PISA and policy-borrowing: A philosophical perspective on their interplay in mathematics education

Cantley, I. (2019). PISA and policy-borrowing: A philosophical perspective on their interplay in mathematics education. *Educational Philosophy and Theory*, *51*(12), 1200-1215. https://doi.org/10.1080/00131857.2018.1523005

Published in:
Educational Philosophy and Theory

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

PISA and policy-borrowing: A philosophical perspective on their interplay in mathematics education

Ian Cantley

School of Social Sciences, Education and Social Work

Queen's University Belfast

ADDRESS: School of Social Sciences, Education and Social Work, Queen's University Belfast, 69-71 University Street, Belfast, United Kingdom, BT7 1HL

EMAIL: i.cantley@qub.ac.uk

TELEPHONE: +44 (0)2890 975936

ORCID: 0000-0002-8995-6281

Abstract

Mathematics achievement in different education systems around the world is assessed periodically in PISA, the Programme for International Student Assessment. PISA is deemed to yield robust international comparisons of mathematical attainment that enable individual countries and regions to monitor the performance of their education systems relative to standards being achieved internationally, with a view to informing their mathematics education policy decisions. Initially, the role of PISA in instigating mathematics education policy borrowing is outlined using England as a case study, and some existing technical critiques of PISA are then reviewed. Following this, aspects of Ludwig Wittgenstein's later philosophy of mind are used to reason that an over-reliance on the use of PISA to inform policy decisions in mathematics education may be problematic. It is suggested that, when PISA is viewed through a later Wittgensteinian lens, a potential deficiency in the underpinning psychometric model, pertaining to the inherent indeterminism in unmeasured mathematical abilities, may weaken PISA's utility in guiding mathematics education policy decisions. It is concluded that, whilst PISA mathematics scores may give some indication of the mathematical proficiency of a nation's students, caution is required before mathematics education policies are borrowed from other jurisdictions on the basis of PISA performance. Implications for the other PISA domains are also outlined.

Keywords

Mathematics education; policy-borrowing; PISA; Wittgenstein; robustness; item response theory

Introduction

Internationally, mathematics achievement has become an important metric for gauging educational success (Plank & Johnson, 2011), and the mathematical proficiency of a country's students has significant implications for its overall international competitiveness (OECD, 2014; Valero, 2017). Mathematics is often considered to more readily lend itself to international comparisons since it is less influenced by language and cultural differences than, for example, literacy or history (Heyneman & Lee, 2014), and mathematics therefore features prominently in international comparisons of educational achievement.

Currently, the mathematical skills of 15 year-old students around the globe are assessed by the triennial Programme for International Student Assessment (PISA), which was inaugurated by the Organisation for Economic Co-operation and Development (OECD) in 2000, and completed its most recent cycle in 2015. Unlike the other main international comparative assessment of mathematical achievement, the Trends in International Mathematics and Science Study (TIMSS), which adopts a curriculum approach to the assessment of mathematical knowledge, PISA focuses on assessing if students can apply mathematics to real-life problem-solving scenarios, and is thus deemed to be curriculum-independent (Labaree, 2014). PISA purports to yield robust comparisons of students' mathematical proficiencies (as well as their knowledge and skills in reading and scientific literacy), thereby providing policymakers in participating countries and jurisdictions with reliable international comparative data to monitor the mathematical achievement of their students, and potentially to make improvements to the relevant mathematics education policies. However, this paper problematizes the possibilities and limitations of using PISA findings as a basis for informing policy decisions on mathematics curricula and pedagogical approaches. It is argued that, when the psychometric model which underpins PISA is viewed through a later Wittgensteinian lens, some tensions emerge concerning PISA's robustness for measuring and comparing mathematical attainment in different international contexts. It is suggested that, whilst PISA outcomes may give some indication of the mathematical competence of a nation's students, the weaknesses associated with the measurement model upon which it is predicated may limit PISA's utility in mathematics education policy discourse.

High-performing countries in PISA often take great pride in their achievements and attribute them to the success of their mathematics education policies (Dossey & Wu, 2013). By contrast, in lower-performing countries, there are frequently widespread calls for reform

of mathematics education policies by both the public and the relevant governments. Politically, lower than desired performance has the potential to legitimise increased government involvement in re-orientating mathematics education policies (Dossey & Wu, 2013; Ertl, 2006; Takayama, 2008), although PISA results may also be used to validate existing policy proposals (Nortvedt, 2018). As Dossey and Wu (2013) noted, "the very existence of rankings provided by assessments such as the … OECD studies provide a perceived base of scientific rationality for policy proposals and their public explanations" (p. 1019). Even if it is assumed that the PISA measurement model has a high level of technical fidelity, basing policy decisions on PISA performance is problematic. For example, the education systems of participating countries are diverse and they advocate different approaches to mathematical pedagogy, the national mathematics curricula may not be closely aligned with the PISA framework for mathematical literacy, and approaches to student sampling are inconsistent, thus creating challenges in relation to drawing direct comparisons between PISA performance levels. However, policy decisions predicated on PISA outcomes would be potentially even more problematic if the underpinning psychometric model were found to have inherent weaknesses. This paper examines the technical fidelity of the PISA measurement model from a fundamental philosophical perspective, and explores the consequent implications for PISA-informed policymaking. Whilst the technical robustness of PISA has been questioned by numerous other researchers (e.g. Goldstein, 2004; Kreiner & Christensen, 2014), there is a dearth of fundamental theoretical critiques such as the one advanced in this paper. Irrespective of whether differences in the PISA performance of participating countries are attributable to the different mathematics curricula and pedagogical approaches employed by teachers, different sociocultural contexts, deficiencies in the underpinning psychometric model, or some combination of these factors, lower than desired achievement in PISA can have significant policy ramifications.

In lower-achieving countries, the "shock" that is generated by the publication of PISA results often leads to a reappraisal of mathematics education policies, and there is an increased tendency to 'borrow' policies from high-performing nations (Grek, 2009; Phillips & Ochs, 2003). In an attempt to improve a country's international standing, its mathematics education policymakers may borrow policies which were developed, and perceived to be effective, in other nations without due regard for the contexts in which the policies were initially operating (Nguyen, Elliott, Terlouw, & Pilot, 2009; Winstanley, 2012). Limited cognisance may be taken of the support structures that ensured the success of the policies, the

different cultural contexts in which they were situated, the effect of policy borrowing on the coherence of existing mathematics education provision (Clapham & Vickers, 2018), or the power of PISA scores to predict students' future mathematical behaviour. The propensity of some countries to borrow mathematics education policies from high-performing jurisdictions in PISA is exemplified by England's recent preoccupation with mathematical education in East Asian regions, such as Shanghai and Singapore. Accordingly, the following section outlines the recent focus of policymakers in England on East Asian policies and practices.

Policy-borrowing in mathematics education: The situation in England

In recent PISA cycles, a number of East Asian regions, such as Singapore, Shanghai and Hong Kong, have dominated the associated international league tables of mathematical performance. For example, in PISA 2015, Singapore's mean mathematics score was 564, while England's equivalent score was 493, and the differential between the mathematics scores of the two countries was consistent with those reported in the previous two PISA cycles, in 2009 and 2012. Similarly, in PISA 2012, Shanghai's mean mathematics score was 613, which was very significantly greater than England's corresponding score of 495. It is unsurprising, therefore, that East Asian mathematics education policies and practices have attracted considerable interest from education policymakers in England. The PISA mathematics performance of East Asian regions has been a significant factor in motivating English policymakers to 'borrow' mathematics education policies from these high performing jurisdictions (Clapham & Vickers, 2018).

In an attempt to identify factors contributing to the significant differentials in mathematics performance between East Asian jurisdictions and England, the mathematics curricula and pedagogical approaches employed in East Asia have come under scrutiny by English educationalists. For example, in 2012, the Department for Education conducted a review of the mathematics curricula in a number of high-performing jurisdictions, including some East Asian regions, to inform impending changes to the national curriculum for mathematics in England (Department for Education, 2012). Furthermore, a teacher exchange programme for teachers from England and Shanghai, linked to the Government's 'maths hubs' initiative (networks of schools tasked with promoting excellence in mathematics teaching), was launched in 2014 (Department for Education, 2014). The programme was designed to permit a number of mathematics teachers from England to learn about Shanghai

approaches to teaching mathematics while working in Shanghai schools for at least a month, and also to allow a number of English-speaking mathematics teachers from Shanghai to run master classes and provide training for teachers during their exchange visit to England. There has also been a high degree of fascination with East Asian mathematics textbooks, and English language versions of some of the books have been trialled in English schools (Department for Education, 2015).

A 'mastery' approach to teaching mathematics is employed by the top-performing Asian regions such as Singapore, Shanghai and Hong Kong. This entails careful planning and whole-class interactive teaching, which makes frequent use of concrete resources and visual representations of mathematical concepts, and the predominantly teacher-led instruction proceeds with the expectation that all students should achieve a certain level of mastery before the class progresses. Mastery teaching is purported to prioritise the development of conceptual understanding and procedural fluency to ensure that students are not simply following a set of "rules without reasons" (Skemp, 2006, p. 89), and are better positioned to apply their mathematical knowledge to problem-solving. The British Government's commitment to implementing the mastery approach to teaching mathematics in England is clearly demonstrated by the announcement, in July 2016, of £41 million funding to support its introduction in primary schools. In the press release that detailed the investment, the Schools Minister, Nick Gibb, commented:

> The significant expansion of the south Asian maths mastery approach can only add to the positive momentum, with thousands more young people having access to specialist teachers and quality textbooks. I am confident that the steps we are taking now will ensure young people are properly prepared for further study and the 21st century workplace, and that the too-often heard phrase 'can't do maths' is consigned to the past. (Department for Education, 2016)

Whilst the aims of mastery teaching are very laudable, its operationalisation in England conflicted with significant aspects of the prevailing sociocultural context. For example, primary mathematics is taught by subject specialists in East Asian regions but by generalists in England, and this is further exacerbated by significant differences in the structures of the school day and attitudes to using textbook-based exercises between England and East Asian jurisdictions (Clapham & Vickers, 2018). Furthermore, Clapham and Vickers (2018) suggested that the implementation of mastery teaching in England was compromised by differences between the learning cultures of England and East Asia, in terms of the

commitment and work ethic of students. Worryingly, Clapham and Vickers (2018) also noted concerns about weaker students being left behind in the English implementation of mastery teaching. This problem is further compounded by a lack of time within the school day for teachers in England to provide additional support for struggling students, and a lower emphasis on home tutoring as a form of remedial support in England compared to East Asian regions.

Given the centrality of PISA in influencing mathematics education policies in this way, it is imperative that PISA scores should accurately reflect students' mathematical competence. However, a number of concerns have been raised about the robustness of PISA, some of which are reviewed in the next section.

Existing critiques of PISA

PISA has been subject to extensive critique in the educational literature from a range of perspectives (e.g. De Lange, 2006; Dohn, 2007; Eivers, 2010; Goldstein, 2004; Kreiner & Christensen, 2014; Labaree, 2014; Prais, 2003), and some scholars have called for a more measured approach to the reporting and interpretation of PISA results that takes account of the assessment's inherent limitations (Rutkowski & Rutkowski, 2016). PISA's focus on assessing mathematical literacy rather than curricular knowledge and skills has attracted some criticism because of its neglect of the more formal aspects of mathematics that feature in some national curricula (de Lange, 2006). Indeed, Labaree (2014) argued that, to resolve the problem of how to assess student achievement across a diverse set of international education systems with different curricula, PISA measures a set of skills that are deemed to be economically useful rather than the extent to which students learn what they are actually taught in school: "Because they can't compare school systems based on what they teach, they invent a skill set that no one teaches and then use mastery of it as the measure of effective schools" (p. 5).

Some studies have found that students' responses to PISA mathematics test items may be significantly influenced by factors other than a lack of mathematical competence (Baucal, Pavlović Babić, & Jošić, 2018; Radišić & Baucal, 2018; Selleri & Carugati, 2018), such as a lack of familiarity with contextual questions of the type that feature in PISA. It has also been reported that substantial increases in PISA scores can occur without any real improvement in the quality of a country's education. For example, Pavlović Babić and Baucal (2011)

attributed a significant increase in Serbia's PISA reading scores between 2006 and 2009 to the national context within which the assessments were conducted rather than any significant improvement in the overall quality of educational provision. Relative to the 2006 Serbian participants in PISA, the 2009 cohort had been educated in more stable social conditions, free from civil unrest, which Pavlović Babić and Baucal (2011) suggest led to the enhanced performance in 2009. Worryingly, Sjøberg (2018) indicated that PISA science scores actually correlate negatively with pedagogical practices that are deemed to be highly desirable aspects of a good scientific education, such as inquiry-based learning, and he feared that schools may be forced to sacrifice enriching educational experiences in the politically-motivated drive to improve PISA performance.

A number of researchers have reported that the difficulty of PISA test items can be differentially influenced by language effects (Eivers, 2010; Roth, Erickan, Simon, & Fola, 2015). Given that language is a central factor when seeking to draw assessment comparisons between students who communicate in different mother tongues, doubt has been cast upon the possibility of constructing PISA questions in a manner that guarantees equivalent meanings across the range of participating countries. Although linguists may perceive test items written in different languages to be identical, students from different countries may actually interpret the items differently, thus leading to differences in item difficulties and compromising the validity and reliability of PISA scores. However, some studies have found that, whilst there may be differential language effects that favour particular groups of students for individual items on a PISA test, these effects balance out at the overall test level, resulting in fair and accurate comparisons (Oliveri, Olson, Ercikan, & Zumbo, 2012).

In principle, all participating countries can propose new test items for an upcoming PISA study, and new items are validated by teachers from different countries and piloted one year before the main study to ensure they function similarly in different countries (OECD, 2017). Although there may be items in the main PISA study that function differently in some countries, attempts are made to ensure the overall test is relatively fair (OECD, 2017), thus corroborating the findings reported by Oliveri et al. (2012). Whilst such rigorous piloting of PISA items is commendable, I suggest it also admits the possibility of critique in its own right, since striving for measurement comparability may compromise the relevance of the selected PISA items for the educational context of a particular nation(s). For example, methodological rigour in the selection of PISA items could potentially guarantee that all items perform similarly in different countries, but it does not guarantee that included items

assess the most relevant aspects of mathematical literacy in different countries. In other words, the PISA approach favours knowledge and skills that are relevant in different countries, but they may be the most relevant and important knowledge and skills in one particular country yet peripheral in others. In such a situation, although metrically PISA results enable us to compare numbers, the implications and relevancy of these numbers in two countries might be quite different. This is a dilemma that cannot be resolved by any measurement model.

The inclusion and exclusion criteria adopted in the PISA sampling frame have also been subject to criticism (Prais, 2003; Schuelka, 2013). For example, Prais (2003) suggested that PISA's use of a 12 month birth period for selecting participants is unsatisfactory since, due to issues such as students repeating or skipping a year of schooling in some countries, the resulting sample could include students with various levels of exposure to the curriculum relative to a typical 15 year-old student in a given country. However, it is important to highlight that use of a 12 month birth period for selecting the PISA sample rather than, for example, sampling whole classes, reduces the probability of the results being confounded by age effects. Furthermore, Schuelka (2013) cautioned that the deliberate exclusion of disabled students from PISA samples leads to further marginalisation and presents a barrier to the consideration of their interests in any policies relating to educational equity that emerge from the PISA enterprise. Interestingly, Tom Loveless (as cited in Dews, 2013) has claimed the high Shanghai PISA scores are almost meaningless since they only represent the performance of a single Chinese province with an atypically higher concentration of economically elite individuals than the rest of China. Loveless makes the point that many educationally disadvantaged children of migrant workers are excluded from participation in PISA by a policy that forces many of them to return to their parents' hometowns to complete their high school education, thus ensuring that only students from more affluent, educationally advantaged backgrounds are included in the PISA sample.

Dohn (2007) raised significant methodological concerns about the PISA project and fundamentally questioned its capacity to generate valid measures of absolute levels of knowledge. She argued that, although PISA test items may be constructed to take cognisance of the absolute levels of knowledge required to respond correctly to them, the resulting measurement scale is statistically constructed relative to the performance of those who take the PISA tests. Dohn therefore concluded that PISA's attempt to make inferences about absolute levels of knowledge on the basis of a relative evaluation is fundamentally flawed.

Technical critiques of PISA, such as the one presented later in the current paper, often refer to the psychometric model underpinning PISA, and it is therefore instructive to briefly outline the essential aspects of this model, which is based on item response theory (IRT). In IRT, it is posited that, although mathematical abilities cannot be measured directly, they are functionally related to the patterns of students' responses to test items that assess the abilities. A mathematical function, subsequently referred to as the item response function (IRF), is used to model the probability of a student giving a correct answer to a test item in terms of the student's mathematical ability and characteristics of the item, such as its difficulty. However, the properties of the test items are deemed to be totally independent of mathematical ability scores, and sophisticated mathematical techniques have been devised to estimate mathematical abilities from students' test responses (Raykov & Marcoulides, 2011).

Goldstein (2004) argued that insufficient attention has been given to the appropriateness of PISA's reliance on a one-parameter IRT model, also known as the Rasch model (Rasch, 1960), which he views as "conceptually simplistic" (Goldstein, 2004, p. 328). Similar concerns have been proffered by Kreiner and Christensen (2014) who, on the basis of their analysis, concluded that there is convincing evidence of model misfit, even when the Rasch model fit was tested separately for each country. Goldstein (2004) also questioned the logic of measuring such a multifarious construct as mathematical ability on a unidimensional scale of the type employed in PISA and he suggested that, whilst such an approach may yield appropriate measures for purposes such as student certification, it is unlikely to be appropriate in international comparative studies, where the incorporation of multiple dimensions would facilitate more accurate comparisons to be made between countries. However, it is noteworthy that PISA 2015 adopted a slightly more sophisticated IRT model than the one-parameter Rasch model that was used in previous cycles of the assessment (OECD, 2016, p. 306), which may alleviate some of the concerns raised by Goldstein (2004) and Kreiner and Christensen (2014). Despite this attempt to improve the technical fidelity of PISA, grave concerns re-emerge about the philosophical robustness of its underpinning measurement model when IRT is viewed through a later Wittgensteinian lens. The author draws on Wittgenstein's later philosophy of mind to argue that a potential deficiency in IRT, relating to the irreducible uncertainty in unmeasured mathematical abilities, may restrict its usefulness in guiding mathematics education policy decisions. As a precursor to this, the following section uses some pivotal ideas from Wittgenstein's later philosophy of psychology to argue that there is inherent indeterminism in unmeasured psychological predicates.

Indeterminism in unmeasured psychological predicates

Wittgenstein's 'private language argument' (PLA), in which he contrasts his view of thought and cognition with that of cognitive psychology, is one of the most widely-known aspects of his later philosophy of psychology. In cognitive psychology, cognitions, including attention, language use, memory, perception, problem-solving, creativity and thinking, are viewed as mental entities that provide causal explanations for other human behaviour or action. As such, mental entities and human actions are considered separate phenomena. In an apparent attempt to undermine this Cartesian dualism, Wittgenstein (2009) invites the reader to consider the plausibility of a 'private language':

> [A private language is] a language in which a person could write down or give voice to his inner experiences – his feelings, moods, and so on ... The words of this language are to refer to what only the speaker can know – to his immediate private sensations. So another person cannot understand the language. (Wittgenstein, 2009, §243)

The PLA has been extensively analysed in the secondary literature and it has been interpreted in various ways. Some scholars have, for example, suggested that the PLA is supposed to show that a private language is impossible (e.g. Malcolm, 1986), while others have construed the PLA as demonstrating that the idea of a private language is unintelligible (e.g. Hacker, 1986). In this section, however, I will give an account of the PLA which can be used to support my thesis that unmeasured mathematical abilities are ontologically indeterminate.

Wittgenstein's attack on the concept of a private language as something removed from the environment, commences in the *Philosophical Investigations* with an example of how a certain sensation is associated with a sign 'S':

> Let's imagine the following case. I want to keep a diary about the recurrence of a certain sensation. To this end I associate it with the sign 'S' and write this sign in a calendar for every day on which I have the sensation. – I first want to observe that a definition of the sign cannot be formulated. But all the same, I can give one to myself as a kind of ostensive definition! – How? Can I point to the sensation? – Not in the ordinary sense. But I speak, or write the sign down, and at the same time I concentrate my attention on the sensation – and so, as it were, point to it inwardly. – But what is this ceremony for? For that is all it seems to be! A definition serves to lay down the meaning of a sign, doesn't it? – Well, that is done precisely by concentrating my

> attention; for in this way I commit to memory the connection between the sign and the sensation. – But 'I commit it to memory' can only mean: this process brings it about that I remember the connection *correctly* in the future. But in the present case, I have no criterion of correctness. One would like to say: whatever is going to seem correct to me is correct. And that only means that here we can't talk about 'correct'. (Wittgenstein, 2009, §258)

Here, Wittgenstein demonstrates that the private language user, subsequently referred to as the 'private linguist', fails to attach a meaning to the symbol 'S'. The private linguist defines 'S' by connecting it with a sensation but, since the original sensation has disappeared when they attempt to use 'S' in future, there will be no standard of correctness against which the future use of 'S' can be assessed.

The above analysis (Wittgenstein, 2009, §258) leads to the logical conclusion that the symbol 'S' has no meaning since there is no mechanism for gauging if a future use of 'S' is correct. Wittgenstein suggested that the only solution to this dilemma is to establish public criteria for the use of 'S'. As McGinn (1997) noted, "it is not that the first-person use of 'S' must itself be guided by these public criteria, but there must be public criteria against which the subject's application of 'S' in a new case can be checked for correctness" (p. 129). Wittgenstein's comments on private language therefore highlight that the meaningfulness of any descriptive label for a psychological concept (such as a thought or sensation) is dependent upon the existence of public criteria for its application.

> It follows [from the private language argument] that any genuine (rule-governed) language must refer only to things and properties whose presence can be publicly verified: in particular, there must be public criteria for the presence of sensations if meaningful sensation words are to be possible. And in point of fact such criteria *do* figure in our actual acquisition of sensation language, since we use (e.g.) 'pain' precisely as a *replacement* for the kind of behaviour which provides others with a warrant for ascribing pain to us. (McGinn, 1984, pp. 48-49)

Similarly to the PLA itself, Wittgenstein's discussion of criteria for the ascription of psychological concepts, such as sensations and thoughts, has been extensively debated in the secondary literature (Witherspoon, 2011). According to Grayling (2001, p. 101), Wittgenstein viewed a criterion of correctness for ascribing a psychological concept to an individual as being "half-way between deductive and inductive grounds" for the ascription. Ascription on

deductive grounds would mandate that the ascription is made when those grounds are manifest. For example, pain would be ascribed to a person on purely behavioural grounds when he or she groans or winces. However, such a simplistic approach clearly fails to take account of the possibility of feigned pain behaviour, which exposes the weaknesses associated with ascribing psychological concepts on deductive grounds. On the other hand, the ascription of psychological concepts on inductive grounds would, for example, treat groaning or wincing as symptoms from which it may be inferred that an individual is in pain, where pain itself is viewed as a private, mental phenomenon. Grayling (2001) makes the point that, for Wittgenstein, recognising the role played by behaviours such as groaning or wincing in the ascription of pain, including the identification of pretence, is an integral aspect of understanding the meaning of 'pain', rather than offering either deductive or inductive grounds for the ascription. On this reading, the public behavioural criteria for ascribing psychological concepts to an individual are logically linked to, and non-separable from, the mental states for which they are criterial. However, this should not be construed as denying the existence of the mental processes and states. Rather, it simply means that, *prior* to their public manifestation, there is inherent indeterminism in psychological concepts, and that such concepts only become determinate with the appearance of public behaviours (which can form the basis of ascription criteria).

The reading of the PLA outlined above implies that psychological phenomena cannot be defined by introspection and, ultimately, that the concept of a private language is impossible. This interpretation of Wittgenstein's PLA will now be used to develop the notion that psychological concepts such as cognitive abilities adhere to a first-person/third-person asymmetry. It will be argued that first-person and third-person ascriptions of abilities are asymmetrical in terms of how they are confirmed since third-person ascriptions are based on public criteria, while first-person ascriptions cannot be predicated on public criteria. Furthermore, the case will be advanced that, since there are no criteria for first-person ascriptions of psychological predicates, cognitive abilities are necessarily indefinite before they are measured. This has significant ramifications for IRT, which postulates that psychological constructs such as abilities have definite values prior to measurement.

Although PISA focuses on mathematical literacy, a number of researchers (e.g. Lee & Chen, 2009) have demonstrated that mathematical content knowledge is very strongly correlated with mathematical problem-solving skills, and therefore the first-person/third-person asymmetry argument will, for the sake of simplicity, be developed using the

mathematical operation of addition. Consider the case of Michael who takes a test consisting, again for the sake of simplicity, of just two addition questions – q_1 and q_2 – each of which is entirely new to Michael. The test is designed to measure Michael's mathematical ability with respect to these two questions. Suppose that the teaching Michael received in arithmetic was highly successful so that Michael might proclaim with confidence that he has 'learned the concept 'add''. Suppose that Michael gives the correct answer for q_1. What can be said about Michael's ability (or his grasp of the meaning of 'add') after he has answered q_1 but before he answers q_2? In other words, what can be said about his ability as a thing-in-itself, when it is not being measured; when he is 'between questions'? Surely ability is something that is inner? Moreover, when thinking of Michael's ability as a thing-in-itself, it is difficult to escape the idea that Michael's responses to the questions are somehow inferior, since they are simply the *application* of his ability. It is tempting to suggest that, before he answers q_2, it is to his mind one should look in search of his ability even if, from his *first-person perspective*, Michael can find nothing in his mind that tells him how he should respond to q_2.

The reading of Wittgenstein's PLA discussed above implies that, prior to actually answering q_2, there are no criteria for judging the correctness or otherwise of a potential answer to the question, and any answer can be deemed to agree with the requirements of the rule for addition if a relevant interpretation of the rule is adopted. For example, *before* Michael responds to the question '12 + 13 = ?', all of the facts about him (both his history of responses to questions about addition and his inner, mental history) are consistent with both the answer '25' *and* the answer '1213' (for example). Before Michael answers q_2, the totality of facts dictate that he is both correct *and* incorrect with respect to that question. From his first-person perspective, Michael has no criterion for choosing among the multitude of possible responses. The terms 'correct' and 'incorrect' cease to have meaning, and Michael would have no grounds for deeming '25' the correct answer and '1213' an incorrect answer:

> This was our paradox: no course of action could be determined by a rule, because every course of action can be brought into accord with the rule. The answer was: if every course of action can be brought into accord with the rule, then it can also be brought into conflict with it. And so there would be neither accord nor conflict here. (Wittgenstein, 2009, §201)

Here, Wittgenstein is making the point that, prior to an answer being proffered, a rule can be interpreted in multiple ways, so that there are no criteria for a correct application of the rule.

He also implies that the problem is not resolved by attaching a further interpretation to the initial interpretation as this simply leads to an infinite regress:

> That there is a misunderstanding here is shown by the mere fact that in this chain of reasoning we place one interpretation behind another, as if each one contented us at least for a moment, until we thought of yet another lying behind it. (Wittgenstein, 2009, §201)

Wittgenstein (2009) stresses that "'following a rule' is a practice" (§202), so that the correct application of the rule is determined by a public practice or custom rather than a particular interpretation of the rule in the mental realm. Therefore, the moment Michael gives his answer to the addition problem, he is *either* correct *or* incorrect. Michael's teacher, from her *third-person perspective* can compare his answer with the consensus view of the community of mathematicians, and pronounce '25' correct and '1213' incorrect, and I suggest this is in keeping with Wittgenstein's contention that "there is a way of grasping a rule which is not an interpretation" (Wittgenstein, 2009, §201). It thus appears that statements such as 'can add 12 and 13' are ascribed by the student to himself (before he answers) *without criterion*, while the same statement is ascribed by the teacher to the student (after he answers) *with a criterion*. While it is meaningful to ask for evidence in respect of third-person ascriptions, it is not in respect of first-person ascriptions. Further support for this reading is offered by Wittgenstein's claim that "it's not possible to follow a rule 'privately'; otherwise, thinking one was following a rule would be the same thing as following it" (Wittgenstein, 2009, §202).

One is forced to conclude that 'ability' is a relational (rather than an innate) attribute. In other words, his response is Michael's relation to q_2 rather than an innate property of Michael. It is problematic to refer to Michael's ability as a thing-in-itself divorced from the measuring instrument (the practice of addition). Given that what can be said about his 'ability' changes radically with the environmental contingencies, in this case the question being asked, one cannot meaningfully speak of ability divorced from the instrument of measurement. Furthermore, the argument presented above implies that Michael's ability relative to a specific question is indefinite prior to answering it, but becomes definite when he gives his answer.

It is conceivable that a rebuttal of the hypothesised indeterminism in psychological predicates prior to their measurement might be attempted by appealing to neural processes in

the brain. Whilst such processes can be measured using advanced technology such as fMRI scans, the resulting measurements adhere to first-person/third-person symmetry, as opposed to the first-person/third-person asymmetry which Wittgenstein (2009) deemed to be an essential characteristic of the psychological domain. For physical phenomena, like neural processes in the brain, both first-person and third-person ascriptions would be governed by the use of criteria.

> The original metaphorical use of the word 'inner' reflects the realization that you and I stand on a different logical level in regard to what *I* think and feel. But the view that thoughts and feelings are brain-processes *abolishes* this logical difference. If this view were true, you and I would stand on the *same* level in regard to what I think and feel. In order to ascertain my thoughts and feelings you *and* I would equally have to rely on advanced technology and scientific theory. (Malcolm, 1986, p. 191)

This suggests that, if psychological predicates such as cognitive abilities were measured using fMRI scans, for example, the resulting measurements would pertain to totally different types of entities from our common conceptions of psychological phenomena. In short, entities that are subject to first-person/third-person asymmetry cannot be meaningfully replaced by entities that are symmetrical with respect to first and third-person ascriptions. However, a deep philosophical dilemma persists even if it is assumed that mental states are correlated with brain states, which is a very reasonable assumption. The above reading of Wittgenstein's PLA implies that it is impossible to reduce mental states to brain states since there are no criteria to identify subjective mental states. Therefore it would seem that, prior to Michael giving an answer to a question, the uncertainty in his mathematical ability is irreducible since it cannot even be reduced by measuring his brain states due to the inherent indeterminism in the associated mental states.

Some philosophers, e.g. Winch (2002), have suggested that Wittgenstein's intent in the PLA was to undermine representational theories of learning, such as those associated with cognitivism, in favour of a social conception of learning. Under a social model of learning, the meaning of a word such as 'add' is bound up with the non-linguistic actions of the practice of 'adding', which take place in the publicly available space. Learners are initiated into the practice of addition through the use of examples and training to equip them with the know-how to perform addition calculations. Such a reading of the PLA, obviates the need for problematic interpretations of mental representations, thus banishing the asymmetry between first- and third-person perspectives of learning since both perspectives are focused on acting

in agreement with the practice (or otherwise). However, despite its intuitive appeal, I contend that this reading of the PLA is somewhat at variance with Wittgenstein's original intention since he actually explicitly acknowledged the asymmetry between first- and third-person perspectives of psychological concepts: "Psychological verbs [are] characterized by the fact that the third person of the present is to be verified by observation, the first person not" (Wittgenstein, 1981, §472). I therefore suggest that my reading of the PLA is not incongruous with Wittgenstein's intent.

The next section examines the implications of the indeterminism in unmeasured psychological predicates, and cognitive abilities in particular, for the philosophical coherence of item response theory.

Philosophical incoherence of item response theory

The apparent indeterminism in psychological attributes prior to measurement suggests that the probabilities linked to measuring these attributes are necessarily objective as opposed to subjective. Subjective probabilities are used to describe an event where the associated uncertainty is reflective of an incomplete knowledge of factors influencing the event, such as the outcome of tossing a fair die. With complete knowledge of the die, and the variables that affect its motion, the result of a die-tossing experiment could, in principle, be predicted with certainty. However, ignorance of relevant information mandates the use of subjective probabilities to characterise the possible outcomes of the experiment. Objective probabilities, on the other hand, are associated with events where there is irreducible uncertainty that is not symptomatic of an incomplete knowledge of all relevant factors. The Wittgensteinian reasoning presented above seems to imply that there is irreducible uncertainty in the ascription of mental predicates, which does not emanate from human or measuring instrument restrictions. Even if the mental states of another person could be directly accessed, no further relevant information would be divulged.

The IRT literature appears to imply that the IRF models a subjective probability rather than an objective probability, and that ability levels have definite, albeit unknown, values prior to measurement. This is exemplified by Lord's claim that an objective probabilistic interpretation of the IRF is 'unsuitable':

> The trouble comes from an unsuitable interpretation of the practical meaning of the item response function … If we try to interpret $P_i(\theta_A)$ as the probability that a

particular examinee A [θ_A is the ability of examinee A] will answer a particular item i correctly, we are likely to reach absurd conclusions. (Lord, 1980, p. 228)

The hypothesized 'absurd' conclusions relate to the following two paradoxes which, according to Lord (1980), who was one of the founding fathers of IRT, manifest themselves if one attaches an objective probabilistic interpretation to an IRF (in the sense that abilities are ontologically indeterminate prior to measurement):

i. Two examinees who omit the same item would be of equal ability;

ii. Conflicting information regarding the abilities of two examinees, A and B, who respond to two test items, 1 and 2, that measure the same ability and have identical item response functions, in the case where A knows the answer to item 1 but does not know the answer to item 2, whereas B knows the answer to 2 but does not know the answer to 1. Consideration of the probability of a correct response to item 1 would appear to imply that A is of greater ability than B, whereas the situation with item 2 implies that A is of lower ability than B.

(Lord, 1980, pp. 227-228)

Lord (1980, p. 228) thus suggested that an objective probabilistic interpretation of the IRF should be eschewed in favour of a subjective probabilistic interpretation:

To obtain useful results, we may properly

1. Interpret $P_i(\theta_A)$ as the probability that a particular examinee A will give the right answer to a randomly chosen item whose parameters are a_i, b_i, and c_i.
2. Interpret $P_i(\theta_A)$ as the probability that a randomly chosen examinee at ability level θ_A will answer a particular item i correctly.
3. Make both of these interpretations simultaneously.

In other words, Lord (1980) was stressing that, to avoid 'absurd' conclusions when interpreting the practical meaning of the IRF for a particular item i, it is necessary to assume that an examinee A has a certain position on the ability scale prior to measurement of his/her ability. Under this interpretation, the IRF thus represents a subjective probability since the ability of the examinee is definite, albeit unknown, before a measurement of it is made. This is in stark contrast to an objective probabilistic interpretation of the IRF where, before measurement, not only would the examinee's position on the ability scale be unknown, but

his/her position on the scale would be irreducibly uncertain, only becoming certain at the point of measurement.

The same stance was also taken by Holland (1990), who rejected a 'stochastic subject' interpretation, which attributes the probabilities that feature in IRT to irreducible uncertainty in the individual examinee, in favour of a 'repeated sampling' model. In the repeated sampling interpretation, an ensemble of examinees of equal ability levels, but with unknown levels of other hidden attributes such as fatigue or motivation, is considered and the probability of selecting an examinee at random who responds correctly is then calculated. Probabilities in the stochastic subject interpretation are objective, while probabilities in the repeated sampling model are subjective (since they arise due to ignorance of the hidden attributes of examinees).

However, the reasoning presented previously suggests that it may be problematic to reject the construal of the IRF as modelling an objective probability in favour of a subjective probability interpretation. The Wittgensteinian analysis presented above implies that, prior to measurement, the ability level is ontologically indeterminate. The two paradoxes that Lord (1980) identified to emanate from an objective probabilistic interpretation of the IRF only arise if ability is deemed to have a definite value before a measurement of it is made. As argued above, a consideration of Wittgenstein's later philosophy of psychology implies that unmeasured abilities are indefinite (and not just unknown), which therefore renders an objective probabilistic interpretation of the IRF to be appropriate. This philosophical incoherence in IRT has potentially grave consequences for PISA, which is based on an item response model, and these implications are explored in the concluding section of the paper.

Conclusion: Implications for use of PISA outcomes in mathematics education policy decisions

PISA measurements of mathematical abilities are underpinned by IRT. It has been argued above that Wittgenstein's later philosophy of psychology implies that, prior to measurement, psychological constructs such as mathematical abilities are indefinite (in the sense that they do not have certain positions on the ability scale rather than just being unknown) and that they only become definite when they are actually measured. The uncertainty prior to measurement is irreducible since it cannot be removed by using more precise measuring instruments such as fMRI scans. A similar situation to this arises in the

physical sciences when, for example, the position of an electron is measured. Before measurement, the position of the electron is irreducibly uncertain and is actually described by a 'probability wave' (that contains information on potential measurement outcomes and their corresponding objective probabilities), which collapses to give a definite value of the electron's position at the instant measurement occurs (Cantley, 2015). Similarly, Cantley (2017) suggested that, prior to a student answering a mathematical question, for example, they are in a superposition of states since their answer is both correct *and* incorrect. There is irreducible uncertainty regarding which outcome will occur before a measurement is made of the student's ability relative to the question, which is only removed when an answer is finally proffered: the answer is either correct *or* incorrect (Cantley, 2017). However, it has also been argued that IRT is predicated on the assumption that abilities are definite in the sense that they have certain, albeit unknown, locations on the ability scale before they are measured. One of the architects of IRT, Frederic Lord, acknowledged that paradoxical problems arise if ability levels are not assumed to have definite values prior to measurement (Lord, 1980, pp. 227-228).

It is indisputable that IRT has evolved into an extremely sophisticated theory that utilises elegant mathematical models and advanced computational strategies. Alas, if one of the fundamental assumptions of IRT is questionable, this casts doubt on the validity of the mathematical ability scores that are calculated in PISA. There may be some form of relationship between the computed mathematical ability level of an individual student and their mathematical proficiency but, on the basis of the reasoning presented in this paper, it may not be so explicit as to support the OECD's claim that their published international league tables of mean mathematics scores are robust. The differences in relative positions of particular countries in the PISA league tables may, at least partially, be attributable to the dubious theoretical foundations of IRT that emerge when educational measurement is viewed through a (later) Wittgensteinian lens.

All tests purporting to measure mathematical ability are placed under tension by the Wittgensteinian critique presented in this paper, but those predicated on IRT are particularly vulnerable because of IRT's reliance on hypothetical test-independent mathematical ability constructs that are assumed to have definite values prior to measurement. Furthermore, the critique poses considerable challenges for the use of IRT as a tool for estimating individual abilities of any type, including those that feature in the other PISA domains.

Whilst the critique of IRT in this paper undermines the validity of PISA in its current cross-sectional guise, faith in the OECD assessment programme may be restored if it were remodelled to incorporate a longitudinal dimension. If the same students were followed over a number of years, it may be possible to establish predictive validity of the PISA mathematical ability measures in terms of their potential to accurately predict later mathematical behaviour, for example at the end of secondary education or in employment scenarios. Indeed, due to the difficulties associated with establishing causal inferences, cross-sectional studies of the type currently conducted by PISA do not generally permit the effect of different educational policies or practices on academic achievement to be studied (Hutchison & Schagen, 2007). On the other hand, a longitudinal study would facilitate an investigation of the extent to which individual students' PISA mathematics scores predict later mathematical performance. If PISA scores were shown to be strong predictors of subsequent mathematical outcomes (at the level of the whole PISA study rather than in selected countries), it would extirpate the potential concerns about the technical fidelity of PISA that emanate from the Wittgensteinian analysis presented in this paper. Furthermore, if PISA scores were confirmed to be technically robust using such an approach, a longitudinal study would permit a more rigorous investigation of the effects of different national education policies on student achievement. The apparent influence of different policies or practices on student achievement can change very significantly when cognisance is taken of prior achievement, which has a major bearing on subsequent performance. Given the prominent role of studies such as PISA in influencing the mathematics education policies and practices of individual nations, it is suggested that a reorientation of PISA to include a longitudinal aspect should be a priority, although the associated practical and logistical challenges are acknowledged.

Given the philosophical tensions surrounding the psychometric measurement model that underpins PISA, I suggest that mathematics education policy borrowing on the basis of favourable PISA rankings is problematic. Even if PISA scores were confirmed to accurately predict future mathematical behaviour, the nature of the relationship between mathematics education policies and practices and student achievement is unclear. The high levels of mathematical achievement of students from East Asian regions could well bear some relationship to the pedagogical approaches employed by their mathematics teachers, but some scholars have argued that the league-topping PISA performance of these economies is more intimately entwined with the importance attached to educational achievement in the

Confucian culture of East Asia (Jerrim, 2015; Leung, 2014). For example, Leung (2014) argued that the high achievement levels of East Asian students in international large-scale assessments such as PISA are attributable to factors such as parents' high expectations regarding their children's academic achievement, the extrinsic motivation for studying that is provided by public examinations, the genuine belief in effort as a vehicle for self-improvement, and the priority given to practice and memorisation. Against this backdrop, and in view of the critique of the PISA measurement model presented in this paper, it is problematic that education policymakers in England appear to have been largely motivated by the strong PISA performance of East Asian jurisdictions to wholeheartedly, and uncritically, embrace their approaches to mathematical pedagogy. As Winstanley (2012, p. 528) noted, "evidence is becoming especially difficult to interpret as global data are lifted out of context and presented as comparative numbers and sound bites". It is even more alarming if those comparative data are not reflective of philosophically robust measurement practices. Although blind educational policy borrowing of any type is extremely problematic (even if motivated by PISA results that are confirmed to be technically robust), policy borrowing could possibly be condoned if care were taken to ensure comparable sociocultural contexts in both the 'borrowing' and 'lending' nations.

Therefore, to avoid the potential misdirection of education policy, and valuable public resources, I suggest policymakers should predicate their decision to borrow educational policies from other nations on the basis of robust evidence, derived from relevant studies, of their capacity to make a *very significant* positive impact on students' learning rather than dazzling PISA rankings. Whilst England's borrowing of mathematics education policies from East Asia has been used as a case study in this paper to exemplify the tensions that can emerge as a consequence of careless policy borrowing, fuelled by a desire to improve PISA outcomes, the argument I have advanced is generalizable to all of the PISA domains in all countries that participate in PISA. In summary, extreme caution is required before educational policies are borrowed from other jurisdictions simply on the basis of their performance in PISA, particularly if the longer-term effects of such policy borrowing have not been investigated.

References

Baucal, A., Pavlović Babić, D., & Jošić, S. (2018). Dialogical PISA: correct answers are all alike, every incorrect answer is incorrect in its own way. *European Journal of Psychology of Education*, *33*(3), 467-487.

Cantley, I. (2015). How secure is a Newtonian paradigm for psychological and educational measurement? *Theory & Psychology*, *25*(1), 117-138.

Cantley, I. (2017). A quantum measurement paradigm for educational predicates: Implications for validity in educational measurement. *Educational Philosophy and Theory*, *49*(4), 405-421.

Clapham, A., & Vickers, R. (2018). Neither a borrower nor a lender be: exploring 'teaching for mastery' policy borrowing. *Oxford Review of Education*. Advance online publication. DOI: 10.1080/03054985.2018.1450745

De Lange, J. (2006). Mathematical literacy for living from OECD-PISA perspective. *Tsukuba Journal of Educational Study in Mathematics*, *25*, 13-35.

Department for Education. (2012). Review of the national curriculum in England: What can we learn from the English, mathematics and science curricula of high-performing jurisdictions? *Department for Education Research Report DFE-RR178*. London: Author. Retrieved on 28/12/2017 from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/184064/DFE-RR178.pdf

Department for Education. (2014). Network of 32 maths hubs across England aims to raise standards. *Department for Education press release*. London: Author. Retrieved on 28/12/2017 from https://www.gov.uk/government/news/network-of-32-maths-hubs-across-england-aims-to-raise-standards

Department for Education. (2015). How to get more high-quality textbooks into classrooms. *Department for Education speech*. London: Author. Retrieved on 28/12/2017 from https://www.gov.uk/government/speeches/how-to-get-more-high-quality-textbooks-into-classrooms

Department for Education. (2016). South Asian method of teaching maths to be rolled out in schools. *Department for Education press release*. London: Author. Retrieved on

28/12/2017 from https://www.gov.uk/government/news/south-asian-method-of-teaching-maths-to-be-rolled-out-in-schools

Dews, F. (2013, December 3). Tom Loveless: Shanghai PISA test scores almost meaningless; hukou a factor [Web log post]. Retrieved on 28/12/2017 from https://www.brookings.edu/blog/brookings-now/2013/12/03/tom-loveless-shanghai-pisa-test-scores-almost-meaningless-hukou-a-factor/

Dohn, N. B. (2007). Knowledge and skills for PISA – assessing the assessment. *Journal of Philosophy of Education*, *41*(1), 1-16.

Dossey, J. A., & Wu, M. L. (2013). Implications of international studies for national and local policy in mathematics education. In M. A. (Ken) Clements, A. Bishop, C. Keitel-Kreidt, J. Kilpatrick, & F. K. S. Leung (Eds.), *Third international handbook of mathematics education* (pp. 1009-1042). New York, NY: Springer.

Eivers, E. (2010). PISA: Issues in implementation and interpretation. *Irish Journal of Education / Iris Eireannach an Oideachais*, *38*, 94–118.

Ertl, H. (2006). Educational standards and the changing discourse on education: the reception and consequences of the PISA study in Germany. *Oxford Review of Education*, *32*(5), 619-634.

Goldstein, H. (2004). International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 319-330.

Grayling, A. C. (2001). *Wittgenstein: A very short introduction*. Oxford: Oxford University Press.

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

Hacker, P. M. S. (1986). *Insight and illusion: Themes in the philosophy of Wittgenstein* (revised edition). Oxford: Oxford University Press.

Heyneman, S. P., & Lee, B. (2014). The impact of international studies of academic achievement on policy and research. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 37-72). Boca Raton, FL: CRC Press.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*(4), 577-601.

Hutchison, D., & Schagen, I. (2007). Comparisons between PISA and TIMSS – Are we the man with two watches? In T. Loveless (Ed.), *Lessons learned: What international assessments tell us about math achievement* (pp. 227-261). Washington, DC: Brookings Institute Press.

Jerrim, J. (2015). Why do East Asian children perform so well in PISA? An investigation of Western-born children of East Asian descent. *Oxford Review of Education*, *41*(3), 310-333.

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.

Labaree, D. F. (2014). Let's measure what no one teaches: PISA, NCLB, and the shrinking aims of education. *Teachers College Record*, *116*(9), 1-14.

Lee, C.-Y., & Chen, M.-P. (2009). A computer game as a context for non-routine mathematical problem solving: The effects of type of question prompt and level of prior knowledge. *Computers & Education*, *52*(3), 530-542.

Leung, F. K. S. (2014). What can and should we learn from international studies of mathematics achievement? *Mathematics Education Research Journal*, *26*(3), 579-605.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates.

Malcolm, N. (1986). *Wittgenstein: Nothing is hidden*. Oxford: Blackwell.

McGinn, C. (1984). *Wittgenstein on meaning*. Oxford: Blackwell.

McGinn, M. (1997). *Routledge philosophy guidebook to Wittgenstein and the Philosophical Investigations*. Abingdon: Routledge.

Nguyen, P.-M., Elliott, J. G., Terlouw, C., & Pilot, A. (2009). Neocolonialism in education: Cooperative learning in an Asian context. *Comparative Education*, *45*(1), 109-130.

Nortvedt, G. A. (2018). Policy impact of PISA on mathematics education: the case of Norway. *European Journal of Psychology of Education*, *33*(3), 427-444.

OECD (2014). *PISA 2012 results – What students know and can do: Student performance in mathematics, reading and science (volume I, revised edition, February 2014)*. Paris: Organisation for Economic Cooperation and Development.

OECD (2016). *PISA 2015 results (volume I): Excellence and equity in education.* Paris: Organisation for Economic Cooperation and Development.

OECD (2017). *PISA 2015 technical report.* Paris: Organisation for Economic Cooperation and Development.

Oliveri, M. E., Olson, B. F., Ercikan, K., & Zumbo, B. D. (2012). Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing*, *12*(3), 203-223.

Pavlović Babić, D., & Baucal, A. (2011). The big improvement in PISA 2009 reading achievements in Serbia: Improvement of the quality of education or something else? *Center for Educational Policy Studies Journal*, *1*(3), 53-74.

Phillips, D., & Ochs, K. (2003). Processes of policy borrowing in education: Some explanatory and analytic devices. *Comparative Education*, *39*(4), 451-464.

Plank, D. N., & Johnson, B. L. (2011). Curriculum policy and educational productivity. In D. E. Mitchell, R. L. Crowson, & D. Shipps (Eds.), *Shaping education policy: Power and process* (pp. 167-188). New York, NY: Routledge.

Prais, S. J. (2003). Cautions on OECD's recent educational survey (PISA). *Oxford Review of Education*, *29*(2), 139-163.

Radišić, J., & Baucal, A. (2018). Teachers' reflection on PISA items and why they are so hard for students in Serbia. *European Journal of Psychology of Education*, *33*(3), 445-466.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Raykov, T., & Marcoulides, G.A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge.

Roth, W., Ercikan, K., Simon, M., & Fola, R. (2015). The assessment of mathematical literacy of linguistic minority students: Results of a multi-method investigation. *The Journal of Mathematical Behavior*, *40*, 88-105.

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, *45*(4), 252-257.

Schuelka, M. J. (2013). Excluding students with disabilities from the culture of achievement: The case of the TIMSS, PIRLS, and PISA. *Journal of Education Policy*, *28*(2), 216-230.

Selleri, P., & Carugati, F. (2018). Errare humanum est! A socio-psychological approach to a "Climbing Mount Fuji" PISA question. *European Journal of Psychology of Education*, *33*(3), 489-504.

Sjøberg, S. (2018). The power and paradoxes of PISA: Should inquiry-based science education be sacrificed to climb on the rankings? *NorDiNa*, *14*(2), 186-202.

Skemp, R. R. (2006). Relational understanding and instrumental understanding. *Mathematics Teaching in the Middle School*, *12*(2), 88-95.

Takayama, K. (2008). The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education*, *44*(4), 387-407.

Valero, P. (2017). Mathematics for all, economic growth, and the making of the citizen-worker. In T. S. Popkewitz, J. Diaz, & C. Kirchgasler (Eds.), *Political sociology and transnational educational studies: The styles of reason governing teaching, curriculum and teacher education* (pp. 117-132). New York, NY: Routledge.

Winch, C. (2002). *The philosophy of human learning*. London: Routledge.

Winstanley, C. (2012). Alluring ideas: Cherry picking policy from around the world. *Journal of Philosophy of Education*, *46*(4), 516-531.

Witherspoon, E. (2011). Wittgenstein on criteria and the problem of other minds. In O. Kuusela, & M. McGinn (Eds.), *The Oxford handbook of Wittgenstein* (pp. 472-498). Oxford: Oxford University Press.

Wittgenstein, L. (1981). *Zettel* (Edited by G. E. M. Anscombe & G. H. von Wright; translated by G. E. M. Anscombe; 2nd edition). Oxford: Blackwell.

Wittgenstein, L. (2009). *Philosophical investigations: The German text with an English translation* (Edited by P. M. S. Hacker & J. Schulte; translated by G. E. M. Anscombe, P. M. S. Hacker & J. Schulte; 4th edition). Chichester: Wiley-Blackwell.