in PISA in Ireland, Cosgrove and Cartwright found that there were differences in the PISA cohorts between the 2000 and 2009 samples. In 2009 there were eight "outlier" schools with very low average reading scores while the 2000 sample did not have any outlier schools. There were several possible reasons for the cohort change. First, the number of immigrant students participating in PISA had increased from 2.3 % in 2000 to 8.3 % in 2009 in Ireland. Second, the retention rates of students in schools for 15 year-olds were higher in 2009 than in 2000. Third, children with special educational needs (SEN) had been integrated into mainstream schools since 2000.

While it would not be possible to pinpoint factors contributing to the score decline in Ireland in the 2009 PISA reading score, Cosgrove and Cartwright described the observed decline as the result of a "perfect storm". That is, multiple factors all happened at the same time to produce a large effect. In addition to the factors observed by Cosgrove and Cartwright, whatever reasons there may be for the observed general decline in English speaking and western/northern European countries (as shown in Fig. 1: PISA reading trends between 2000 and 2009) could also have contributed to a substantial decline for Ireland. Further, the PISA survey is sample based so there is always a chance element as to which schools are chosen, thus further contributing to this "perfect storm". Cosgrove and Cartwright summarized the lesson learned from the case study for Ireland as follows:

> This topic has relevance for other countries as well as for other large international assessments of education. PISA's potential use by policy-makers

> to monitor education systems and effect policy changes on the basis of the results implies that a good understanding of what the results mean is required for appropriate policy interventions, while misinterpretation or partial interpretation could result in misguided and erroneous interventions.
>
> (Cosgrove and Cartwright 2014, p. 2).

The key to this case study is that the 30-score-point decline should not be seen as a real decline in educational standards in Ireland. We have made an attempt to explain this decline through multiple factors unrelated to real trends in student performance.

Another example of misinterpretation of country rankings can be found in the Grattan report which stated that:

> Only 11 years ago, Hong Kong ranked 17th in assessments of reading literacy (PIRLS) and Singapore was ranked 15th. Just 5 years later (in 2006) they ranked 2nd and 4th.
>
> (Jensen et al. 2012, p. 2).

It should be noted in fact that Hong Kong was ranked 4th, 4th and 3rd in the 1995, 1999 and 2003 TIMSS mathematics respectively, while Singapore was ranked 1st in 1995, 1999 and 2003 in TIMSS mathematics. Singapore was also ranked 1st, 2nd and 1st in 1995, 1999 and 2003 TIMSS science. In PISA 2000, Hong Kong ranked 6th in reading, 1st in mathematics and 3rd in Science. Singapore did not participate in PISA 2000. From these outstanding performance results of Hong Kong and Singapore over the past two decades, any attempt to paint a picture of Hong Kong and Singapore as low performing countries in the year 2001 is difficult to credit. The low PIRLS 2001 results for Hong Kong and Singapore appear to be an anomaly rather than the norm for these two countries over the long history of international testing. Hong Kong did have an apparent improvement in its science scores in TIMSS, ranking 16th, 15th, 4th and 3rd in 1995, 1999, 2003 and 2007 TIMSS science assessments respectively. At a forum on large-scale assessments held at the National Taiwan Normal University on December 7, 2014, Dr Wong, Hong Kong’s TIMSS Science national research coordinator, attributed the improvement of Hong Kong’s science results to a major science curriculum overhaul over the testing period. The authors of the Grattan Institute report do not seem to have considered the performance data of Hong Kong and Singapore in PISA and IEA studies other than PIRLS, an inclusion which may have led to a somewhat different interpretation of the effect of policy implementations at the system level.

In fact, it is not unexpected that, if information is limited to what may be obtained at the system and school levels, government officials and school principals will be likely to claim credit for why their students perform well by pointing to the “successful” policies they have implemented. If researchers could examine wider evidence including cultural and family values, the conclusions might be quite different.

## Conclusion

In this paper, we have considered the ways in which think tanks and policy actors more broadly utilize data from surveys such as PISA. We have noted the ways in which data may be misunderstood or misinterpreted to support policy advice or actual policy formulations. We have shown how policies may be poorly supported by evidence derived from PISA which cannot be used for these purposes. This may be due to a number of technical data related issues that need to be considered before interpreting and using large scale assessment study survey data such as PISA for policy reform purposes. It may also be due to the ill-advised practice of making causal links between student achievement and success factors in large scale assessment survey data such as PISA and a related lack of caution in interpreting country rankings.

We would argue therefore that there is an urgent need to raise awareness among policy-makers that any trend data reported in international student assessments may or may not be related to the efficiency of the education system and that there may be many other important factors implicated, some of which may be difficult to measure or even identify. This is a challenge requiring integrated and coherent education policy research. Moreover, it should be noted that change in student performance at the country level is only likely to be visible in the long term. Short-term variations in achievement, which are often highlighted unfairly in media reports, would most likely be caused by factors unrelated to achievement such as those discussed in this paper. Given the lack of statistical literacy amongst much of the public, including some policy makers and the media, it is vital that high profile commentators, research and policy organisations (including the OECD) and government consider and adequately acknowledge such limitations when using PISA data for policy advice and for advocacy purposes.

## References

Alexander, R. J. (2010). World class schools—noble aspirations or globalised hokum? *Compare: A Journal of Comparative and International Education, 40*(6), 801–817.

Ashenden, D. (2013). Class sizes and the dead hand of history, inside story: Current affairs and culture from Australia and beyond, viewed 20th February 2015 http://insidestory.org.au/class-sizes-and-the-dead-hand-of-history/.

Ball, S. (1990). Self-doubt and soft data: social and technical trajectories in ethnographic fieldwork. *International Journal of Qualitative Studies in Education, 3*(2), 157–171.

Ball, S., & Exley, S. (2010). Making policy with 'good ideas': Policy networks and the 'intellectuals' of New Labour. *Journal of Education Policy, 25*(2), 151–169.

Bogdandy, A. V., & Goldmann, M. (2012). Taming and framing indicators: A legal reconstruction of the OECD's program for international student assessment (PISA). In K. E. Davis, A. Fisher, B. Kingsbury, & S. E. Merry (Eds.), *Governance by Indicators/Global Power through Quantification and Rankings*. Oxford: Oxford University Press.

Buckingham, J. (2012). Keeping PISA in perspective: Why australian education policy should not be driven by international test results, *Issue Analysis*, No. 136. Centre for Independent Studies, Sydney. Retrieved from 21st February http://www.cis.org.au/images/stories/issue-analysis/ia136.pdf.

Byrne, B., Coventry, W. L., Olson, R. K., Wadsworth, S. J., Samuelsson, S., Petrill, S. A., et al. (2009). Teacher effects in early literacy development: Evidence from a study of twins. *Journal of Educational Psychology, 102*(1), 32–42. doi:10.1037/a0017288.

Chen, C., & Stevenson, H. W. (1995). Motivation and mathematics achievement: A comparative study of Asian-American. *Caucasian-American and East Asian high school students, Child Development, 66*(4), 1214–1234.

Cosgrove, J., & Cartwright, F. (2014). Changes in achievement on PISA: The case of Ireland and implications for international assessment practice. *Large-scale Assessments in Education, 2014*(2), 2. doi:10.1186/2196-0739-2-2.

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

Figazzolo, L. (2009). Impact of PISA 2006 on the Education Policy Debate. Brusells: Education International, Retrieved from 21st February 2015 http://download.ei-ie.org/docs/IRISDocuments/Research%20Website%20Documents/2009-00036-01-E.pdf

Gorur, R., & Wu, M. (2014). Leaning too far? PISA, policy and Australia's 'top five' ambitions. *Discourse: Studies in the Cultural Politics of Education,*. doi:10.1080/01596306.2014.930020.

Goyette, K., & Xie, Y. (1999). Educational expectations of Asian American youths: Determinants and ethics differences. *Sociology of Education, 72*(1), 22–36.

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

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. In M. von Davier & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large scale assessments* (Vol. 2). Hamburg: IER Institute.

Hao, L., & Bonstead-Bruns, M. (1998). Parent-child differences in educational expectations and the academic achievement of immigrant and native students. *Sociology of Education, 71*(3), 175–198.

Hsin, A., & Xie, Y. (2014). Explaining Asian Americans' academic advantage over whites. The Proceedings of the National Academy of Sciences of the USA, June 10, 2014. *111*(23), 8416–8421. http://www.pnas.org/content/111/23/8416.full.pdf+html

Hwang, T. (2015). The studying and striving of secondary students, chapter 7. In S. Hsu & Y.-Y. Wu (Eds.), *Education as cultivation in Chinese culture, Education in the Asia-Pacific region: Issues, concerns and prospects*. Singapore: Springer Science + Business Media.

Jensen, B. (2010). *Investigating our teachers, investing in our money*. Melbourne: Grattan Institute. Retrieved from 21st February, 2015. http://grattan.edu.au/wp-content/uploads/2014/04/057_report_education_investing_teachers.pdf

Jensen, B. (2012). *Catching up: learning from the best school systems in East Asia: Summary Report*. Melbourne: Grattan Institute.

Jensen, B., Hunter, A., Sonnemann, J., & Burns, T. (2012). *Catching up: learning from the best school systems in East Asia*. Melbourne: Grattan Institute.

Lee, J. (2014). Minister apologies for 12 year compulsory education confusion, The China Post. Retrieved from 23rd July, 2015. http://www.chinapost.com.tw/print/409965.htm.

Leigh, A. (2010). Estimating teacher effectiveness from two-year changes in students' test scores. *Economics of Education Review, 29*(3), 480–488.

Liu, J. (2012). Does cram schooling matter? Who goes to cram schools? Evidence from Taiwan. *International Journal of Educational Development, 32*, 46–52.

Liu, J. (2013). An overview of student achievement and the related factors in Taiwan, Chapter 9.13. In J. Hattie & E. Anderman (Eds.), *International guide to student achievement*. New York: Routledge.

Ministry of Education. (2013). Education in Taiwan. Taipei: Ministry of Education, http://www.edu.tw

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement, 8*(3), 323–335.

OECD (2010a). *PISA 2009 results: Learning trends: Changes in student performance since* 2000 *(Volume V)*. Retrieved May 15, 2012, from http://dx.doi.org/10.1787/9789264091580-en

OECD. (2010b). *PISA 2009 results: What makes a school successful?—Resources, policies and practices (Volume IV)*. Retrieved May 15, 2012, from http://dx.doi.org/10.1787/9789264091559-en

Pyne, C. (2012). Aust obsessed with small class sizes: The Australian, 2nd August, Retrieved from 21st February 2015, http://www.theaustralian.com.au/news/latest-news/aust-obsessed-with-small-class-sizes-pyne/story-fn3dxiwe-1226428178883

Rizvi, F., & Lingard, B. (2010). *Globalizing education policy*. Abingdon: Routledge.

Robertson, R. (1995). Glocalization: Time-space and homogeneity-heterogeneity. In M. Featherstone & R. Robertson (Eds.), *Global modernities* (pp. 25–44). London: Sage.

The Economist. (2012). Lessons from East Asia: The classroom crush, Why class sizes in England may be sent to expand, Retrieved from 21st February 2015, http://www.economist.com/node/21547854

Thrupp, M. (2014). When PISA meets politics—a lesson from New Zealand. *The Conversation*, 20 May 2014. Viewed 21st February 2015, http://theconversation.com/when-pisa-meets-politics-a-lesson-from-new-zealand-26539

Tran, T. T. (2012). Is the learning approach of students from Confucian heritage culture problematic? *Education Research for Policy and Practice,*. doi:10.1007/s10671-012-9131-3.

Tuchman, B. W. (1984). *The march of folly: from Troy to Vietnam*. New York: Ballantine Books.

Urbach, D. (2013). An investigation of Australian OECD Pisa Trend results. In M. Prenzel, M. Kobarg, K. Schöps & S. Rönnebeck (Eds.), *Research on PISA: Research Outcomes of the PISA conference 2009*. Springer, doi: 10.1007/978-94-007-4458-5.

Watkins, T., Kirk, S., Small, V., & Levy, D. (2012). Backlash forces government class size U-turn. Report by Stuff.co.nz at http://www.stuff.co.nz/national/education/7059177/Backlash-forces-Government-class-size-U-turn

Wei, M., & Eisenhart, C. (2011). Why do Taiwanese children excel at math? *Kappan, 93*(1), 74–76.

Wu, M. (2009). A comparison of PISA and TIMSS 2003 achievement results in mathematics. *Prospect, 39*, 33–46. doi:10.1007/s11125-009-9109-y.

Wu, M. (2014). Evidence-based policy making in education. *International Journal of Contemporary Educational Research, 1*(1), 1–8.

Zyngier, D. (2014). Class size and academic results, with a focus on children from culturally, linguistically and economically disenfranchised communities. *Evidence Base, 1*, 1–23.

**Shelley Gillis** is an Associate Professor and Deputy Director of the Centre for Vocational and Educational Policy at the Melbourne Graduate School of Education in the Univeristy of Melbourne. She has successfully completed over 50 commissioned research studies focusing on educational assessment and reporting, quaification frameworks and quality assurance issues.

**John Polesel** is Professor and Associate Dean International in the Melbourne Graduate School of Education in the University of Melbourne. He has written over 100 journal articles, book chapters and commissioned reports, including articles in Oxford Review of Education, Comparative Education and Journal of Education Policy. His research focusses on youth transitions. He is currently leading a national study of the partnerships schools form to deliver vocational learning.

**Margaret Wu** has a background in statistics and psychometrics. She has been involved in large-scale assessments over the past 20 years. Margaret has taught item response modelling and quantitative methods at the University of Melbourne and Victoria University . She has co-authored two item response modelling software packages and published numerous journal articles and book chapters. Currently Margaret is involved in a number of international consultancies in the area of student assessment.