NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS
INSTITUTE OF EDUCATION

as a manuscript

Tatiana Khavenson

Methodology of Using Large-Scale Assessment Studies in Education for Educational Policy

Executive summary of a PhD thesis (the thesis comprises five published research papers)

PhD Dissertation Summary for the purpose of obtaining
Doctor of Philosophy in Education HSE

Academic advisor:
Professor Martin Carnoy,
PhD

Moscow - 2019

The work on thesis was completed in the Institute of Education at National Research University Higher School of Economics.

The thesis comprises five published research papers (appendices 1-5):

1. Carnoy, M., Khavenson, T., & Ivanova, A. (2015). Using TIMSS and PISA results to inform educational policy: a study of Russia and its neighbours. *Compare: A Journal of Comparative and International Education*, *45*(2), 248–271. (Scopus – Q1, WoS – Q2)

2. Carnoy, M., Marotta, L., Louzano, P., Khavenson, T., Guimaraes, F. R. F., & Carnauba, F. (2017). International Comparative Education: What State Differences in Student Achievement Can Teach Us about Improving Education—the Case of Brazil. *Comparative Education Review*, *61*(4), 726–759. (Scopus – Q1, WoS – Q2)

3. Khavenson, T., & Carnoy, M. (2016). The unintended and intended academic consequences of educational reforms: the cases of Post-Soviet Estonia, Latvia and Russia. *Oxford Review of Education*, *42*(2), 178–199. (Scopus – Q1, WoS – Q2)

4. 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. *International Journal of Educational Development*, *60*, 60–79. (Scopus – Q1, WoS – Q2)

5. Khavenson T. (2019). School integration through the curriculum reform in Latvia and Estonia // Educational studies (*in Russian*) (forthcoming). (Scopus – Q3)

Selected conferences, where research results were presented

1. 61st Annual Conference Comparative and International Education Society. March 5-9, 2017. Atlanta, GA, USA. Talk: “Educational inequality in Russia: The role of socio-economic status and academic achievements”.
2. International conference “Achievements in large-scale assessment studies (TIMSS and PISA) as indicators of changes taking place in education system”. February 1st 2017. Moscow, Russia. Talk: “Participating in international large-scale assessment. Main findings for education policy”.
3. Laboratory for Comparative Social Research International Annual Conference. November 2015. Moscow, Russia. Talk: “The Unintended and Intended Academic Consequences of Educational Reforms: The Cases of Post-Soviet Estonia and Latvia”.
4. 59th Annual Conference Comparative and International Education Society. March 8-13, 2015. Washington, DC, USA. Talk: “Analyzing the Impact of Educational Reforms on Russian- Medium Students in the Baltic Countries and Russia: A "Natural Experiment”.
5. XV April International Academic Conference on Economic and Social Development. April 2-4 2014. Moscow, Russia. Talk: “The relationship between teachers’ characteristics and students’ achievement gain between TIMSS 2011 and PISA 2012. The study of Russian case”.
6. Pedagogical conference “PISA survey: A Reflection of Estonian Education”. August 25th 2014. Narva, Estonia. Talk: “Russian PISA results”.
7. XX All-Russian conference «Developmental pedagogy: critique and modern challenges». April 25-27 2013. Krasnoyarsk, Russia. Talk: “Approaches for using research data for education policy”.
8. International Conference “Russian Education in the Mirror of the International Comparative Studies”. June 19-20 2013. Moscow, Russia. Talk: “Can international test score comparisons inform educational policy? A closer look at student performance in Russia and its neighbors”.
9. The fourth Conference of the European Survey Research Association. July 18-22, 2011. Lausanne, Switzerland. Talk: “Combining different types of data in educational research. Data organization issues”.
10. VIII International conference "Trends in Education Development: issues in management and quality evaluation”. February 18th 2011. Moscow, Russia. Talk: “Achievement change in TIMSS and PISA and the explanatory capacity of contextual data. The case of Russia and Eastern European countries”.

Introduction

International Large-Scale Assessments (ILSA) that aim to assess the educational achievements (such as TIMSS[1], PIRLS[2], PISA[3], and others) have been a potentially valuable source of data for evidence based educational policy from their initial introduction (Howie and Plomp 2005; Martens, Niemann, and Teltemann 2016; Mullis, Martin, and Loveless 2016). Their mission has been to provide researchers, educational policy makers and practitioners with an extensive comparative data base to make informed decisions and to enhance the quality of education worldwide. Among the functions they were intended to fill from the scientific point of view were description, benchmarking, monitoring, and stimulating cross-national research (Plomp 1998). These were later expanded by policy interests which included: cultural (e.g., understanding the foundations of education), historical (empirical descriptions of schooling that allowed the past to be compared with the present), international comparison, meeting future challenges in education, accountability, economic, policy and administrative perspectives (e.g., how well are the aims and principles communicated across levels?) (Howie and Plomp 2005). ILSA are also useful in defining relevant outcomes, identifying promising interventions, and targeting specific populations of interest even if a country has its own large-scale educational studies (Schneider et al. 2007). Appearance of these data shifted the global agenda in education to a new level of information usage both in research and education policy. In general evidence based educational policy is the approach that is considered as an important practice commonly used in most developed countries. In the UK the specific procedure of incorporating research in policy making were introduced in 1999 (Shaxson 2005). The idea behind that was to provide decision makers with the best available real-life data. In (Borer and Lawn 2013) the process of developing the evidence-based policy approach during the 20 century is described. As well as the role of such organizations like the World Bank, UNESCO, the OECD, and others are playing in this process.

In Russia, ILSA have been part of a system that assesses the quality of education since the mid 1990s, when the TIMSS test was first applied to a sample of Russian students. This helps explains the growing interest among Russian educational policy makers and educators more generally in such international assessments (Tyumeneva 2013; Болотов 2018; Ковалева 2011; Решетникова 2017).

From the historical perspective two reasons gave rise to conducting comparative research in education regularly in early 1950s. Firstly, social scientists were interested in how social interaction in the educational system is arranged and what factors shape student achievement

[1] Trends in International Mathematics and Science Study.
[2] Progress in International Reading Literacy Study.
[3] Programme for International Student Assessment.

among nations. The comparative perspective was essential to study this issue. Secondly, comparative research enjoyed a growing interest among scholars and educators. They saw an opportunity to conduct more profound research on the role of curricula, classroom practices and other factors contributing to educational achievements of students and the system as a whole. Countries were assessed in terms of their educational performance. Studies that involved testing students' knowledge of various subjects constituted a further development of comparative studies in the late 1960s. Educational policy-makers and managers were eager to use the results of these studies to develop their own educational policies (Carnoy 2019; Verger, Parcerisa, and Fontdevila 2019).

The then-founded International Association for the Evaluation of Educational Achievement (IEA) had two objectives in mind from such comparative large-scale assessments: 1) providing educators and educational decision-makers with information about the quality of the educational system in *their* own countries and 2) studying and analyzing differences in educational systems. The IEA emphasized that although the second objective could be useful for educational decision-makers, its main focus was research (Plomp 1998). With these objectives in mind, they conducted several surveys such as FIMS (First International Mathematics Study) in the 1960s and the Second International Mathematics Study in the early 1980s. These were followed by studies in science (the first one was conducted between 1966 and 1973, the second, in 1983-1989) and others (Howie and Plomp 2005; Medrich and Griffith 1992). Since the mid-1990s ILSA were conducted regularly over equal periods of time. That led to an increased interest of educators and educational policy experts and also served another objective: monitoring the development of educational systems (Addey et al. 2017).

At present, ILSA are widely used not only to serve the objectives they were created for (for comparative research on a wide range of issues in education), but also to provide rationales for certain educational policy measures (Hopfenbeck et al. 2018; Lewis 2017; Tuijnman and Bottani 1994; Джуринский 2013). However, there are some pitfalls that underlie using the ILSA with the goal of doing meaningful policy research. Scholars discuss the misuse of comparative studies to serve objectives for which they were not initially intended, both in terms of methodology and content. First, it is becoming a common practice to borrow educational practices from countries with a higher level of performance in ILSA. This leads to an overly simplified understanding of the educational process (Grek 2009; Johansson 2016; Phillips and Ochs 2003). In the case of reforming the educational system, this approach neglects the fact that the countries' cultural characteristics and differences in the educational systems make it unreasonable to transfer pedagogical or systemic practices from one country to another (Medrich and Griffith 1992;

Агранович 2017; Carnoy et al. 2016; Carnoy, Garcia, and Khavenson 2015; Elliott et al. 2018; Fuchs and Wößmann 2007).

Secondly, the test content in ILSA does not always reflect the intended national curriculum and that is why ILSA cannot be used to assess accurately the quality of the educational system, interventions, and other changes objectively and rigorously in terms of methodology (Jerrim et al. 2017; Smith 2002). This is especially relevant regarding PISA tests (Hopfenbeck et al. 2018). This also concerns contextual questionnaires, which are not country-specific (Caro, Sandoval-Hernández, and Lüdtke 2014; Medrich and Griffith 1992).

Thirdly, research focuses on a country's average scores and countries' ranking based on the average scores for each country. It is more appropriate to take into consideration the results for significant parameters that specify the educational system and society in general (Carnoy et al. 2016; Klemenčič and Mirazchiyski 2018). Moreover, it is not always reasonable to examine a decentralized educational system of a country as one educational system. Regions or different groups of schools can differ drastically in terms of performance in ILSA (Carnoy and Rothstein 2013).

Fourthly, relying only on studies of this kind, we cannot make inferences about causal relationships. The only results that can be obtained from these data are so-called *weak inferences* that are informative about correlational relationships and can help to hypothesize about possible reasons for lower or higher results. These data cannot contribute to addressing local issues concerning educational policy or to uncovering 'how' and 'why' when analyzing school performance. They do not provide direct answers concerning the management of the educational system (Elliott et al. 2018; Loveless, Ladd, and Rouse 1998; Plomp 1998; Urick 2018).

Yet, ILSA provide information for revealing areas of concern, for raising research questions and for conducting more profound research (Plomp 1998). To make stronger inferences, we need to conduct additional research that would include ILSA (Aloisi and Tymms 2018; Choi and Jerrim 2016; Medrich and Griffith 1992; Smith 2002).

Two main organizations that conduct large-scale comparative studies have different views on opportunities and limitations concerning the use of the data from these studies to develop educational policy. IEA (the organization that conducts TIMSS and PIRLS) mostly pursues scholarly objectives and does not make any recommendations relying on the research results. It is assumed that countries may develop their own recommendations taking into account not only the research results but also their own context or their additional research results. On the other hand, the Organization for Economic Cooperation and Development (OECD is responsible for PISA) operates on a premise that its research provides information that is directly connected with

the quality of education and the performance of the educational system. This methodologically unwarranted premise that underpins PISA determines the vector of using ILSA nowadays (Carnoy 2019; Grek 2009; Klemenčič and Mirazchiyski 2018; Loveless 2001; Pettersson, Popkewitz, and Lindblad 2017).

Moreover, countries justify their use of ILSA for deriving educational policy in terms of their considerable investment in these expensive and resource-intensive studies as well as in terms of scale and range of data regarding the educational system that they provide. However, when making decisions relying on the research results (that is when inferences concerning casual relationships are made about certain interventions leading to concrete results), it is important to use ILSA data correctly both in terms of methodology and content. There should be as few inaccuracies as possible in inferences and interpretation as that can render ineffective and irrelevant the reform measures taken (Postlethwaite and Leung 2007; Shaxson 2005).

The goal of the thesis is based on the abovementioned contradiction between the initial goals set for ILSA and their current usage. The goal is an elaboration and testing several methodological approaches to the ILSA data analysis allowing generation of empirical evidence that is methodologically correct and relevant for evidence-based policy making.

Theoretical framework of the thesis is driven by the evidence-based educational policy paradigm, also called "governing by numbers" (Borer and Lawn 2013; Grek 2009; McDonald 2010; Shaxson 2005). According to this paradigm research evidence is of central importance in informing education policy required to make intelligent decisions (Robinson and Evans, n.d.). Methodological requirements should be elaborated not only for data collection but also for the incorporating the data, including ILSA, into governing tools for conducting evidence-based policies (Borer and Lawn 2013). A second basis for the theoretical framework is the theory of change that describes the process of elaboration and implementation of the reforms or other initiatives in different spheres. The theory of change explains the process' mechanics and goals: how and why it works. The process can be understood both as a large-scale reform and as a local-scale change at an individual, organizational, or community level (Connell and Klem 2000; Laing and Todd 2015). The theory also assumes that possible reasons, changes and consequences are not unidimensional but indeed multifold. Any research of the reform process should take this and the broader context into account and explore it (Barnett and Gregorowski 2013; Laing and Todd 2015).

This leads to this thesis's research question: what are the approaches for using ILSA data in a more nuanced or sophisticated way for producing relevant and useful for educational policy conclusions?

The main issues that limit using ILSA data when developing educational policy can be roughly divided into several types: 1) statistical and psychometrics issues in processing large-scale research data; 2) the gap between the ILSA data and the context of a concrete educational system; 3) comparability of educational systems of different countries; 4) descriptive data that cannot be used to make inferences about casual relationships.

Statistical and psychometrics issues are beyond the scope of our research. As for eliminating or reducing the influence of the other three issues, we will provide an extensive description of several methodological approaches that we validated in our research (table 1). For example, a *mixed method design* and *comparing TIMSS and PISA results for a country* help to embed quantitative ILSA results in a broader context of the educational system in which these ILSA test results were obtained. *Fixing sample parameters when studying the dynamics of the results* and *comparing regional educational systems in a country* allow to conduct comparative studies in which an opportunity and appropriateness to compare the systems in question are ensured. *Natural experiment*s with a number of conditions helps to approach casual inferences, which, in turn, leads to assessing the effectiveness of some reform measures.

Table 1. The correspondence of the issues that limit using ILSA data and the suggested methodological approaches.

The issues	The methodological approaches
The gap between the ILSA data and the context of a concrete educational system	Mixed method design (papers 3, 4 and 5); Comparing TIMSS and PISA results for a country (papers 1 and 4).
The non-comparability of educational systems of different countries or of different time periods	Fixing sample parameters when studying the dynamics of the results (papers 1 and 4); Comparing regional educational systems in a country (papers 2 and 4).
Descriptive character of the data	Natural experiment (papers 3 and 5).

The methodological approaches that were validated in our research are explained in papers presented for the defense.

Key results in the papers

Before we describe concrete methodological approaches that allow for more effectively integrating ILSA into developing educational policy, it is wise to cover a common methodological principle that most researchers agree on. Any academic achievements and postsecondary educational trajectories are always in some way connected with individual characteristics of students and are largely determined by them (Carnoy and Rothstein 2013; Raudenbush and Kim 2002). In many countries the impact of these extra-school factors is greater

than that of the school (Elliott et al. 2018). It is important to consider these factors when making inferences concerning educational policy, as these characteristics are not subject to educational policy interventions (Aloisi and Tymms 2018; Капуза et al. 2017). If the analysis of the connection between school and institutional factors of academic achievement is carried out without taking relevant individual student characteristics into account, it can lead to a biased assessment of this connection and, consequently, to wrong inferences and reform measures (Cochran 1968; Raudenbush and Kim 2002).

Furthermore, individuals' social context within countries is heterogeneous and the relation of schooling and other factors to student performance can vary across groups and may imply that an educational policy may be effective for one group but not another. When analyzing student academic achievement and factors that contribute to student performance improving or declining, the most important characteristics for division into groups are students' socio-economic family status, the size of the settlement, gender, immigrant status, and, in some countries, ethnicity or race. Results may differ greatly among these groups and analysis of the country's overall performance may conceal differing cultural and social context effects among these groups and their interaction with educational inputs (Carnoy, Garcia, and Khavenson 2015).

Mixed method design

For ILSA, combining quantitative and qualitative methods in one study suggests using statistical analysis of ILSA data together with qualitative methods, such as interviews with educational policy experts, actors in the educational process, analysis of documents, etc. Mixed method design provides an opportunity to analyze the environment and the immediate context in which educational achievements that students show in ILSA are formed, explain them considering the processes that are taking place within a certain educational systems. Besides, mixed method design in education allows to assess the consistency and validity of the results (Chesnut, Hitchcock, and Onwuegbuzie 2018).

The paper (Хавенсон 2019) addresses the national curriculum reform in Russian-medium schools in Estonia and Latvia. The curriculum is seen as a three-layer structure: 1. *the intended curriculum* – that which is stipulated in official documents and what society would like to see taught; 2. *the implemented curriculum* – what is actually taught in the classroom and how teachers integrate the curriculum elements into the educational process; 3. *the attained curriculum* – what students have learned. The three-layered curriculum is used in this study as a lens to examine the process of implementing the new curriculum as a core aspect of integrating Russian-medium schools in the national school system in Latvia and Estonia.

As ILSA provide data on the attained curriculum only, it is impossible to study how close the attained curriculum is to the other two, if we rely exclusively on these data. Thus, as the *intended, implemented* and *attained* curricula have different content, they cannot be studied with only one methodological approach. That was why we used a partially mixed methods concurrent equal status design. This research design suggests that, on the one hand, quantitative and qualitative parts have their own goals and, on the other hand, combining the obtained results allows for meta-inferences. *The intended curriculum* was explored through the analysis of official documents concerning the content and integration of the new national curriculum as part of Russian-medium school reform in Estonia and Latvia. To investigate *the implemented curriculum,* we conducted interviews with teachers and school principals to assess to what extent they implemented the intended curriculum and what their attitudes were. To assess *the attained curriculum,* we looked at trends in educational performance in mathematics in Latvia, Estonia and Russia using PISA for the period from 2006 to 2015 (the reform period). As Baltic countries' intended curriculum is rather PISA-style it seems possible to measure the attained curriculum through PISA test. In (Хавенсон 2019) the qualitative part preceded the quantitative one and was focused on describing the ideas that underlie the Russian-medium school curriculum reform in Latvia and Estonia as well as on describing how these reforms were implemented in the schools' everyday practice. The quantitative analysis of PISA data was aimed at assessing the effect of the reforms' interventions.

The research design (Khavenson and Carnoy 2016) used in this paper was also mixed methods, but the order of the qualitative and quantitative parts was reversed. First, the PISA data analysis allowed us to reveal the trends of result changes during the reform period (2006-2012). Then, relying on interviews with educational policy experts and school principals, we were able to infer how and how much these reforms influenced the change of the results and which of these reform effects were intended and which of them were unintended consequences.

A third paper (Morsy, Khavenson, and Carnoy 2018) deals with ILSA potential to explain the reasons for lower performance of Australian students in PISA and TIMSS. We again used mixed method design where qualitative methods were followed by quantitative ones. We conducted a series of qualitative interviews with officials from the federal and regional ministries of education—individuals involved in developing educational policy as well as experts responsible for conducting and analyzing the ILSA results in Australia. The focus of the interviews was not so much on explaining lower performance of students than on asking these experts why they thought the scores were declining. From this, we developed several hypotheses for the large decrease in student performance, especially in mathematics. These included possible changes in

social context (student social class, student immigrant status), educational context changes, such as in the number of math lessons in the secondary school (time on task), changes in the educational labour market (an increase in the number of teachers without a degree in mathematics), and shifts from public schools to subsidized private schools. These "expert-derived" hypotheses were tested using PISA and TIMSS quantitative data for 2000-2015 and 1999-2015 respectively across Australian states, since both PISA and TIMSS surveys in Australia are randomized within states. TIMSS data provided an opportunity to analyze information about teachers over an almost twenty-year period and to assess whether reasons for lower performance hypothesized by interviewees could be substantiated with quantitative results. Students' characteristics were taken into account when we analyzed the dynamics of PISA results of different socio-economic groups and of students from schools of various types for the country as a whole and across states. They allowed us to compare the timeframe of changes in the educational policy and changes of academic performance. They also helped to check whether the explanations offered by interviewees could have an impact on the academic performance.

In addition to analyzing PISA and TIMSS data, the mixed method design in all these cases helped to extract information that made the quantitative results much more relevant for educational policy. Qualitative methods contributed to a better understanding of the context in which changes and the educational reform were taking place, to formulating hypotheses, to providing a more thorough interpretation and possible explanations of the results that were obtained with the help of quantitative methods. Quantitative analysis of ILSA data allowed to analyze the overall trends in students' achievement in different countries and it also helped to validate several hypotheses that were formulated when qualitative data were collected and documents were analyzed. Therefore, on the one hand, ILSA data were used in accordance with their objectives, on the other hand, they were embedded in a wider context and, thus, their usefulness to develop educational policy increased.

Comparing TIMSS and PISA results for a country

Some countries take part in several ILSA and, thus, they get a more nuanced picture of their educational process. Firstly, studies can be conducted at different levels: PIRLS and TIMSS 4th grade target primary schools, whereas TIMSS 8th grade and PISA – the secondary graduate or high school level (it depends on the country) (Treviño and Órdenes 2017; Tyumeneva 2013; Wu 2010). Secondly, these studies cover a wide range of attainments on the subject as their tests' content focus is different even if, formally, they test achievement in one subject: for example, mathematics or science in TIMSS and PISA. The difference is that the TIMSS test is curriculum-based and measures what students have learned at school, whereas PISA is literacy-based and

captures how well students are ready for an adult life (Wu 2010; Klieme 2016). Asian and Eastern European countries tend to have higher TIMSS scores than PISA. While Western European nations show the opposite trend (Wu 2010). Consequently, on the one hand, it is arguable that data from different ILSA measure the quality of school education in the country (Klieme 2016; Loveless 2001). On the other hand, results from several ILSA allow for more profound and multidirectional inferences for educational policy (for example, Jakubowski 2010; Jakubowski and Pokropek 2015; Sollerman and Pettersson 2016; Treviño and Órdenes 2017).

The Australian paper (Morsy, Khavenson, and Carnoy 2018) compares Australian results for TIMSS and PISA to reveal a better picture of changes in Australian students' performance. Australia takes part in both studies over the whole period in which these tests have been conducted. The researchers found it important that the dynamics of achievement in mathematics differed over the period under consideration: from 1999 (TIMSS) and 2000 (PISA) to 2015. Initially, both TIMSS and PISA results were high. However, PISA results declined steadily over the whole period while a downward trend for TIMSS results was registered only from 1999 to 2007. In 2007-2015, Australian student performance on TIMSS increased. The content of the two tests as well as the samples are different. They also correspond to the school curriculum differently. Considering these three factors, we managed to focus our analysis on different aspects and identified a bigger number of reasons that could explain one set of results, but not the other. It is worth mentioning that PISA is more prominent in Australia for educational decision-makers and for the mass media. This is due to the fact that the OECD focuses on developing recommendations to improve educational performance and the test content reflects skills that are relevant for the 21st century. TIMSS tests the school curriculum on a certain subject and, thus, is seen as an international measure of what academic curriculum experts believe should be taught in schools in mathematics and science in a particular grade level. In mathematics, for example, the 8th grade TIMSS evaluates standard sub-subject matter such as number problems, algebra, geometry, and statistics.

Our research in which we used data for Russia (Carnoy, Khavenson, and Ivanova 2015) shows that data of different ILSA for the secondary and high schools in Russia diverge even further than in Australia. TIMSS results in the 8th grade are consistently high while PISA results are consistently low in comparison with other countries. Further, the PISA results for Russian 15 year-olds show little positive trend except in the latest PISA cycle we covered—2009 (this positive trend continued in 2012 and 2015). Thus, Russian students and system as a whole are more "successful" in one type of tested achievement, but less "successful" in the other.

The comparative study of the dynamics in some European countries and Russia (Carnoy, Khavenson, and Ivanova 2015) for the periods of 1999-2011 (TIMSS) and of 2000-2009 (PISA) illustrated that the available results of the two studies in different countries differ and show more or less divergent trends. A detailed analysis of these trends helped us draw a series of inferences regarding the most appropriate ILSA to use in developing educational policies to improve student outcomes: 1. Which test content reflects the intended curriculum? and 2. Which measure of achievement is more important? (the latter question is particularly relevant, if the reform intends to change the content of what is taught in schools). On the other hand, in order to assess the current educational system, it is advisable to consider which survey better reflects current school curriculum. For most East European countries and Russia, TIMSS better measures what is taught in the school curriculum. This does not mean that these countries should not participate in PISA, but it is important to understand that accounting for PISA results with school factors may be more difficult than in the case of TIMSS. Indeed, the structure of the PISA survey—no teacher questionnaires; lack of linkage between students in the sample and their teachers (PISA is a school based sample; TIMSS is a school based sample with a classroom sample clustered in each school) is such that explaining PISA student outcomes with classroom factor variation is not possible.

Therefore, data on educational performance in the various studies I have presented here allows me to state the following: 1. They enhance the understanding of educational processes and increase the validity of the inferences; 2. They allow for a more nuanced multidimensional analysis. These various dimensions should be taken into account when developing educational policy, since such policy often depends on which measure of achievement the policy targets. Comparing ILSA results and national tests increases the information pool and helps to embed ILSA in the system to assess the quality of education (Carnoy et al. 2015; Jakubowski and Pokropek 2015; Sollerman and Pettersson 2016; Treviño and Órdenes 2017).

Fixing sample parameters when studying the dynamics of the results

As ILSA results are often used to study the dynamics of educational performance, it is necessary to remember that the sample may change over time concerning some characteristics of students, teachers and schools (for example, the proportion of children with low socio-economic status or of certain type of schools) may vary across years. This is especially relevant if the effectiveness of certain interventions is assessed relying on the dynamics of the results. The study (Aloisi and Tymms 2018) showed that sample changes concerning socio-economic status and demographic characteristics from one cycle to another have a larger impact on PISA results than the reform of the school curriculum. The change in the configuration of variables connected with the results

(gender, migration status, parents' education and occupation, the number of books at home) from one PISA cycle to another in Germany, France, Sweden and the UK significantly changed the inferences about educational inequality and its level, which several reforms aimed to reduce (Lenkeit, Schwippert, and Knigge 2017). Taking into account the dynamics of socio-demographic characteristics in the sample of some US states in TIMSS helped to correctly identify educational systems that were experiencing growth or decline of results (Carnoy and Rothstein 2013).

In our studies (Carnoy, Khavenson, and Ivanova 2015) and (Morsy, Khavenson, and Carnoy 2018) we fixed the results in PISA and TIMSS as if the sample demographics did not change over time. To do this, each year's results were weighed by socio-economic status in the control year: in this case it was 2000 in (Carnoy, Khavenson, and Ivanova 2015) and 2012 in (Morsy, Khavenson, and Carnoy 2018).

The Australia PISA results adjusted for the 2012 sample composition showed the same trends, but they helped to specify the degree of the changes, because they turned out to be lower than the officially reported average scores. That means that the explanation for lower performance over the period in question is not only the decline of the quality of education, but also the sample demographic changes. Over time, the samples were characterized by higher proportions of lower socio-economic status students, and these usually perform more poorly than others.

In a comparative study of Russia and several European countries, we fixed the sample not only across survey years, but also among countries. Each country's results were weighed by Russia's socio-economic distribution in the year 2000, and that allowed us to estimate 'purer' results as we adjusted for the effect on PISA scores of Russia's students' low socio-economic status compared to other European countries.

Fixing the sample parameters when studying the dynamics of results helps to achieve a better comparability of potential impact of the educational system and to take into account the effect of the change of characteristics that are not influenceable by educational policy. Therefore, the identified differences in the test scores can be more directly attributed to educational policy changes (holding other factors constant). Yet, in some cases family socio-economic status can be influenced by education policy, e.g. changes in school admission rules or general enrollment interventions. In such cases these interventions can be of a special interest and specific efforts for studying them can be undertaken.

Comparing regional educational systems in a given country

Considering a high impact of cultural characteristics on the way the educational system functions (Fuchs and Wößmann 2007) and different correspondence of ILSA tests to national curricula, comparison across countries may not provide as valuable information relevant for the educational policy, as comparing results within a country. Within country comparisons may be more informative for planning valid reforms. In fact, we may have considerable variation among educational systems within a country: for example, education in different regions in a country with a decentralized (federal) educational system, or education that targets various language groups in a multinational state with several state languages. It is obvious that this situation warrants an educational policy that takes into account these differences. Differences between regions or language groups are less significant than between countries (Carnoy and Rothstein 2013; Postlethwaite and Leung 2007). Consequently, comparing regions or groups of schools within the country, even with a very decentralized education system, the inferences made are more relevant for these national context and education system, and therefore it may allow for drawing more relevant inferences for educational policy in that country. Besides, with the sample that is representative for various regions, one can perform analysis taking into account these regional characteristics (Hippe, Jakubowski, and Araújo 2018). Borrowing educational practices may be more effective if such borrowing is among reforms that have been shown to have positive effects in the same country, especially if there are large differences in student achievement performance over time (corrected for socio-economic differences) in some regions/states than others (Carnoy, Garcia, and Khavenson 2015).

When participating in ILSA, Australia draws not only the national sample, but also a sample that is representative for each of the eight Australian states. This kind of data for PISA is available for all cycles from 2000 to 2015 and for TIMSS cycles in 2011 and 2015. In the Australian paper (Morsy, Khavenson, and Carnoy 2018), where Australian data is analyzed to identify the reasons for declines in educational performance, we could check experts' hypotheses against data for different states with different socio-demographic characteristics and with trends concerning the changes of a teacher's profile, etc. The decline in student performance differed in different states. In two Australian states – New South Wales and Victoria – the dynamics of PISA results and the trend of inequality declining were different: student performance declined more and was characterized by greater inequality in results across social class groups over time in New South Wales than in Victoria. In the period under discussion, New South Wales saw the fall in the performance of students with low and high socio-economic status, while in Victoria the performance of students with low socio-economic status increased by 2015. The national data

would have concealed these trends and that would reduce the relevance of inferences and validity of explanations of potential reasons for declines.

In another paper (Carnoy et al. 2017), we analyzed the results of the Brazilian National Evaluation System of Basic Education (SAEB), using the data from this large Latin American country with a federal political structure and decentralized educational system. We showed that inter-state comparison can help to explain the reasons underlying the results and to suggest educational policies that may be effective in improving the quality of education. Characteristics taken into account were the state where the school was located and where students lived, and cooperation between the municipal and state administrative levels of education.

Our studies with ILSA data and regional/state data showed three things. First, for comparative studies in education, it is important to provide rationalization for comparison at the national, regional, municipal levels, etc. The larger and more culturally diverse the units of comparison, the less reliable are the inferences that we can draw regarding educational policies, and the more difficult it is to transfer effective educational policy in one unit's context in order to increase the quality of education in another unit. Secondly, intranational comparison in federal countries are more theoretically sound as controlling the implementation of the educational process and often of the educational content is the responsibility of regions. Consequently, regions play an important role in influencing the quality of education. Thirdly, the advantage of intranational comparisons is that regions have much in common at the national level. For example, they function in similar conditions in terms of school funding, the teaching labour pool, cultural and national peculiarities in and outside schools. This is not the case for comparing the educational systems internationally. In terms of methodology, it is easier to identify effective educational practices and reform measures when there are fewer contextual differences. That is why intranational comparisons can focus on local issues of managing the educational system.

Despite the fact that Russia's educational system is centralized and regional differences concerning the curriculum are not as great as in Brazil and particularly the U.S., regions differ socio-economically. That is why, when analyzing the changes or when planning them, considering regional characteristics would help to adjust educational policy to a particular educational environment (Yastrebov et al. 2014; Агранович 2011).

Natural Experiment

The papers (Хавенсон 2019) and (Khavenson and Carnoy 2016) exploited the condition of a natural experiment that occurred after the collapse of the Soviet Union. At that time, every ex-Soviet country abandoned one educational system and conducted its own reforms and achieved results that differ from those of other countries. In addition to that, Latvia and Estonia had two

educational systems after leaving the Soviet Union – schools with the Latvian or Estonian language as the medium of instruction and Russian-medium schools. Reforms in the former started right after the collapse of the USSR, while in the latter – much later and very quickly. This historical external "jolt" provided an opportunity to apply the natural experimental approach.

When studying the effect of a reform it is problematic to assess how much the interventions implemented during the reforms are connected to the changes of student performance on a test. The reform effect is difficult to differentiate from other processes that occur during that period. Interventions are introduced gradually and that blurs the whole picture; there often appear other reasons that make it difficult to find the connection between reform and educational performance. But historical conditions and the way the reform was implemented in Russian-medium schools in Baltic countries allowed us to compare Latvia's and Estonia's educational systems with the system in Russia, on which Russian-medium schools base their curriculum and teaching practices. The paper by Khavenson and Carnoy (2016) analyzed students' scores in two types of schools in Latvia, and Estonia, plus schools in Russia using the PISA data for 2006-2012 (the period right after or during the reforms). Natural experiment methodology helped to study the reform of the educational system in the two countries as well as to assess the effect these interventions produced (both those that were planned when developing the reform and those that were unintended). The paper (Хавенсон 2019) also analyzed the PISA data for the same groups for 2006-2015, but it focused on which curriculum changes in Russian-medium schools brought about improving students' performance and how long-term these effects were.

On the one hand, historical events that allowed to conduct this kind of experiment do not happen often. On the other hand, this methodology can be applied when reforms that differ content-wise or local changes of educational practices are implemented in regions of one country but not another. As experimental research is largely not applicable in educational situations, especially comparative ones, natural experiment situations should be exploited when possible. ILSA in this case serve as a regular measurement of educational achievements. Their dynamics can serve as a measure that records the reform effect, if other methodological principles are observed.

Main findings:

The main principles of applying ILSA to develop the educational policy can be summarized in the following:

Utilizing ILSA data with the suggested methodological approaches allows broadening the list of policy issues that can be covered within evidence-based policy paradigm.

The mixed method design, where one of the elements is ILSA data analysis and the other is qualitative methods, contributes to accounting for the context of the educational process and to developing explanatory models for the changes with provision for the features of a certain educational system. That means extracting information which is more relevant for educational policy than the analysis of PISA and TIMSS data only.

Comparing students' achievements in various ILSA and in national tests in one country allows for a multidimensional analysis of the results with provision for the features of each test, their correspondence to the curriculum, and their connection with the reforms.

Accounting for sample changes concerning the socio-economic status from one cycle to another in ILSA helps to achieve greater comparability of the results when studying long-term trends in ILSA achievements.

ILSA data analysis on samples that are representative for a country's regions provides an opportunity to conduct comparative studies that can help assess the reform effects or reasons for changes of student results in a given national educational system and to avoid challenges that are typical for inter-nation comparisons. As a result, this analytical approach that can help produce more locally relevant educational policies (for example, region-specific, targeting various language groups or types of schools (if they are relatively autonomous)).

Incorporating ILSA data in experimental research in natural experiment situations allows for inferences to some extent closer to causal ones that are made in cross sectional design research such as ILSA.

Substantive results of each paper in which the methodological approaches were elaborated and tested

Although in my thesis I focus on methodological approaches for using ILSA to inform evidence-based education policy, it seems relevant to shortly describe the substantive findings that were obtained in the published research.

1. In (Carnoy, Khavenson, and Ivanova 2015) we analysed Russian and its neighbours data in PISA and TIMSS. Our focus was on Russia, where students do relatively well on the TIMSS mathematics test but not on the PISA, and which has had a good reputation in teaching students mathematics. We attempt to understand why Russian students do not score well on the PISA and use the results of our analysis to draw some tentative educational policy lessons.

The message is indeed mixed on how well Russian students do in mathematics. If their TIMSS score is a good measure of how well they are learning mathematics, they are performing quite

well at all levels of family academic resources, better than their counterparts in most neighbouring countries. If their PISA performance is a good measure of their mathematics skills, Russian advantaged and middle family-resource students are indeed performing relatively poorly and advantaged Russian students are making little progress in improving how much they learn when measured by the PISA definition of maths knowledge.

Through this comparison we draw an attention that ILSA cannot be considered as a quality measures as students can achieve differently in different tests. For Russian policy-makers it is warning that education should not be driven by test scores in ILSA. Interventions and policy changes can be reflected differently in two types of tests.

Implementing the approach of fixing the sample by Russian SES distribution allowed first to control for changes in the sample from this point of view. And second, what is even most important to take the differences in students' SES into consideration. If students tested in various countries live, on average, in family environments that differ considerably, comparisons of average student performance could incorrectly attribute outcomes to educational policies when they may be the result of differing outside-of-school influences. Furthermore, educational policies may affect students from different environments differently. By comparing the academic performance of students in particular family and social environments over time, we can better understand the nuances of educational policies in various countries. Such comparisons are the core of our analysis in this study.

This also allowed unravelling unusual pattern of Russian PISA test scores differences in high and low SES groups. The smaller gaps in other countries on the PISA are mainly the result of high scores of students in the most disadvantaged groups, but in Russia, the small gap is mainly the result of relatively low scores for advantaged students. In contrast, Russia's disadvantaged students perform similarly compared to disadvantaged students in a number of other countries.

2. In (Carnoy et al. 2017) we studied one of the feature of the political organization of Brazilian education – the division between the state and municipalities in administering schooling – and how state policies mediating this division might have affected student performance in each state. We also focused on the possible differences in the "effectiveness" of state education administrations in delivering education. We measured state effectiveness by students' mathematics achievement gains on a national test in 1999–2013. We also examined the possible reasons why gains differ greatly in states with similar demographic characteristics.

The results of our analysis showed that Brazilian students' achievement in the ninth grade adjusted for student and school socioeconomic differences, as well as some key school resource differences, varies considerably across states and, within some states, across the separate state

and municipal school systems. Successive cohorts of students in a number of states have greatly increased their mathematics scores on the national SAEB test in 2001–13 and 2003–13. At the same time, successive cohorts of students in other states' state and municipal systems have seen their adjusted scores stay level or decline.

From the methodological perspective we made an argument in this article that in comparative education analysis, there is persuasive support in political theory to consider subnational state comparisons in federalist nations and that such comparisons can yield valuable insights for improving education in the federal nation-state as a whole. We showed how analyzing the comparative "effectiveness" of state education systems within one such federalist country, Brazil, in the first decade of the twenty-first century (1999-2013) can yield important insights into educational change.

Our research had important implications for comparative education research. It suggested that in searching for the reasons that students in some political entities (such as nation-states) score higher on tests than in others, it is crucial to provide a political/administrative justification for comparing students across such entities, whether nations or school districts. Intranational comparisons in federalist countries make theoretical sense, since subnational political units have major juridical responsibility for delivering education. Our research also suggested that there are important benefits of intranational comparisons. A main one is that they can be less concerned with differences in overall educational systems and their subcomponents, such as teacher labor markets, financing arrangements, and educational culture inside and outside schools, since these are much more similar in intranational comparisons than in international comparisons of national educational systems. Intranational comparisons can therefore focus more on educational management issues or particular subnational state led interventions, since these still tend to vary considerably among subnational units in federal systems. Methodologically, it is thus easier to identify effective educational policies and practices when the contextual variations in which those policies and practices occur are greatly reduced.

3. In (Khavenson and Carnoy 2016) we scrutinised the process of planning and implementing the educational reforms as well as their intended and unintended academic effects through the lens of compensatory legitimation theory in Post-Soviet Estonia, Latvia and Russia.

We found that relative to students in Russia, Russian-medium students in the Baltics made significant gains in the PISA test. In Latvia, these appear to be an unintended effect of somewhat 'softened' state language policies, the conditions surrounding minority rights, and the general context of maintaining social cohesion. In Estonia relative gains of Russian students appeared to be an intended effect of locally grown educational (and language) policies and increased, more

effective cooperation with Russian medium schools to further improve PISA performance in a relatively high scoring, PISA-focused country. The education of Russian students in Russia was also subject to compensatory legitimation reforms in this period, but based on a review of these reforms and interviews with policy makers, principals and teachers in Russian schools, curriculum and teaching changes have been very limited compared to those introduced indirectly by language policies in Latvia and more directly to raise student achievement (as measured by PISA) in Estonia.

We do not claim to make direct links between compensatory legitimation policies in Russia and the Baltics and patterns of student achievement gains in Russian medium schools in the three countries studied. Yet, we can make a compelling case that the policy of imposing Latvian language requirements on Russian medium schools in the 1990s and 2000s as part of a larger effort to legitimate the state had an 'unintended' and large impact on Russian students' achievement gains because of the positive effects of bilingual education on learning.

One lesson to learn here is that states enact educational reforms that are consistent with a broader goal of state legitimation, and even if the reforms have little direct relation to raising student achievement outcomes, they may still improve student learning. A second already well-known lesson is that if the intended effect of the compensatory legitimation reform is to raise student achievement on a particular type of test, the best way to do so is to link instruction to the test. A third lesson, illustrated by the Estonian case (but also potentially applicable to Latvia), is that this linkage is easier to accomplish in a small education system than in a large one, since accepting and implementing instructional reforms school by school is the key to successfully raising test scores.

Such in-depth analysis was able to be performed due to historical condition allowed to implement natural experiment methodology. This helped to link the changes in school education with the changes in PISA test scores in all subjects. Mixed method approach was used to measure the planned interventions and their implementation in everyday school process.

4. In (Morsy, Khavenson, and Carnoy 2018) we examined the educational experts' opinion on the reasons for the pervasive decline in the Australian PISA and TIMSS scores. We then used PISA and TIMSS test scores and students and teachers data to test whether our informants' explanations were supported empirically. We were aided in our analysis by the federal nature of Australia's education system (the jurisdiction of education is in the hands of the states, although the federal government has considerable influence over educational policy). Australian students' PISA scores in mathematics have fallen steadily since their relatively high performance the first time the test was applied, in 2000. The decline persists when students' performance is adjusted

for changes in the family academic resources of students sampled, and we observe it for students in all Australia's eight states, for both socially advantaged and disadvantaged students, and in both government and private schools, although the trends vary somewhat from state to state and in different types of schools. The declines in scores are large enough that they should be quite easy to explain, and there is no shortage of reasons given by Australian experts for why Australian mathematics education (and Australian education more generally) is getting worse.

However, the bunch of methodological approaches allowed us to perform more nuanced analysis and demonstrate that these reasons largely failed to stand up to empirical scrutiny. Indeed, there were some surprises. For example, many experts in Australia assume that student performance in government schools has worsened more than in private schools, and that government school test score decline is pulling down the national average. Our estimates showed that, adjusting for students' family academic resource differences, students in Catholic schools had the largest decline in mathematics scores in 2003–2015. The decrease in time spent on mathematics appears to be real, at least according to PISA students' reports. However, according to our estimates, the smaller reported exposure to mathematics has only a small effect on average PISA mathematics scores. Taken together, our assessment of the teacher quality and math time explanations suggests that the "quality" of mathematics teaching in Australia may have, in fact, declined somewhat (less time on math), but the estimated effects are small. And at least in some states, teaching quality may have even improved because of state policies.

All this shows that despite the considerable insight that Australia's educational experts have into what is occurring in Australian schools, many of their views are off the mark, especially in explaining the reasons for the decline in Australia's PISA scores. Misidentifying these reasons is not benign. For example, based on the notion that government schools are leading the drive downward, some politicians are suggesting funding cuts for government schools, when, given our results, precisely the opposite policy—to increase their funding—may be appropriate.

5. In (Хавенсон 2019) the process of curriculum reform in Russian-medium schools in Latvia and Estonia is investigated. The research focused on whether those curriculum reforms were successful from the perspective of schools' interiorisation of new curriculum and PISA performance improvement.

There is a common view that a curriculum has three layers: the intended curriculum – 'what society would like to see taught'; the implemented curriculum – what is actually taught in the classroom and how teachers bring all the curriculum's elements into play; and the attained curriculum – what students have learned (Livingstone et al. 1986).

The research question of our study was whether integration has been more or less achieved or at least whether the gap between the three levels of curriculum has been narrowing since the Russian-medium schools reform began. To measure all three layers mixed method design was implemented as a methodological approach. A series of in-depth interviews in Russian-medium schools, in conjunction with the PISA 2003-2015 trends analysis, were conducted.

To be able to measure the extent to which the intended curriculum was implemented in schools and was attained by students in Latvia and Estonia, the study used the natural experiment that occurred after the break-up of the Soviet Union. During the Soviet period, considerable effort was directed towards the unification of the educational systems in the 15 constituent republics. By the end of the 1980s, this had generally been accomplished. Education systems were quite similar over all parts of the USSR (Herbst and Wojciuk 2017; Mitter 1992).

As newly independent states started building their own educational systems in the 1990s, a situation of natural experiment arose, in which initially similar groups began to live in different circumstances and under different transformation processes. While the Russian Federation largely maintained its previous curriculum standards, Latvia and Estonia changed their national curricula quite substantially with an orientation to European integration. We believe that this approach allowed us to answer more precisely the question of how post-socialist transformations can appear, giving examples of such transformation and describing the ways they have been approached. Using Russia as a reference country allows for the attribution of national academic achievement (attained curriculum) with respect to the reform measures (intended curriculum) and to the process of the new curriculum implementation (implemented curriculum).

We concluded that intended, implemented and attained curricula are drawing closer together. The intended curriculum described in official documents is clearly reflected in everyday school practices in Russian medium schools. Schools actively implement new-style teaching, including expansion of tasks aimed at developing competences related to application and reasoning, functional reading, active learning and outside-school classes, individualisation and respect for students, among others. At the same time, PISA performance has been constantly improving, showing that the attained curriculum is approaching that intended.

The study has also shown that a positive emotional background facilitates implementation of the reform intentions. There is some discrepancy in this process between Estonia and Latvia. In Estonia, intensive interventions spanned a shorter period, focusing on particular teaching approaches and curriculum elements, and educational officials spent more time and effort to get school principals and vice-principals on their side. According to the interviews, the curriculum transformations were more positively accepted by school principals and teachers in Estonia than

those in Latvia. The former have interiorised the proposed changes to a greater degree. In Latvia, the reforms were implemented in schools as in Estonia. However, there seemed to be more initial resistance there. School teachers and principals did not feel like active participants in the ongoing reforms even when they agreed with the new approaches. The quality and the depth of education reform implementation strongly depend on whether all actors accept this new wave. Even though this is well known, this step is often skipped in the planning of reforms. Involvement of all actors in the reform can facilitate the process and make it run more smoothly, ultimately saving resources in a broad sense (Khavenson 2018).

Conclusion

ILSA design hinders their direct use to develop evidence based educational policy. The thesis describes the possible pitfalls which can be caused by the non-critical utilizing of ILSA data, both for policy evidence and for substantive inferences. It is necessary to work out special methodological approaches in order to integrate these studies into the system of assessing the quality of education, elaborating recommendations with reference to them, or to inform educational policy. The studies we discuss here that used ILSA data and a combination of methodological approaches allowed for results that are more relevant for educational policy than simply drawing inferences from the analysis of ILSA data themselves. The results of the studies discussed made it possible not only to analyze the trends of educational achievements taking place in the educational systems in Russia and other countries, but also to determine their reasons and possible changes that are crucial for achieving this or that result, while still operating within the paradigm of evidence based educational policy. The methodological approaches that have been described are of universal inter-nation character and can be applied to national monitoring of the quality of education.

In Russia, ILSA can be even more useful as a tool of the national system to assess the quality as compared to their inter-nation comparative potential. First, Russia’s educational system has a large number of unique features which hinders the search for relevant reference countries to conduct comparative research. Secondly, its regions differ greatly concerning a number of characteristics and conducting comparative studies within Russia will allow us to assess the effect of a large number of factors. It is also important to point out that besides a large amount of various data on the educational system and the context in which it functions, countries’ involvement in ILSA develops the infrastructure of conducting studies and analyzing their results not only in educational policy research, but also in sociology, psychology, educational economy (Addey et al. 2017; Johansson 2016; Porter and Gamoran 2002). Russia lacks this kind of infrastructure and, especially, staff (Джуринский 2013).

No doubts all these approaches have their limitations which either bound their scope of usage or weaken the results. First, none of them allow to make pure causal inferences. Even natural experiment which eliminates possible confounders to some extent does not guarantee that all experimental conditions were held. Mixed method design heavily depends on the choice and access to experts and subjected to all the pitfalls of qualitative data analysis. Fixing the sample can be problematic when changes in socio-economic status or other out-of-school variables were intentionally managed by the policies. In this case they become a matter of substantive analysis rather than a matter of control for the sample amendments. Access to the regional data and to the different national or international tests do broaden the scope of possible analysis. However, the first are not widely spread across the countries participated in ILSA. And to some extent are subjected to the same pitfalls as between-countries studies requiring for the justification of the feasibility of comparisons. The latter can assume going deeper in pedagogical practices or national curricula to identify which part of the students' knowledge is captured by different tests. There are other limitations that have to be considered when using suggested approaches. However, their positive investment in widening the usage of ILSA and making it more transparent and relevant for education policy overweighs the limitations. The latter are the subject of future research rather than restrictions.

All these studies I have discussed helped to develop approaches to using ILSA to analyze educational policy more effectively and obtain more relevant analytical information. Consequently, they expand these studies' application: from being solely descriptive to providing explanatory inferences or even measuring effectiveness of changes and allowing us to make suggestions about necessary interventions or broadly inform educational practitioners.

References:

Addey, Camilla, Sam Sellar, Gita Steiner-Khamsi, Bob Lingard, and Antoni Verger. 2017. 'The Rise of International Large-Scale Assessments and Rationales for Participation'. *Compare: A Journal of Comparative and International Education* 47 (3): 434–52. https://doi.org/10.1080/03057925.2017.1301399.

Aloisi, Cesare, and Peter Tymms. 2018. 'PISA Trends, Social Changes, and Education Reforms'. *Educational Research and Evaluation* 23 (5–6): 180–220. https://doi.org/10.1080/13803611.2017.1455290.

Barnett, C., and R. Gregorowski. 2013. 'Learning about Theories of Change for the Monitoring and Evaluation of Research Uptake'. 14. IDS Practice Paper in Brief. Brighton: Institute of Development Studies. https://opendocs.ids.ac.uk/opendocs/bitstream/handle/123456789/2995/PP%20InBrief%2

014%20FINAL.pdf;jsessionid=78C1F0CA3E3DA2E7AE3D2E8A96CD933C?sequence=1.

Borer, Valérie Lussi, and Martin Lawn. 2013. 'Governing Education Systems by Shaping Data: From the Past to the Present, from National to International Perspectives'. *European Educational Research Journal* 12 (1): 48–52. https://doi.org/10.2304/eerj.2013.12.1.48.

Carnoy, Martin. 2019. *The Transformation of Comparative and International Education, 1965-2015*. Stanford, California: Stanford University Press.

Carnoy, Martin, Emma Garcia, and Tatiana Khavenson. 2015. 'Bringing It Back Home. Why State Comparisons Are More Useful than International Comparisons for Improving U.S. Education Policy'. 410. EPI Briefing Paper. Washington, D.C.: Economic Policy Institute. http://www.epi.org/publication/bringing-it-back-home-why-state-comparisons-are-more-useful-than-international-comparisons-for-improving-u-s-education-policy/.

Carnoy, Martin, Tatiana Khavenson, Izabel Fonseca, Leandro Costa, and Luana Marotta. 2015. 'Is Brazilian education improving? Evidence from PISA and SAEB'. *Cadernos de Pesquisa* 45 (157): 450–85.

Carnoy, Martin, Tatiana Khavenson, and Alina Ivanova. 2015. 'Using TIMSS and PISA Results to Inform Educational Policy: A Study of Russia and Its Neighbours'. *Compare: A Journal of Comparative and International Education* 45 (2): 248–71. https://doi.org/10.1080/03057925.2013.855002.

Carnoy, Martin, Tatiana Khavenson, P. Loyalka, W. H. Schmidt, and A. Zakharov. 2016. 'Revisiting the Relationship Between International Assessment Outcomes and Educational Production: Evidence From a Longitudinal PISA-TIMSS Sample'. *American Educational Research Journal* 53 (4): 1054–85. https://doi.org/10.3102/0002831216653180.

Carnoy, Martin, Luana Marotta, Paula Louzano, Tatiana Khavenson, Filipe Recch Franca Guimaraes, and Fernando Carnauba. 2017. 'Intranational Comparative Education: What State Differences in Student Achievement Can Teach Us about Improving Education—the Case of Brazil'. *Comparative Education Review* 61 (4): 726–59. https://doi.org/10.1086/693981.

Carnoy, Martin, and Richard Rothstein. 2013. 'What Do International Tests Really Show about US Student Performance'. *Retrieved from Economic Policy Institute Website: Http://Www. Epi. Org/Publication/Us-Student-Performance-Testing*. http://www.aboutweston.com/2013international-testsperformance.pdf.

Caro, Daniel H., Andrés Sandoval-Hernández, and Oliver Lüdtke. 2014. 'Cultural, Social, and Economic Capital Constructs in International Assessments: An Evaluation Using

Exploratory Structural Equation Modeling'. *School Effectiveness and School Improvement* 25 (3): 433–50. https://doi.org/10.1080/09243453.2013.812568.

Chesnut, Colleen E., John H. Hitchcock, and Anthony J. Onwuegbuzie. 2018. 'Using Mixed Methods to Inform Education Policy Research'. In *Complementary Research Methods for Educational Leadership and Policy Studies*, edited by Chad R. Lochmiller, 307–24. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-93539-3_15.

Choi, Álvaro, and John Jerrim. 2016. 'The Use (and Misuse) of PISA in Guiding Policy Reform: The Case of Spain'. *Comparative Education* 52 (2): 230–45. https://doi.org/10.1080/03050068.2016.1142739.

Cochran, W. G. 1968. 'The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies'. *Biometrics* 24 (2): 295. https://doi.org/10.2307/2528036.

Connell, James P, and Adena M Klem. 2000. 'You Can Get There From Here: Using a Theory of Change Approach to Plan Urban Education Reform'. *Journal of Educational and Psychological Consultation* 11 (1): 93–120.

Elliott, Julian, Lazar Stankov, Jihyun Lee, and Jens F. Beckmann. 2018. 'What Did PISA and TIMSS Ever Do for Us?: The Potential of Large Scale Datasets for Understanding and Improving Educational Practice'. *Comparative Education* 0 (0): 1–23. https://doi.org/10.1080/03050068.2018.1545386.

Fuchs, Thomas, and Ludger Wößmann. 2007. 'What Accounts for International Differences in Student Performance? A Re-Examination Using PISA Data'. *Empirical Economics* 32 (2–3): 433–64. https://doi.org/10.1007/s00181-006-0087-0.

Grek, Sotiria. 2009. 'Governing by Numbers: The PISA "Effect" in Europe'. *Journal of Education Policy* 24 (1): 23–37. https://doi.org/10.1080/02680930802412669.

Herbst, Mikołaj, and Anna Wojciuk. 2017. 'Common Legacy, Different Paths: The Transformation of Educational Systems in the Czech Republic, Slovakia, Hungary and Poland'. *Compare: A Journal of Comparative and International Education* 47 (1): 118–32. https://doi.org/10.1080/03057925.2016.1153410.

Hippe, Ralph, Maciej Jakubowski, and Luísa Araújo. 2018. 'Regional Inequalities in PISA: The Case of Italy and Spain'. *Luxembourg (Luxembourg): Publications Office of the European Union*.

Hopfenbeck, Therese N., Jenny Lenkeit, Yasmine El Masri, Kate Cantrell, Jeanne Ryan, and Jo-Anne Baird. 2018. 'Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment'. *Scandinavian Journal of Educational Research* 62 (3): 333–53. https://doi.org/10.1080/00313831.2016.1258726.

Howie, Sarah, and Tjeerd Plomp. 2005. ‘International Comparative Studies of Education and Large-Scale Change’. In *International Handbook of Educational Policy*, 75–99. Springer. http://link.springer.com/chapter/10.1007/1-4020-3201-3_4.

Jakubowski, Maciej. 2010. ‘Institutional Tracking and Achievement Growth: Exploring Difference-in-Differences Approach to PIRLS, TIMSS, and PISA Data’. In *Quality and Inequality of Education*, edited by Jaap Dronkers, 41–81. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-90-481-3993-4_3.

Jakubowski, Maciej, and Artur Pokropek. 2015. ‘Reading Achievement Progress across Countries’. *International Journal of Educational Development* 45 (November): 77–88. https://doi.org/10.1016/j.ijedudev.2015.09.011.

Jerrim, John, Luis Lopez-Agudo, Oscar D. Marcenaro-Gutierrez, and Dominique Shure. 2017. ‘What Happens When Econometrics and Psychometrics Collide? An Example Using the PISA Data’. IZA Discussion Paper 10847. IZA – Discussion Paper Series. Bonn, Germany: IZA Institute of Labor Economics.

Johansson, Stefan. 2016. ‘International Large-Scale Assessments: What Uses, What Consequences?’ *Educational Research* 0 (0): 1–10. https://doi.org/10.1080/00131881.2016.1165559.

Khavenson, Tatiana. 2018. ‘Postsocialist Transformations, Everyday School Life, and Country Performance in PISA: Analysis of Curriculum Education Reform in Latvia and Estonia’. In *Comparing Post-Socialist Transformations: Purposes, Policies, and Practices in Education*, edited by Maia Chankseliani and Iveta Silova, 85–103. Oxford Studies in Comparative Education. Oxford, UK: Symposium Books. http://www.symposium-books.co.uk/bookdetails/104/#394.

Khavenson, Tatiana, and Martin Carnoy. 2016. ‘The Unintended and Intended Academic Consequences of Educational Reforms: The Cases of Post-Soviet Estonia, Latvia and Russia’. *Oxford Review of Education* 42 (2): 178–99. https://doi.org/10.1080/03054985.2016.1157063.

Klemenčič, Eva, and Plamen Vladkov Mirazchiyski. 2018. ‘League Tables in Educational Evidence-Based Policy-Making: Can We Stop the Horse Race, Please?’ *Comparative Education* 54 (3): 309–24. https://doi.org/10.1080/03050068.2017.1383082.

Klieme, Eckhard. 2016. ‘TIMSS 2015 and PISA 2015 How Are They Related on the Country Level ?’ Working Paper. DIPF Working Paper. German Institute for International Educational Research (DIPF). https://www.degruyter.com/view/j/bd.1992.26.issue-4/bd.1992.26.4.525/bd.1992.26.4.525.xml.

Laing, Edited Karen, and Liz Todd. 2015. ‘Theory-Based Methodology: Using Theories of Change in Educational Development, Research and Evaluation’. Newcastle University: Research Centre for Learning and Teaching.

Lenkeit, Jenny, Knut Schwippert, and Michel Knigge. 2017. ‘Configurations of Multiple Disparities in Reading Performance: Longitudinal Observations across France, Germany, Sweden and the United Kingdom’. *Assessment in Education: Principles, Policy & Practice* 0 (0): 1–35. https://doi.org/10.1080/0969594X.2017.1309352.

Lewis, Steven. 2017. ‘PISA “Yet To Come”: Governing Schooling through Time, Difference and Potential’. *British Journal of Sociology of Education* 0 (0): 1–15. https://doi.org/10.1080/01425692.2017.1406338.

Livingstone, I.D., N.T. Postlethwaite, K.J. Travers, and L.E. Suter. 1986. ‘Second International Mathematics Study. Perceptions of the Intended and Implemented Mathematics Curriculum’. Contractor’s Report CS-86-212. Washington, DC: Center for Statistics (OERI/ED).

Loveless, Tom. 2001. ‘International Tests Are Not All the Same’. *Brookings* (blog). 30 November 2001. https://www.brookings.edu/research/international-tests-are-not-all-the-same/.

Loveless, Tom, Helen F. Ladd, and Cecilia Rouse. 1998. ‘The Use and Misuse of Research in Educational Reform’. *Brookings Papers on Education Policy*, no. 1: 279–317.

Martens, Kerstin, Dennis Niemann, and Janna Teltemann. 2016. ‘Effects of International Assessments in Education – a Multidisciplinary Review’. *European Educational Research Journal* 15 (5): 516–22. https://doi.org/10.1177/1474904116668886.

McDonald, Rhona. 2010. ‘The Government Chief Scientific Adviser’s Guidelines on the Use of Scientific and Engineering Advice in Policy Making’. URN 10/669. UK: Government Office for Science.

Medrich, E.A., and J.E. Griffith. 1992. *International Mathematics and Science Assessments: What Have We Learned?* Washington, DC: Office of Educational Research and Improvement, Department of Education.

Mitter, Wolfgang. 1992. ‘Education in Eastern Europe and The-Former Soviet Union in a Period of Revolutionary Change: An Approach to Comparative Analysis’. In *Education and Economic Change in Eastern Europe and the Former Soviet Union*, edited by David Phillips and Michael Kaser, 2 (1):15–28. Oxford Studies in Comparative Education. Wallingford, UK: Triangle Books.

Morsy, Leila, Tatiana Khavenson, and Martin Carnoy. 2018. ‘How International Tests Fail to Inform Policy: The Unsolved Mystery of Australia’s Steady Decline in PISA Scores’.

International Journal of Educational Development 60 (May): 60–79. https://doi.org/10.1016/j.ijedudev.2017.10.018.

Mullis, Ina V. S, Michael O Martin, and Tom Loveless. 2016. *20 Years of TIMSS International Trends in Mathematics and Science Achievement, Curriculum, and Instruction*. Chestnut Hill, Ma: TIMSS & PIRLS.

Pettersson, Daniel, Thomas S. Popkewitz, and Sverker Lindblad. 2017. 'In the Grey Zone: Large-Scale Assessment-Based Activities Betwixt and between Policy, Research and Practice'. *Nordic Journal of Studies in Educational Policy* 3 (1): 29–41. https://doi.org/10.1080/20020317.2017.1316181.

Phillips, David, and Kimberly Ochs. 2003. 'Processes of Policy Borrowing in Education: Some Explanatory and Analytical Devices'. *Comparative Education* 39 (4): 451–61. https://doi.org/10.1080/0305006032000162020.

Plomp, Tjeerd. 1998. 'The Potential of International Comparative Studies to Monitor the Quality of Education'. *Prospects* 28 (1): 45–59.

Porter, Andrew C., and Adam Gamoran, eds. 2002. *Methodological Advances in Cross-National Surveys of Educational Achievement*. Washington, DC: National Academies Press.

Postlethwaite, T. Neville, and Frederick Leung. 2007. 'Comparing Educational Achievements'. In *Comparative Education Research: Approaches and Methods*, edited by Mark Bray, Bob Adamson, and Mark Mason, 215–39. Dordrecht: Springer Netherlands. http://dx.doi.org/10.1007/978-1-4020-6189-9_9.

Raudenbush, Stephen W., and Ji-Soo Kim. 2002. 'Statistical Issues in Analysis of International Comparisons of Educational Achievement'. In *Methodological Advances in Cross-National Surveys of Educational Achievement*, by Andrew C. Porter and Adam Gamoran, 267–94. Washington, DC: National Academies Press.

Robinson, Mark, and Will Evans. n.d. 'Assessing the Strength of Evidence in the Education Sector'. Building Evidence in Education.

Schneider, Barbara, Martin Carnoy, Jeremy Kilpatrick, William Schmidt, and Richard Shavelson. 2007. *Estimating Causal Effects: Using Experimental and Observational Designs*. A Think Tank White Paper. American Educational Research Association.

Shaxson, Louise. 2005. 'Is Your Evidence Robust Enough? Questions for Policy Makers and Practitioners'. *Evidence & Policy* 1 (1): 101–12. https://doi.org/10.1332/1744264052703177.

Smith, Marshall S. 2002. 'Drawing Inferences for National Policy from Large-Scale Cross-National Education Surveys'. In *Methodological Advances in Cross-National Surveys of*

Educational Achievement, by Andrew C. Porter and Adam Gamoran, 295–320. Washington, DC: National Academies Press.

Sollerman, S., and A. Pettersson. 2016. *Focusing on mathematics. Analysis of coherence between Swedish governance documents and the international study TIMSS 2015*. Stockholm: Skolverket.

Treviño, Ernesto, and Miguel Órdenes. 2017. 'Exploring Commonalities and Differences in Regional and International Assessments'. UIS Information Paper 48. Montreal, Quebec, Canada: UNESCO Institute for Statistics. http://uis.unesco.org/sites/default/files/documents/ip48-exploring-commonalities-differences-regional-international-assessments-2017-en.pdf.

Tuijnman, Albert, and Norberto Bottani, eds. 1994. *Making Education Count: Developing and Using International Indicators*. Paris: OECD Publications and Information Centre.

Tyumeneva, Y. 2013. *Disseminating and Using Student Assessment Information in Russia*. Vol. 10. SABER – SYSTEMS APPROACH FOR BETTER EDUCATION RESULTS. Washington, DC, US: The World Bank. https://www.hse.ru/mirror/pubs/lib/data/access/ram/ticket/6/15354026944b2fac632e1a84e9e0559607fe6559dc/WP10_Russia_web.pdf.

Urick, Angela. 2018. 'Secondary Data Analysis in the Field of Educational Leadership and Policy Studies'. In *Complementary Research Methods for Educational Leadership and Policy Studies*, edited by Chad R. Lochmiller, 143–71. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-93539-3_8.

Verger, Antoni, Lluís Parcerisa, and Clara Fontdevila. 2019. 'The Growth and Spread of Large-Scale Assessments and Test-Based Accountabilities: A Political Sociology of Global Education Reforms'. *Educational Review* 71 (1): 5–30. https://doi.org/10.1080/00131911.2019.1522045.

Wu, Margaret. 2010. 'Comparing the Similarities and Differences of PISA 2003 and TIMSS'. OECD Education Working Papers 32. http://www.oecd-ilibrary.org/education/comparing-the-similarities-and-differences-of-pisa-2003-and-timss_5km4psnm13nx-en.

Yastrebov, Gordey A., Alexey R. Bessudnov, Marina A. Pinskaya, and Sergey G. Kosaretsky. 2014. 'Contextualizing Academic Performance in Russian Schools: School Characteristics, the Composition of Student Body and Local Deprivation'. SSRN Scholarly Paper ID 2531276. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2531276.

Агранович, М.Л. 2011. ‘Оценка и управление качеством образования в территориальных системах’. *Управление образованием: теория и практика*, no. 1 (1): 67–72.

———. 2017. ‘Индикаторы Достижения Целей Устойчивого Развития в Сфере Образования и Национальная Образовательная Политика’. *Вопросы Образования//Educational Studies*, no. 4: 242–264.

Болотов, В.А. 2018. ‘Прошлое, Настоящее и Возможное Будущее Российской Системы Оценки Качества Образования’. *Вопросы Образования//Educational Studies*, no. 3: 287–97. https://doi.org/DOI: 10.17323/1814-9545-2018-3-287-297.

Джуринский, А.Н. 2013. *Сравнительная Педагогика. Взгляд Из России*. Москва: МПГУ Издательство Прометей. https://www.litres.ru/aleksandr-naumovich-dzhurinskiy/sravnitelnaya-pedagogika-vzglyad-iz-rossii/chitat-onlayn/.

Капуза, А., Ю. Керша, А. Захаров, and Татьяна Хавенсон. 2017. ‘Образовательные Результаты и Социальное Неравенство в России: Динамика и Связь с Образовательной Политикой’. *Вопросы Образования*, no. 4: 10–35. https://doi.org/10.17323/1814-9545-2017-4-10-35.

Ковалева, Г.С. 2011. ‘Образовательные Достижения Учащихся в Системе Оценки Качества Образования (Презентация)’. *Управление Образованием: Теория и Практика* 1 (1): 62–67.

Решетникова, О.А. 2017. ‘Национальные Экзамены в Российской Федерации и Международные Сравнительные Исследования Качества Образования: Точки Соприкосновения’. *Педагогические Измерения*, no. 2: 5–8.

Хавенсон, Татьяна. 2019. ‘Интеграция Школ в Латвии и Эстонии Через Реформу Содержания Образования’. *Вопросы Образования*.