International Journal of Educational Development 73 (2020) 102149

Contents lists available at ScienceDirect

International Journal of Educational Development

journal homepage: www.elsevier.com/locate/ijedudev

Human Capital Index and the hidden penalty for non-participation in ILSAs

Ji Liu[a,*], Gita Steiner-Khamsi[b]

[a] School of Education, Shaanxi Normal University, PR China
[b] Teachers College, Columbia University, USA

ARTICLE INFO

Keywords:
Human Capital Index
International Large-Scale Assessments
Education policy
Development
Comparative education
Testing and accountability

ABSTRACT

The World Bank's Human Capital Index (HCI) aims to provide new information regarding future productivity of each country's workforce, by synchronizing available International Large-Scale Assessment (ILSA) and regional test program results. Linking the literature on ILSA participation, this study questions the problematic nature of this approach and revisits the comparability issue of ILSA results. We find that education systems are imposed upon a score penalty depending on which ILSA or regional test program they choose to partake in. In particular, our results show that (i) test-overlap systems used in the score synchronization procedure are systematically different from systems that only choose to participate in one ILSA or exclusively in regional tests, (ii) inter-test score exchange rate is volatile due to sampling design and cohort effects, (iii) test participation type alone accounts for about 57.8 percent of the variation in synchronized scores, and the score penalty is especially salient for systems that exclusively participate in regional test programs; the majority of which are low-income and lower-middle income countries. Findings in this study show how various intra- and extrapolations to compensate for missing data in effect introduce large score penalties for systems that either did not participate or only partially participated in ILSAs. Finally, this study contributes to research on reasons for participation in ILSAs and the global rise of test-based accountability reform, under which the World Bank's new HCI may be seen as a tool to incentivize participation in ILSAs by penalizing those governments that have chosen alternative, non-standardized paths for measuring learning outcomes of students.

1. Introduction

Economists have long been interested in studying human capital, and this enthusiasm was propelled forward in part by the theoretical foundations laid out by Schultz (1961) and Becker (1962), and coincided with the expansion of the global development project since World War II and more recent donor interests in education interventions (Chabbott, 2007; Heyneman, 2003; Shields and Menashy, 2019; Steiner-Khamsi, 2006). To date, there have been various attempts to conceptualize and quantify human capital, for instance, United Nations Development Programme (UNDP) has been routinely measuring and releasing the Human Development Index since early 1990s, whereas Barro and Lee (1993) are among the first academics to construct a cross-nationally comparable educational attainment dataset using census and household survey. Until only recently, the focus on human capital measurement has been primarily confined to education quantity measurements that rely on enrollment or attainment data. Notwithstanding, the global growth of International Large-Scale Assessments (ILSAs) and regional student testing programs has prompted recent attempts to examine student learning outcomes more closely (see Barro and Lee, 2001; Cohen and Soto, 2007; Hanushek and Kimko, 2000; Hanushek and Woessmann, 2012; Heyneman and Lee, 2012).

As part of its global call-to-action to invest more in people, the World Bank released its first Human Capital Index (HCI) results in October 2018, ambitiously covering 157 economies and close to 98 percent of the world's population (World Bank, 2018a). While the index captures public health information such as infant and adult mortality, its main novelty according to the World Bank (2018b, p. 4), lies in its ability to provide "direct measure of school quality and human capital." To this end, the World Bank (2018b) envisions the index's education component as a new macro indicator to help countries keep track of progress in education quality or the lack thereof. As indicated in several key publications, the World Bank (2018a, 2018b, 2018c, 2018d, 2019a) plans for the index to have far-reaching global impacts, and indicated its future priority in setting HCI as a central metric in education planning, public financing, and government accountability. More specifically, the Bank plans on utilizing HCI to strategically promote new measurement initiatives, inform lending programs, and address shortcomings in political incentives (World Bank, 2018a, p.10).

While the HCI holds a grand promise, there are many intricacies

* Corresponding author at: Shaanxi Normal University, 199 Chang'An South Ave., Tin Ka Ping Education Building, Xian City, Shaanxi Province, 710062, PR China.
E-mail addresses: jiliu@snnu.edu.cn (J. Liu), gs174@tc.columbia.edu (G. Steiner-Khamsi).

https://doi.org/10.1016/j.ijedudev.2019.102149
Received 21 July 2019; Received in revised form 28 October 2019; Accepted 16 December 2019

worth examining in detail. Importantly, the key education component in the HCI – Harmonized Learning Outcomes (HLO) – is extrapolated from various ILSAs and regional test programs by relying on an inter-test exchange rate that is calculated using test-overlap systems for given pairs of test programs. However, differences in test design, timing, sampling, non-school factors can substantially limit the usefulness of such an exercise. To illustrate, the World Bank (2018b) imputes HLO scores for Lesotho, which did not participate in any ILSA programs, using an inter-test exchange rate that is calculated with Botswana's scores. However, Lesotho and Botswana are far from similar; for instance, the per capita Gross Domestic Product (GDP, 2017 PPP) amount in Botswana is $17,024, or 5.5 times more than that of Lesotho ($2932) (World Bank, 2019b). The obvious leap of faith is that this inter-test exchange rate is country-invariant and transcends borders and contextual factors such as demographics, income, culture, etc. As such, there are several important implications that arise with centering ILSA score harmonization at the core of the new Human Capital Index. Firstly, the reliance on synchronizing ILSA results to produce an HLO score implies that countries who choose not to participate in any ILSA or regional test programs will not receive a score. It is conceivable that a score penalty, which can substantively limit future policy, financing, and programming prospects, will act as policy levers in pushing countries towards reconsidering ILSA non-participation decisions. Secondly, because the core component of HCI involves test result conversion based on test-overlap systems, therefore score conversion rates could be biased upwards, neutral, or downwards depending on which ILSAs or regional test program and which year a system chooses to participate in. As the expansive literature on ILSA participation and comparability has shown (see Addey et al., 2017; Addey and Sellar, 2019), the political, economic, and social costs often can both inhibit or incentivize participation depending on country context, which further amplifies the comparability issue inherent to ILSAs through inter-test score synchronization exercises.

More specifically in this study, we examine the 'penalty' complex in further detail, and address the following research questions: (i) How reliable is the intra-test exchange rate, which depends on the comparability of test-overlap systems? (ii) Does participation in different test programs matter for Harmonized Learning Outcome (HLO) results, and how much is the associated penalty? In this investigation, we argue that the introduction of HCI as a new education quality tracking device, centering on synchronizing ILSA scores, should be critically examined. In particular, the algorithmic procedures devised to impute the index adds a new dimension of complexity – penalties – to the widely documented 'reasons for ILSA engagement' literature, in that systems are penalized for both non-participation and partial-participation in ILSA benchmarking. While existing studies have explored in depth the demand-side reasons why countries choose to participate in ILSAs, our current study contributes to understanding new supply-side intricacies that are introduced through HCI calculations. By examining the non-participation and partial-participation effect, we show how extrapolations to compensate for missing data in effect introduces considerable score penalties for countries that either did not participate or only partially participated in ILSAs.

2. Coercive rationales for ILSA engagement: a focus on the supply-side

The spectacular growth of education systems participating in ILSAs is noticeable and deserves theorizing. In this list, some of the most influential ILSAs include *Programme for International Student Assessment* (PISA), *International Reading Literacy Study* (PIRLS), *Trends in International Mathematics and Science Study* (TIMSS), and regional assessments such as *Latin American Laboratory for Assessment of Quality in Education* (LLECE), *Southern and Eastern Africa Consortium for Monitoring Educational Quality* (SAQMEC), *Early Grade Reading Assessment* (EGRA), and so on (Fischman et al., 2019). In terms of PISA alone, 43 countries participated in 2000, 72 economies in 2015, and 80 economies in 2018. In analyzing the explosive growth of PISA and other ILSAs, researchers have proposed several explanations, ranging from broad ones, such as globalization and political pressure to be part of a larger international educational space, to very concrete ones, such as an ever-increasing number of evidence-driven policy actors who rely on international comparison for measuring the quality of their educational system (see Addey et al., 2017; Engel, 2015; Liu, 2019; Verger et al., 2019). It is noticeable that this body of research has focused on national governments and explored why they buy into ILSAs, that is, explain the rationales for their participation.

The attractiveness of ILSAs is in part related to the quantification of learning outcomes. Wendy Espeland (2015) and Radhika Gorur (2016) masterfully observe the advantages of numbers over complex narratives because one may attach one's own narratives to numbers. What is especially appealing to policy actors are *Organization for Economic Co-operation and Development* (OECD) and *International Association for the Evaluation of Educational Achievement* (IEA) studies, is the capacity to generate statistics, scores, ranking, and benchmarks that are based on international comparison or on comparison over time. Espeland (2015, p. 56) explains the dual process of *simplification* and *elaboration* involved in the usage of numbers. In a first step, numbers "erase narratives" by systematically removing the persons, institutions, or systems being evaluated by the indicator and the researcher doing the evaluation. This technology of simplification stimulates narratives, or as Espeland astutely observes:

> If the main job of indicators is to classify, reduce, simplify, to make visible certain kinds of knowledge, indicators are also generative in ways we sometimes ignore: the evoke narratives, stories about what the indicators mean, what are their virtues or limitations, who should use them to what effect, their promises and their failings. (Espeland, 2015, p. 65)

A few scholars have focused on the "narrative evoking" phase (Espeland, 2015, p. 65) of such studies and dissected what national governments interpret or project onto OECD reports or other international comparative studies based on their own policy context and agenda (Waldow and Steiner-Khamsi, 2019). Studies on the rationales for participation in ILSAs have typically attempted to address the following contradiction: a great number of governments chooses to continuously participate in ILSAs even though they have already learnt after the first round of participation about the strengths and weaknesses of their systems. Arguably, the focus on rationales for participation has generated two blind spots: (i) rationales for non-participation and (ii) strategies to ensure continuous participation. The first research area focuses on governments (i.e. Botswana, South Africa, Kyrgyzstan, Switzerland) that discontinued, or considered a discontinuation (i.e. China, Mexico, Vietnam), respectively, the participation in an ILSA and the second research area focuses on international organizations that develop, administer, and analyze ILSAs. For the second under-explored research area, OECD's strategies to ensure continued interest is worth mentioning briefly.

In regards to the first area mentioned above, more research is needed on *rationales for non-participation or discontinuation*, respectively. For instance, Kyrgyzstan discontinued its participation for the PISA 2012 round after scoring at the very bottom in PISA 2006 (out of 47 countries and economies) and PISA 2009 (out of 65 countries and economies). In disbelief over the poor results, the Ministry of Education and Science established its own national assessment center. Notwithstanding, not all "ILSA drop-outs" are education systems that score at the bottom of a league table (Addey and Sellar, 2019; Lockheed and Wagemaker, 2013). For instance, South Africa discontinued TIMSS for resource reasons (Wiseman, 2013) and Mongolia dropped out of TIMSS due to technical capacity constraints (Addey, 2015). League leaders also have good reasons to discontinue their participation in ILSAs. For example, Switzerland only participated in two rounds of

TIMSS (Trends in International Mathematics and Science Study): in 1995 and 1999. In contrast, the country participated in all rounds of OECD's PISA, starting in 2000. The Swiss education authorities found it sufficient to participate in PISA rather than in PISA *and* TIMSS given that the two tests assess students' learning outcomes for the same subjects, albeit in different grades or age groups, respectively. The discontinuation of Swiss participation in TIMSS is simply a reflection of their disinterest and not, as sometimes assumed with systems that slip or rank below expectations, an indication of political anxiety about potentially harmful results. In fact, Switzerland scores were consistently above OECD average in mathematics and science since 2006. The country ranks first in mathematics among European countries.

As for the second under-explored area, the exclusive preoccupation with demand-side rationales has distracted attention from the *supply-side*. What do OECD, IEA and other international organizations undertake to keep governments' interest in participation alive? An analysis of strategies, issued by the governing boards of the organizations themselves, would be necessary to address the question (e.g., OECD, 2019). For example, for the PISA test, the OECD has created incentives to participate by adding a third innovative domain onto the test. As such, generating novelty is one of the mechanisms to ensure continued engagement. Another strategy was to expand the scope of participating countries by first allowing non-OECD countries ("partner countries") and then "economies" or municipalities (e.g. Baku-Azerbaijan, Miranda-Venezuela, Shanghai-China) to participate. Similarly, in TIMSS, participation has expanded from traditional IEA partner countries to include "benchmarking participants" (eg. Buenos Aires-Argentina, Dubai-UAE, Quebec-Canada). An important effect of the boom of ILSAs is, as Gorur (2016) has shown, is that governments start to "see like PISA," that is, internalizing the logic of ILSAs by attempting to make learning measurable, calculable, comparable, and accountable. In the same vein, Verger et al. (2019) have documented the global spread of test-based accountability reforms under the influence of ILSA participation. They show that the number of countries participating in ILSAs and the number of tests administered to students in a country have increased exponentially in the new millennium. Periodically, the global testing industry has come under siege for developing tests while simultaneously selling textbooks and teacher training materials in preparation for these tests (Hogan et al., 2016). Unsurprisingly, the proliferation of ILSA testing and the global commercialization of test-based accountability has become an object of intense academic curiosity and inquiry (Verger and Parcerisa, 2017; Lewis, 2017). Some have shown how the global education industry has succeeded in extending its reach by brokering education policies that require periodical testing of students, either through international benchmarking exercises or utilizing standards-based curriculum and accountability reforms. Once the demand has been created and governments start to "see like PISA," global actors sell their tests for an ever-increasing number of subjects, grade levels, and educational systems. In addition to PISA, we now have PISA-D (PISA for Development), PISA for Schools, and a proposed "Baby PISA" (International Early Learning and Child Well-Being study), that are administered all three years worldwide.

As mentioned above, the positive incentives that the OECD and the global testing industry have created to ensure continued participation are currently under-explored. Even less examined, however, are the less visible coercive strategies used to reinforce participation in ILSAs, particularly through generating penalties for non-participation. This current study contributes to the second, supply-side, area of research. Specifically, a considerable body of evidence from our analyses show that World Bank's HCI method of intra- and extrapolations in effect penalizes governments that participate in one test but not in both, or participates exclusively in regional tests.

Table 1
Human Capital Index and its Education Component, Select Countries and Economies.
Source: (World Bank, 2018a, d). Author's compilation.

Countries/ Economies	Human Capital Index	Harmonized Learning Outcomes	Learning-Adjusted Years of School (1)	Expected Years of School (2)	Difference: (1) - (2)
Singapore	0.88	580.9	12.9	13.9	−1.0
South Korea	0.84	563.1	12.2	13.6	−1.3
Japan	0.84	563.4	12.3	13.6	−1.3
Hong Kong	0.82	561.9	12.1	13.4	−1.4
Macao	0.76	545.3	11.0	12.6	−1.6
…	…	…	…	…	…
Dominican Republic	0.49	350.1	6.3	11.3	−5.0
Haiti	0.45	345.4	6.3	11.4	−5.1
Kosovo	0.56	374.8	7.7	12.8	−5.1
Guyana	0.49	346.4	6.7	12.1	−5.4
Ghana	0.44	307.3	5.7	11.6	−5.9

3. Human Capital Index and the long-standing ILSA comparability conundrum

In devising the Human Capital Index (HCI), the World Bank creates two new indicators: *Harmonized Learning Outcome* (HLO) and *Learning Adjusted Years of School* (LAYS). Firstly, the Bank synchronizes student achievement test results drawn from various international and regional test programs by producing a 'learning exchange rate' based on test-overlap systems (World Bank, 2018b, p. 8). Using this approach, the HLO scale maps all available country scores on ILSAs and regional test programs to a TIMSS-equivalent system ranging 300–625 points. Secondly, using a given country's HLO as percentage of a Bank-chosen 'full learning' benchmark: 625 TIMSS-equivalent points, LAYS is calculated to adjust for the amount of learning that occurs during the expected years of schooling in said country, which is a traditional measure of educational quantity. According to the World Bank (2018a, 2018d), this computed result reflects how much effective learning has occurred, relative to the benchmark case. In Table 1, we present descriptive information for the five education systems with the smallest gap and those with the largest gap between LAYS and Expected Years of Schooling, as indicated in the final column of Table 1. For instance, in 2015, Singapore's HLO score was 581 points, which signifies that approximately 93 percent of all potential learning was realized, relative to the 625-point 'full learning' benchmark case. In other words, while Singaporean students are expected to attend 13.9 years of school, the World Bank (2018a) calculations indicate that only 12.9 years of schooling were effective learning. In contrary, students in Ghana are estimated to realize only 5.7 years of their potential learning when the schooling expectation is 11.6 years, in other words, more than half of their time spent at school are considered to be ineffective by this metric.

In this extrapolation exercise, the Bank's central assumption in harmonizing different ILSAs and regional tests into one concerted scale is that these test programs are inherently comparable, and more specifically, the World Bank (2018c, p. 10) believes that all variation in score distributions on each test "can be attributed entirely to the assessments themselves." In other words, successful synchronization of HLO scores requires that the only difference between test programs to be exclusively test difficulty related. However, as the vast literature that critically examines the systemic incomparability of ILSAs has pointed out (see Brown et al., 2007; Ercikan and Koh, 2005; Kreiner and Christensen, 2014; Oliveri and Ercikan, 2011; Rutkowski and Rutkowski, 2018), this is an unwarranted claim and there exists at least four major reasons why this assumption will not hold.

First, while many ILSAs and regional assessments share similar visions of improving educational quality, they are run by different testing agencies and vary substantively in terms of objective, domain, and

design. Broadly speaking, there exists two distinct class of test programs, one that primarily focuses on evaluating skills, literacy, and competencies, and another that assesses learning outcomes that are more closely aligned with curriculum input (Heyneman and Lee, 2014; Lietz et al., 2017). For instance, ILSAs such as PISA, PIRLS, etc. fall into this first category of skills- and competency-based assessments that evaluate broad definitions of scholastic literacy, while TIMSS, SACMEQ, etc. are designed as curriculum-based student achievement tests. Consequently, resulting from having distinct test-specific objectives, ILSAs tend to adopt very different approaches in study domain and test design. On the one hand, PISA comprehensively covers reading, math, and science using a subject-focused three-year rotation cycle (OECD, 1999). On the other hand, TIMSS, PIRLS, SACMEQ are both subject- and curriculum-specific (Martin et al., 2012), such that TIMSS evaluates only math and science, while PIRLS and early versions of SACMEQ only test students on reading. In addition, test design varies drastically among ILSAs. For example, PISA, TIMSS, PIRLS rely on Item Response Theory (IRT) modelling in test administration, such that each student only completes a subset of the total pool of test materials, whereas in SACMEQ, each student completes all designed test materials (Wasanga et al., 2012). Moreover, in terms of item design, the average length of reading texts on PIRLS is about twice the length of those on PISA (Lietz et al., 2017) and test items are also screened more closely for cultural sensitivity (Martin and Mullis, 2012). These test objective, domain, and design differences are intricate but introduce substantial uncertainty in estimating population-level achievement results (Rutkowski and Rutkowski, 2018).

Second, the list of participant countries, target population, and sampling approach vary substantively among test programs. For instance, PISA participants include primarily OECD member countries and several invited OECD partner systems representing a large cluster of the world's largest economies, whereas the latest wave of TIMSS and PIRLS included more than twenty Central Asian, African, and Middle Eastern participant systems (Rutkowski and Rutkowski, 2018). In contrast, LLECE, SACMEQ, and EGRA are regional test programs and are exclusive to participants within a common geography. At the same time, high participation costs pose a crucial barrier for many countries, as total test administration fees could surmount to as high as US$ 2 million per round (Engel and Rutkowski, 2018). While some of this cost is offset by multilateral donors such as the World Bank (Addey and Sellar, 2019; Lockheed, 2013), domestic taxpayers are likely sponsoring a large portion of current and future rounds of ILSA participation (Engel and Rutkowski, 2018). In terms of target population, some ILSAs adopt an age-based approach while others use a grade-based approach, which could lead to as much as two years in the variation in sample age mean at the country-level (O'Leary, 2001). Another key sampling difference is in sampling units. For instance, while PISA and TIMSS both sample schools at the first stage using proportional-to-size sampling, the second stage for PISA randomly selects students within schools, whereas TIMSS surveys entire intact classrooms (Joncas and Foy, 2012). Finally, even within the same country, different test programs may survey drastically different student populations (Morsy et al., 2018). For instance, only in 2015 did PISA and TIMSS coincide in the same test year, which raises concerns regarding timing saliency and cohort differences. In addition, using the case of Viet Nam, Glewwe et al. (2017) find that PISA student samples are biased upwards in terms of socio-economic status, and that the PISA sample excluded about half of all Vietnamese 15-year-olds who were not enrolled in educational institutions. In retrospect, if the sampled population across different test programs vary substantially, the validity of comparisons generated from synchronizing results across test programs is critically jeopardized.

Third, the salience of school quality and non-school factors fluctuates considerably across systems, making the validity of ILSA comparisons questionable. Specifically, a large body of studies examining the cross-national relationship between school quality and student achievement has documented that the relationship is nuanced not only across country contexts but also across time. For instance, Heyneman and Loxley (1983), Baker et al. (2002), Gamoran and Long (2007), Chudgar and Luschei (2009) and Chiu and Chow (2010) have shown that there exist systematic differences in patterns of student achievement that are highly context- and time-dependent. In addition, a related class of studies highlighting the importance of cultural and historical context has directly questioned whether it is appropriate to attribute strong performance on ILSAs to the soundness of education systems or policies (Steiner-Khamsi and Waldow, 2012; Sellar and Lingard, 2013; Steiner-Khamsi, 2013; Andon et al., 2014; Feniger and Lefstein, 2014; Liu, 2016). For one, studies have shown that public inputs up to the time when students partake in such assessments are not the same cross-nationally, especially in terms of early childhood learning opportunities (Beese and Liang, 2010; OECD, 2011; Gamboa and Waltenberg, 2012). For another, child development conditions at home vary substantially across systems, especially family educational expectations and investments (Kao and Thompson, 2003; Martins and Veiga, 2010; Xu and Hampden-Thompson, 2011; Merry, 2013; Hartas, 2015). Moreover, He and van de Vijver (2013), Zhang et al. (2015), and Perera and Asadullah (2019) find that many school-unrelated contextual factors make direct comparison of ILSA results problematic, since observed score differences can be attributed to underlying family-, cultural-, or measurement-related variation rather than exclusively performance-based.

Fourth, differences in measurement technology and technical standards on each type of tests are considerable. To this end, mode effects of test-administration type have been extensively documented in the assessment literature that substantial differences in results exist between 'paper & pencil' and computer-administered tests (Kroehne and Martens, 2011; Jerrim, 2016). In particular, students in 58 countries and economies participated in the computer-administered PISA 2015 while another 15 countries administered the paper and pencil version. In contrast, all TIMSS 2015 and previous waves of other testing programs were exclusively conducted on paper. Using data from PISA 2015-trials, Jerrim et al. (2018) have shown that students who completed the computer-based version performed consistently worse than those who completed the paper-based test, with differences of up to 0.26 standard deviations. This computer-administration induced score penalty equates almost a full year of learning according to OECD (2016) guidance. In addition, technical quality across different test programs in administration, grading, and standardization are reasonably expected to vary due to the high-level of specificity involved with computer-assisted test implementation (Lietz et al., 2017). Therefore, conversions across tests and years without accounting for differences in mode of test-administration effect or technical standards will inevitably introduce bias unrelated to student performance, making cross-country comparison results unreliable.

4. Methodology

In the analytic section, we use *Harmonized Learning Outcome* (HLO) scores information from the World Bank's Human Capital Index (HCI) database, and merge it by-country and by-year to each country's economic and geographic characteristic information available through the 2019 World Development Indicators dataset. In addition, we assemble publicly-available ILSA and regional test program participation records from institutional databases where these tests are organized, namely the OECD, IEA, UNESCO, etc. Our analysis first begins with a descriptive analysis in understanding how test-overlap systems are systematically different from ones that exclusively partake in one test program. Secondly, we examine the (in)consistency of the 'learning exchange rate' which sits at the core of the HLO synchronization procedure. Following World Bank's (2018c) ratio-linking approach, we reproduce the 'learning exchange rate' informatioon between PISA and TIMSS by calculating the ratio of observed scores on the two test programs. Thirdly, we use the PISA-TIMSS pair to illustrate how test-

overlap systems in effect offer little sample 'overlap' and representativeness even though they participate in both test programs.

As a final analytic step, we evaluate whether test program participation is associated with a 'penalty' score for non-overlap systems. Operationally, we fit an *Ordinary Least Squares* (OLS) multiple regression model in assessing the observed differences in HLO scores, measured in TIMSS-equivalent point units, between systems that participate exclusively in one test program, either ILSA or regional, as opposed to systems partaking in both PISA and TIMSS. In detail, we regress HLO scores for system i in year t on a set of *TestType* dummies and control covariates, in the following form:

$$HLO_{it} = \beta_0 + \gamma^k \bullet TestType + X \bullet \delta + \varepsilon_{it}$$

such that β_0, or the constant coefficient, represents the HLO score for the reference group "*Both PISA & TIMSS*," while a cluster of coefficients γ^k indicates the observed differences in HLO points by test-program participation type k, relative to the reference group. In our model specification, we also include a vector of control variables, X, such as test-year and region fixed-effects which account for year- and region-varying observable and unobservable factors, as well as including country-level income information, represented by GDP per capita in 2015 Purchasing Power Parity terms, in order to account for country-level economy driven variation. ε_{it} is the remaining error-term.

5. Findings

In Table 2, using two illustrative pairs of ILSA-to-ILSA and regional-to-ILSA conversion representing upper and lower ILSA grades respectively, we shed light on the extent to which test-overlap systems, those who participate in multiple ILSAs or regional test programs, are systematically different from systems that participate exclusively in only one test program. In Table 2, Panel A, we observe that systems that participate in both PISA and TIMSS in 2015 exhibit much higher income levels, approximately PPP$11,000 to PPP$13,000 more on average, than systems that only choose to participate in either PISA or TIMSS. In addition, these test-overlap systems also appear on average to provide more years of formal schooling opportunities for young children. In Panel B, similar patterns are observed for the only test-overlap country in SACMEQ 2013 and TIMSS 2011, Botswana, whose income is double that of systems that had participated only in SACMEQ. Moreover, Botswana also provides about one additional year of formal schooling by the end of primary school, compared to 'SACMEQ Only' and 'TIMSS Only' counterparts.

Next, we evaluate the (in)consistency of the 'learning exchange rate' which sits at the core of the HLO conversion procedure. Importantly, World Bank's (2018a, 2018b, 2018c) ILSA and regional test synchronization approach relies on the validity and reliability of 'learning exchange rate' that the Bank constructs using test-overlap systems' performance results as linking mechanism. According to the World Bank (2018b, p.8), the learning exchange rate effectively "captures the difference in difficulty between the two assessments" such that it "enables the construction of globally comparable learning outcomes." To put simply, in order for the conversion to be valid and reliable, the learning exchange rate must only reflect test difficulty differences between the two assessments, and consequently cannot be a function of non-difficulty factors such as time, proficiency, schooling level, or data availability, etc.

In other words, one should expect the learning exchange rate to remain relatively constant, which would be consistent with the World Bank's (2018b) assumption that test difficulty is country-invariant. As shown in Fig. 1, if this were the case, one would expect a tight vertical spread on the Y-axis, representing a consistent learning exchange rate across systems. However, the learning exchange rate is observed to vary substantially, ranging from 0.865 (South Korea) to 1.005 (New Zealand). While this range of 0.140 is seemingly small, it can have a large multiplier effect in subsequent test score conversions, especially in cross-test and cross-year computations. For instance, this fluctuation in the exchange rate can represent as much as a 70-point difference on the imputed HLO scale for a system scoring at 500 TIMSS-equivalent points, in other words, as much as 14 percent of a system's score could be varying due to exchange rate instability alone. This is expected to drastically affects the validity and reliability of HLO harmonization across test programs. More importantly, as we observe in Fig. 1, the learning exchange rate between PISA and TIMSS appears to be negatively correlated with TIMSS performance, as indicated by the downward trend in country observations. This goes to suggest that when test-

Table 2
Comparing Test-Overlap and non-Overlap Country Profiles, PISA 2015 and TIMSS 2015.
Source: OECD (2016); Mullis et al. (2016); UNESCO (2018); World Bank (2019b). Authors' compilation.

	List of Countries/Economies	Average GDP per capita (PPP 2015)	Average Formal Schooling (Years): Pre-primary & Primary
Panel A | PISA 2015 & TIMSS 2015 (8th Grade) Linking			
Test-Overlap	Argentina (Bueno Aires), Australia, Canada, Chile, Chinese Taipei, Georgia, Hong Kong, Hungary, Ireland, Israel, Italy, Japan, Jordan, Kazakhstan, Lebanon, Lithuania, Malaysia, Malta, New Zealand, Norway, Qatar, Russian Federation, Singapore, Slovenia, South Korea, Sweden, Thailand, Turkey, USA, United Arab Emirates	29314	8.67
PISA 2015 Only	Albania, Algeria, Belgium, Brazil, China, Colombia, Costa Rica, Dominican Republic, Estonia, Greece, Iceland, Kosovo, Luxembourg, Macedonia, Mexico, Moldova, Montenegro, Peru, Romania, Switzerland, Tunisia, Uruguay, Viet Nam, United Kingdom	18273	8.39
TIMSS 2015 Only	Azerbaijan, Bahrain, Botswana, Cyprus, Egypt, Iran, Kazakhstan, Kuwait, Morocco, Oman, Saudi Arabia, Serbia, South Africa	16691	8.08
Panel B | SACMEQ 2013 & TIMSS 2011 (4th Grade) Linking			
Test-Overlap	Botswana	15357	10
SACMEQ 2013 Only	Kenya, Lesotho, Mauritius, Malawi, Namibia, Seychelles, Uganda, Zambia, Zimbabwe	7530	9.05
TIMSS 2011 Only	Azerbaijan, Bahrain, Botswana, Canada, Chile, Chinese Taipei, Cyprus, Egypt, Georgia, Hong Kong, Hungary, Iran, Ireland, Israel, Italy, Japan, Jordan, Kazakhstan, Kuwait, Lebanon, Lithuania, Malaysia, Malta, Morocco, New Zealand, Norway, Oman, Qatar, Russian Federation, Saudi Arabia, Serbia, Singapore, Slovenia, Sweden, South Africa, South Korea, Thailand, Turkey, USA, United Arab Emirates	21411	8.45

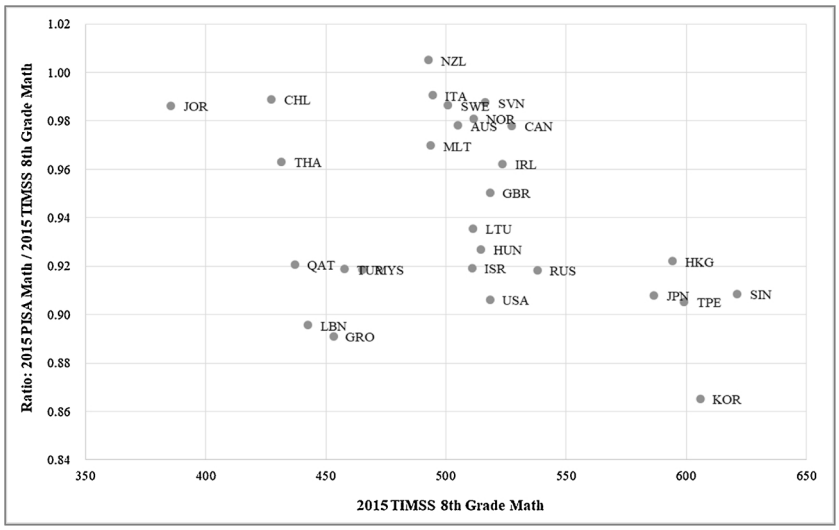

Fig. 1. 'Learning Exchange Rate' among Test-Overlap Systems, PISA and TIMSS 2015.
Source: OECD (2016) and Mullis et al. (2016). Authors' compilation.

overlap systems happen to be a cluster of top-performers, such as Singapore, South Korea, Japan, etc., the conversion ratio tends to be downward-biased; whereas, if test-overlap systems are relatively low-performing, such as Jordan, Chile, Thailand, etc., the exchange rate will be upward-biased. In other words, the imputation results derived using the World Bank's (2018b) approach are highly dependent on which systems are included as test-overlap in a given year, which introduces substantial performance-unrelated noise into conversions made. To this end, our results presented here indicate that relative test difficulty is in fact country-specific, which is in contradiction with the World Bank's (2018b) assumption that test difficulty is country-invariant.

Consequently, in Table 3, focusing on PISA- and TIMSS-overlap systems as example, we try to shed light on why the learning exchange rate could be so inconsistent across systems by presenting the absolute numeric difference in sample characteristics, by country. In particular, we find that even within the same test system and test year, different test programs have chosen drastically different sampling outcomes. Importantly, PISA and TIMSS samples for the same system and year can yield on average an 11 percentage-point difference in target population coverage (Table 3, Column 3) and an average of 1.6 years in age gap (Table 3, Column 6). In other words, drastically different students and cohorts are chosen to participate in these two test programs, and therefore the learning exchange rate reflects not only test difficulty differences but also remains a function of sampling design and cohort effects, among other factors that vary across tests and years.

In Table 4, we quantify the magnitude of non- and partial-participation penalty by presenting results from regressing HLO scores on systems' test participation type for 157 countries and economies included in the World Bank HCI dataset. Specifically, Model 1 reports unstandardized coefficients from the base model with no control variables, Model 2 indicates that with additional test-year and country-region variation accounted for, and results reported in Model 3 includes controlling for GDP per capita information. Firstly, in Model 1, we can observe that about 57.8 percent of the variation in HLO scores is associated with test participation type. Importantly, systems who partake in only PISA is associated with 62.66 HLO points ($p < .001$) less than test-overlap systems who participate in both PISA and TIMSS. This difference in HLO scores is observed to be -71.47 points ($p < .001$) between TIMSS/PIRLS participants and test-overlap systems who participate in both PISA and TIMSS test programs. Accordingly, such associated gaps translate to approximately 2 full years of learning (OECD, 2016). Similar trends are observed for all test types, with the average score difference for 'SACMEQ only' systems to be -94.41 points

Table 3
Comparing PISA 2015 and TIMSS 2015 (8th Grade) Samples, Test-Overlap Systems.
Source: OECD (2016) and Mullis et al. (2016). Authors' compilation.

Countries/ Economies	Target Population Coverage (%)		Difference: |(1)–(2)|	Sample Mean: Age (Years)		Difference: |(3) - (4)|
	PISA 2015 (1)	TIMSS 2015 (2)		PISA 2015 (3)	TIMSS 2015 (4)	
Argentina (Bueno Aires)	55.0	97.3	42.3	15.8	14.1	1.7
Australia	91.0	96.5	5.5	15.8	14.0	1.8
Canada	84.0	95.2	11.2	15.8	14.0	1.8
Chile	80.0	98.1	18.1	15.8	14.3	1.5
Chinese Taipei	85.0	98.3	13.3	15.8	14.3	1.5
Georgia	79.0	94.0	15	15.9	13.7	2.2
Hong Kong SAR, China	89.0	98.4	9.4	15.8	14.2	1.6
Hungary	90.0	94.6	4.6	15.8	14.7	1.1
Ireland	96.0	98.8	2.8	15.7	14.4	1.3
Israel	94.0	77.2	16.8	15.7	14.0	1.7
Italy	80.0	93.9	13.9	15.8	13.8	2.0
Japan	95.0	97.7	2.7	15.8	14.5	1.3
Jordan	86.0	99.0	13	15.9	13.8	2.1
Kazakhstan	91.0	96.2	5.2	15.8	14.3	1.5
Korea, Republic of	92.0	79.0	13	15.7	14.4	1.3
Lebanon	66.0	98.7	32.7	15.8	14.2	1.6
Lithuania	90.0	93.0	3	15.8	14.7	1.1
Malaysia	76.0	95.6	19.6	15.8	14.3	1.5
Malta	98.0	96.5	1.5	15.7	13.9	1.8
New Zealand	90.0	96.9	6.9	15.8	14.1	1.7
Norway	91.0	96.3	5.3	15.8	13.7	2.1
Qatar	93.0	96.8	3.8	15.7	14.1	1.6
Russian Federation	95.0	96.3	1.3	15.8	14.7	1.1
Singapore	96.0	93.0	3	15.8	14.2	1.6
Slovenia	93.0	96.2	3.2	15.7	13.8	1.9
Sweden	94.0	94.5	0.5	15.7	14.7	1.0
Thailand	71.0	99.8	28.8	15.7	14.4	1.3
Turkey	70.0	98.7	28.7	15.8	13.9	1.9
United Arab Emirates	91.0	96.4	5.4	15.8	13.9	1.9
United States	84.0	93.2	9.2	15.8	14.2	1.6
Column Mean	86.4	95.2	11.3	15.8	14.2	1.6

Table 4
OLS regression results on the relationship between ILSA participation type and HLO scores.

Dependent Variable	HLO		
Test Participation Type (Reference = Both PISA & TIMSS)	Model (1)	Model (2)	Model (3)
PISA Only	−62.66***	−49.87***	−37.43***
	(11.48)	(10.01)	(9.11)
TIMSS/PIRLS Only	−71.47***	−36.59**	−28.47**
	(12.21)	(11.62)	(10.39)
SACMEQ Only	−94.41***	−28.37	−7.20
	(15.26)	(28.32)	(25.35)
LLECE Only	−101.16***	−58.95*	−45.48
	(18.42)	(27.55)	(24.53)
PASEC Only	−123.31***	−74.83*	−29.37
	(13.54)	(28.98)	(26.77)
EGRA Only	−141.70***	−109.59***	−72.36***
	(10.25)	(20.55)	(19.23)
Control Covariates			
Test Year Fixed-Effect	No	Yes	Yes
Region Fixed-Effect	No	Yes	Yes
GDP per capita (2015 PPP)	No	No	Yes
Constant	504.43***	480.87***	217.96***
	(6.83)	(10.95)	(58.34)
Adj. R-squared	.578	.729	.787
N	157	157	157

Note: standard errors in parentheses, * denotes p-values < .05, ** denotes p-values < .01, *** denotes p-values < .001.

(p < .001), 'LLECE Only' systems to be -101.16 points (p < .001), 'PASEC Only' and 'EGRA Only' systems are -123.31 (p < .001) and -141.7 points (p < .001) respectively. Such differences on HLO scores are to be interpreted as due to a broad class of factors, including test design, timing, sampling, non-school factors. Secondly, in Model 2, we eliminate time- and geography-varying influences by including test year and region fixed-effects. Results show that the score 'penalty' by test participation persists for all test program types. For instance, 'PISA Only' and 'TIMSS/PIRLS Only' systems exhibit an average of 49.87 lower points (p < .001) and 36.59 lower points (p < .001) relative to test-overlap systems. To be expected, as regional test programs are region-specific, score differences are substantially reduced. Thirdly, in Model 3, where we include purchasing-power adjusted per capita GDP information as additional control, lower HLO scores continue to be substantial and statistically significant for 'PISA Only' (-37.43 points, p < .001), 'TIMSS/PIRLS Only' (-28.47 points, p < .001) and 'EGRA Only' systems (-72.36 points, p < .001), with the first two categories still indicating more than a full year difference in student learning terms. After accounting for per capita GDP, systems that only participate in regional test programs SACMEQ, LLECE, or PASEC are shown to be statistically indifferent from systems that participate in both PISA and TIMSS.

6. Discussion and conclusion

This study began with the following research questions: (i) How reliable is the intra-test exchange rate? (ii) Does participation on different test programs matter for HLO results, and how much is the associated penalty? Firstly, our results show that the learning exchange rate that World Bank devised to synchronize scores across ILSA and regional test programs is highly volatile and can represent as much as a 70-point difference on the imputed scale for a country scoring at 500 TIMSS-equivalent points. This finding implies that the Bank's algorithmic approach is highly inadequate in addressing the inter- and intra-country sample differences, and which systems fall in the test-overlap subset can have large influences on test harmonization. For one, Lockheed (2015) has shown that country participation on ILSAs is highly selective, and the propensity to participate in PISA is markedly higher for systems that are OECD members, located in Europe and Central Asia, high- and upper-middle-income countries, and with prior national or ILSA participation experience. For another, in this study, we find that test-specific sample differences are substantial, for instance, there is little sample-overlap even within test-overlap systems such that there exists on average an 11 percentage-point difference in target population coverage and 1.6 years in age gap between PISA and TIMSS samples for the same country and year.

Secondly, our results indicate that test participation type alone accounts for about 57.8 percent of the variation in HLO scores, and the score penalty effect associated with non- or partial-participation is evaluated to be between 1 to 2 full years of learning, and is especially salient for systems that exclusively participate in regional test programs. To revert to the Swiss example, mentioned above: Switzerland ranks second in the Human Development Index (0.944) but twentieth on the Human Capital Index (0.767). Switzerland's relatively low Learning Adjusted Years of Schooling (11.1 years) is affected by its non-participation in TIMSS, which required PISA-TIMSS score conversion. As we have demonstrated in this study, had Switzerland participated both in PISA and TIMSS, one would expect an average increase of 37 points on the World Bank's HLO scale, which would translate to having almost a full year more learning. Importantly, findings in our study uncover this large yet 'invisible' score penalty embedded in the synchronization procedure, which adds to existing studies that have examined monetary (Engel and Rutkowski, 2018) and non-monetary costs (Addey and Sellar, 2019) to participate in ILSA. This non-trivial score penalty that is contingent on ILSA participation type not only affects country decisions as a hidden cost in choosing to participate in one test over another or not participating at all, but also adds coercive rationales to systems' ILSA engagement in that the World Bank (2018a) plans to use this information to assess and reevaluate its lending programs as well as align political and economic incentives.

As this study has argued, the core issues challenging the validity and reliability of the Bank's HLO score synchronization procedure and subsequent HCI calculations are in three-folds. Firstly, researchers and policy makers should realize that all ILSA and regional programs have specific and differentiated goals, and some are purposefully devised to ask and answer very different sets of questions. Without accounting for the critical differences in test objective, domain, and design, the Bank's HLO synchronization exercise becomes a liability for the new HCI computation and greatly undermines its credibility, as indicated through our results. Secondly, it is crucial to acknowledge that in ILSA test synchronization, country-overlap does not equate sample overlap. As our finding illustrate, this sample overlap assumption is unlikely to hold even for the same country within the same test-year, and inadequate accounting of sample differences creates large volatility in exchange rate computations among test programs. To this end, post hoc adjustments to correct for test-specific sample variation are vital to ensure score convertibility, should the Bank continues its HLO calculations. Thirdly, measurement errors, such as test exchange rate volatility, in each step of the synchronization process may have substantial multiplier effects later on, and introduce large biases due to the structure of the Bank's test-specific score, HLO, LAYS, HCI linking approach. As our results have shown, test exchange rate volatility can affect a median-performing country's HLO score by as much as 14 percent. Therefore, in order for the Human Capital Index to be valid and reliable enough to be act as a global human capital and education quality tracking instrument, these aforementioned challenges should not be taken lightly.

As the World Bank's (2017) flagship publication World Development Report points out, too often meaningful learning is absent in schools. To this end, the Bank's renewed interests in human capital and education development, and its global call-to-action to investing in people marks an important milestone for education experts at the Bank. At the same time, we should not forget that there is no direct

relationship between testing learning outcomes and improving the quality of education in schools. Many changes at the policy and practice levels are needed to have an effect on meaningful and effective learning. In effect, the introduction of test-based accountability reforms has led to a proliferation of standardized student assessment (Verger et al., 2019). The shift towards outcomes-based school reform is not without its critics. It is one of the most visible signposts of the market model in education that, as part and parcel of neo-liberal reforms, encourages businesses to enter the education sector first as providers of goods and services (Robertson and Verger, 2012; see also Steiner-Khamsi and Draxler, 2018), and more recently also as policy advisors (Lubienski, 2019). The reliance on outcomes measurement, coupled with the globalization of the knowledge-based economy, explains the sharp rise of global indicators that measure and compare national developments such as, the Human Capital Index. 'Governance by numbers' (Grek, 2008) or 'soft power' (Niemann and Martens, 2018) have been some of the terms to capture the coercion that is implied with global monitoring, evaluation, and international comparisons.

In addition to the general criticism regarding the global norm-setting for national developments, mentioned above, the new Human Capital Index (HCI), sitting at the core of the Bank's ambitious education agenda, requires caution in its calculation, usage, and interpretation. In effect, the intra- and extrapolations to compensate for missing data introduce a new typology of complexity to ILSA participation, such that systems which either did not participate or only partially participated in ILSAs are subject to large score penalties. On the one hand, existing studies have explored in depth the demand-side reasons why systems choose to participate in ILSAs. This study, on the other hand, contributes to the understanding of supply-side practices that are bound to generate coercive rationales for ILSA participation through score penalties. Our findings illustrate the source and magnitude of such non- and partial-participation penalties, which will inevitably complicate and alter the existing country-ILSA coupling and interaction in the mid-to-long term. Future studies are much needed to examine this topic further, and in particular, how Human Capital Index and its related World Bank programs evolve and project new rationales for systems that are either not or partially participating ILSA and regional testing.

As this study has attempted to demonstrate, the Human Capital Index relies on, and in effect exacerbates test-based accountability. It needs to be considered the new 'soft power' governance tool *par excellence,* because it penalizes systems that choose not to participate in international-large scale student assessment. In effect, their score on the Harmonized Learning Outcome is lowered by 37 points, or equivalent to 1 full year of learning. In an era of global monitoring of national development, these "methodological glitches" of the HCI have far-reaching political and economic consequences that need to be examined in future research.

Funding

This work was supported by the Shaanxi Normal University RenCai Faculty Seed Fund [Grand ID: 1301031829].

Declaration of Competing Interest

No potential conflict of interest was reported by the authors.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijedudev.2019.102149.

References

Addey, C., 2015. Participating in international literacy assessments in Lao PDR and Mongolia: a global ritual of belonging. In: Hamilton, M., Maddox, B., Addey, C. (Eds.), Literacy as Numbers: Researching the Politics and Practices of International Literacy Assessment. Cambridge University Press, Cambridge MA, pp. 147–164.
Addey, C., Sellar, S., Steiner-Khamsi, G., Lingard, B., Verger, A., 2017. The rise of international large-scale assessments and rationales for participation. Comp. J. Comp. Educ. 47 (3), 434–452.
Addey, C., Sellar, S., 2019. Is it worth it? Rationales for (non)participation in international large-scale learning assessments. Education Research and Foresight Working Papers. UNESCO, Paris.
Andon, A., Thompson, C., Becker, B., 2014. A quantitative synthesis of the immigrant achievement gap across OECD countries. Large-scale Assess. Educ. 2 (1), 2–20.
Baker, D., Goesling, B., LeTendre, G., 2002. Socioeconomic status, school quality, and national economic development: a cross-national analysis of the "Heyneman-Loxley Effect" on mathematics and science achievement. Comp. Educ. Rev. 46 (3), 291–312.
Barro, R., Lee, J., 1993. International comparisons of educational attainment. J. Monet. Econ. 32 (3), 363–394.
Barro, R., Lee, J., 2001. International data on educational attainment: updates and implications. Oxf. Econ. Pap. 53 (3), 541–563.
Becker, G., 1962. Investment in human capital: a theoretical analysis. J. Polit. Econ. 70 (5, II), 9–49.
Beese, J., Liang, X., 2010. Do resources matter? PISA science achievement comparisons between students in the United States, Canada and Finland. Improv. Sch. 13 (3), 266–279.
Brown, G., Micklewright, J., Schnepf, S., Waldmann, R., 2007. International surveys of educational achievement: how robust are the findings? J. R. Stat. Soc. 170 (3), 623–646.
Chabbott, C., 2007. Carrot soup, magic bullets, and scientific research for education and development. Comp. Educ. Rev. 51 (1), 71–94.
Chiu, M., Chow, B., 2010. Culture, motivation, and reading achievement: high school students in 41 countries. Learn. Individ. Differ. 20 (6), 579–592.
Chudgar, A., Luschei, T.F., 2009. National income, income inequality, and the importance of schools: a hierarchical cross-national comparison. Am. Educ. Res. J. 46 (3), 626–658.
Cohen, D., Soto, M., 2007. Growth and human capital: good data, good results. J. Econ. Growth 12 (1), 51–76.
Engel, L., 2015. Steering the national: exploring the education policy uses of PISA in Spain. Eur. Educ. 47 (2), 100–116.
Engel, L., Rutkowski, D., 2018. Pay to play: what does PISA participation cost in the US? Discourse Stud. Cult. Politics Educ. 1–13.
Ercikan, K., Koh, K., 2005. Examining the construct comparability of the English and French versions of TIMSS. Int. J. Test. 5 (1), 23–35.
Espeland, W., 2015. Narrating numbers. In: Rottenburg, R., Merry, S., Park, S., Mugler, J. (Eds.), The World of Indicators: The Making of Governmental Knowledge through Quantification, Cambridge Studies in Law and Society. Cambridge University Press, Cambridge, pp. 56–75.
Feniger, Y., Lefstein, A., 2014. How not to reason with PISA data: an ironic investigation. J. Educ. Policy 29 (6), 845–855.
Fischman, G., Topper, A., Silova, I., Goebel, J., Holloway, J., 2019. Examining the influence of international large-scale assessments on national education policies. J. Educ. Policy 34 (4), 470–499.
Gamboa, L., Waltenberg, F., 2012. Inequality of opportunity for educational achievement in Latin America: evidence from PISA 2006–2009. Econ. Educ. Rev. 31 (5), 694–708.
Gamoran, A., Long, D., 2007. Equality of educational opportunity: a 40 year retrospective. In: Teese, R., Lamb, S., Duru-Bellat, M., Helme, S. (Eds.), International Studies in Educational Inequality, Theory and Policy. Springer, Dordrecht, pp. 23–47.
Glewwe, P., Lee, J., Vu, K., Dang, H.A., 2017. What explains vietnam's exceptional performance in education relative to other countries? Analysis of the 2012 PISA data. In: RISE Annual Conference Papers, Center for Global Development. Washington, DC, June.
Gorur, R., 2016. Seeing like PISA: a cautionary tale about the performativity of international assessments. Eur. Educ. Res. J. 15 (5), 598–616.
Grek, S., 2008. From symbols to numbers: the shifting technologies of education governance in Europe. Eur. Educ. Res. J. 7 (2), 208–218.
Hanushek, E., Kimko, D., 2000. Schooling, labor-force quality, and the growth of nations. Am. Econ. Rev. 90 (5), 1184–1208.
Hanushek, E., Woessmann, L., 2012. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. J. Econ. Growth 17 (4), 267–321.
Hartas, D., 2015. Patterns of parental involvement in selected OECD countries: cross-national analyses of PISA. Eur. J. Educ. Res. 4 (4), 185–195.
He, J., van de Vijver, F., 2013. Methodological issues in cross-cultural studies in educational psychology. Advancing cross-cultural perspectives on educational psychology: A festschrift for Dennis McInerney. pp. 39–56.
Heyneman, S., Loxley, W., 1983. The effect of primary-school quality on academic achievement across twenty-nine high-and low-income countries. Am. J. Sociol. 88 (6), 1162–1194.
Heyneman, S., 2003. The history and problems in the making of education policy at the World Bank 1960-2000. Int. J. Educ. Dev. 25 (4), 315–337.
Heyneman, S., Lee, B., 2012. Impact of international studies of academic achievement on policy and research. In: Rutkowski, L., Davier, M., Rutkowski, D. (Eds.), Handbook of International Large Scale Assessment: Background, Technical Issues, and Methods of Data Analysis. Chapman and Hall Publishers, London, pp. 37–74.
Heyneman, S., Lee, B., 2014. International large scale assessments: uses and implications. In: Ladd, H., Goertz, P. (Eds.), Handbook of Research in Education and Finance. Routledge, New York, pp. 105–212.
Hogan, A., Sellar, S., Lingard, B., 2016. Commercialising comparison: Pearson puts the TLC in soft capitalism. J. Educ. Policy 31 (3), 243–258.

Jerrim, J., 2016. PISA 2012: how do results for the paper and computer tests compare? Assess. Educ. Princ. Policy Pract. 23 (4), 495–518.
Jerrim, J., Micklewright, J., Heine, J.H., Salzer, C., McKeown, C., 2018. PISA 2015: how big is the 'mode effect' and what has been done about it? Oxf. Rev. Educ. 44 (4), 476–493.
Joncas, M., Foy, P., 2012. Sample Design in TIMSS and PIRLS Methods and Procedures. TIMSS and PIRLS International Study Center: Lynch School of Education, Boston College, Boston.
Kao, G., Thompson, J., 2003. Racial and ethnic stratification in educational achievement and attainment. Annu. Rev. Sociol. 29 (1), 417–442.
Kreiner, S., Christensen, K., 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.
Kroehne, U., Martens, T., 2011. Computer-based competence tests in the national educational panel study: the challenge of mode effects. Zeitschrift Für Erziehungswissenschaft 14 (2), 169.
Lewis, S., 2017. PISA for schools: respatializing the OECD's global governance of education. In: In: Wiseman, A., Taylor, C. (Eds.), The Impact of the OECD on Education Worldwide (International Perspectives on Education and Society 31 Emerald Publishing Limited, Portland, OR.
Lietz, P., Cresswell, J.C., Rust, K.F., Adams, R.J. (Eds.), 2017. Implementation of Large-Scale Education Assessments. John Wiley & Sons, Hoboken, NJ.
Liu, J., 2016. Student achievement and PISA rankings: policy effects or cultural explanations? In: In: Smith, W. (Ed.), The Global Testing Culture: Shaping Education Policy, Perceptions, and Practice 25. Oxford Studies in Comparative Education. Oxford: Symposium Books, pp. 85–99.
Liu, J., 2019. Government, media, and citizens: understanding engagement with PISA in China (2009-2015). Oxf. Rev. Educ. 45 (3), 315–332.
Lockheed, M., 2013. Causes and consequences of international large-scale assessments in developing countries. In: Meyer, H., Benavot, A. (Eds.), PISA, Power, and Policy: The Emergence of Global Educational Governance. Symposium Books, Wallingford, UK, pp. 163–183.
Lockheed, M., 2015. Why do countries participate in International Large-Scale Assessments? The case of PISA. Policy Research Working Paper 7447. World Bank, Washington, D.C.
Lockheed, M., Wagemaker, H., 2013. International large-scale assessments: thermometers, whips or useful policy tools? Res. Comp. Int. Educ. 8 (3), 296–306.
Lubienski, C., 2019. Advocacy networks and Market models for education. In: Parreira do Amaral, M., Steiner-Khamsi, G., Thompson, C. (Eds.), Researching the Global Education Industry. Palgrave, New York, pp. 69–86.
Martin, M., Mullis, I., 2012. Methods and Procedures in TIMSS and PIRLS 2011. Boston College, Chestnut Hill, MA.
Martin, M., Mullis, I., Foy, P., Stanco, G., 2012. TIMSS 2011 International Results in Science. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.
Mullis, I., Martin, M., Foy, P., Hooper, M., 2016. TIMSS 2015 International Results in Mathematics. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.
Martins, L., Veiga, P., 2010. Do inequalities in parents' education play an important role in PISA students' mathematics achievement test score disparities? Econ. Educ. Rev. 29 (6), 1016–1033.
Merry, J., 2013. Tracing the US deficit in PISA reading skills to early childhood: evidence from the United States and Canada. Sociol. Educ. 86 (3), 234–252.
Morsy, L., Khavenson, T., Carnoy, M., 2018. How international tests fail to inform policy: the unsolved mystery of Australia's steady decline in PISA scores. Int. J. Educ. Dev. 60 (5), 60–79.
Niemann, D., Martens, K., 2018. Soft governance by hard fact? The OECD as a knowledge broker in education policy. Glob. Soc. Policy 18 (3), 267–283.
OECD. (n.d.). A Longer-term Strategy of PISA. Paris: OECD-PISA Governing Board.
OECD, 1999. Measuring Student Knowledge and Skills: A New Framework for Assessment. OECD, Paris.
OECD, 2011. PISA In Focus 1: Does Participation in Pre-primary Education Translate into Better Learning Outcomes at School? OECD, Paris.
OECD, 2016. Education at a Glance 2016: OECD Indicators. OECD Publishing, Paris, France.
O'Leary, M., 2001. The effects of age-based and grade-based sampling on the relative standing of countries in international comparative studies of student achievement. Br. Educ. Res. J. 27 (2), 187–200.
Oliveri, M., Ercikan, K., 2011. Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions? Appl. Meas. Educ. 24 (4), 349–366.
Perera, L., Asadullah, A., 2019. Mind the gap: what explains Malaysia's underperformance in Pisa? Int. J. Educ. Dev. 65 (3), 254–263.
Robertson, S.L., Verger, A., 2012. Governing education through public private partnership. In: Robertson, S.L., Mundy, K., Verger, A., Menashy, F. (Eds.), Public Private Partnership in Education. New Actors and Modes of Governance in a Globalizing World. Edward Elgar, Cheltenham, pp. 21–42.
Rutkowski, L., Rutkowski, D., 2018. Improving the comparability and local usefulness of international assessments: a look back and a way forward. Scand. J. Educ. Res. 62 (3), 354–367.
Schultz, T., 1961. Investment in human capital. Am. Econ. Rev. 1–17.
Sellar, S., Lingard, B., 2013. The OECD and global governance in education. J. Educ. Policy 28 (5), 710–725.
Shields, R., Menashy, F., 2019. The network of bilateral aid to education 2005–2015. Int. J. Educ. Dev. 64 (1), 74–80.
Steiner-Khamsi, G., 2013. What is wrong with the 'what-went-right' approach in educational policy? Eur. Educ. Res. J. 12 (1), 20–33.
Steiner-Khamsi, G., 2006. The development turn in comparative and international education. Eur. Educ.: Issues Stud. 38 (3), 19–47.
Steiner-Khamsi, G., Draxler, A., 2018. Introduction. In: Steiner-Khamsi, G., Draxler, A. (Eds.), The State, Business, and Education: Public-Private Partnerships Revisited. E. Elgar, Cheltenham, UK, pp. 1–15.
Steiner-Khamsi, G., Waldow, F., 2012. World Yearbook of Education 2012: Policy Borrowing and Lending in Education. Routledge, New York.
UNESCO, 2018. UNESCO Institute for Statistics (UIS) Database. UNESCO-UIS, Montreal.
Verger, A., Parcerisa, L., 2017. A difficult relationship. accountability policies and teachers: international evidence and key premises for future research. In: Akiba, M., LeTendre, G. (Eds.), International Handbook of Teacher Quality and Policy. Routledge, New York, pp. 241–254.
Verger, A., Fontdevila, C., Parcerisa, L., 2019. Reforming governance through policy instruments: how and to what extent standards, tests and accountability in education spread worldwide. Discourse Stud. Cult. Politics Educ. 40 (2), 248–270.
Waldow, F., Steiner-Khamsi, G., 2019. Understanding PISA's Attractiveness: Critical Analyses in Comparative Policy Studies. Bloomsbury Academic, London.
Wasanga, P., Ogle, M., Wambua, R., 2012. The SACMEQ III Project in Kenya: a Study of the Conditions of Schooling and the Quality of Education. UNESCO, Paris.
Wiseman, A., 2013. Policy responses to PISA in comparative perspective. In: Meyer, H., Benavot, A. (Eds.), PISA, Power, and Policy. The Emergence of Global Educational Governance. Symposium Books, Wallingford, UK.
World Bank, 2017. World Development Report 2018: Learning to Realize the Promise of Education. World Bank, Washington, D.C.
World Bank, 2018a. The Human Capital Project. World Bank, Washington, D.C.
World Bank, 2018b. Measuring Human Capital. World Bank, Washington, D.C.
World Bank, 2018c. Global Dataset on Education Quality: A Review and Update (2000–2017). World Bank, Washington, D.C.
World Bank, 2018d. Learning-Adjusted Years of Schooling (LAYS): Defining a New Macro Measure of Education. World Bank, Washington, D.C.
World Bank, 2019a. World Development Report 2019: The Changing Nature of Work. World Bank, Washington, D.C.
World Bank, 2019b. World Development Indicators. World Bank, Washington, D.C.
Xu, J., Hampden-Thompson, G., 2011. Cultural reproduction, cultural mobility, cultural resources, or trivial effect? A comparative approach to cultural capital and educational performance. Comp. Educ. Rev. 56 (1), 98–124.
Zhang, L., Khan, G., Tahirsylaj, A., 2015. Student performance, school differentiation, and world cultures: evidence from PISA 2009. Int. J. Educ. Dev. 42 (5), 43–53.