SCANDINAVIAN JOURNAL OF EDUCATIONAL RESEARCH, 2018
VOL. 62, NO. 3, 368–385
https://doi.org/10.1080/00313831.2016.1258729

Routledge
Taylor & Francis Group

The Contribution of International Large-Scale Assessments to Educational Research: Combining Individual and Institutional Data Sources

Rolf Strietholt[a,b,c] and Ronny Scherer[b]

[a]Institute for School Development Research, Technische Universität Dortmund, Dortmund, Germany; [b]Centre for Educational Measurement, University of Oslo, Oslo, Norway; [c]Department of Education and Special Education, University of Gothenburg, Gothenburg, Sweden

ABSTRACT

The present paper aims to discuss how data from international large-scale assessments (ILSAs) can be utilized and combined, even with other existing data sources, in order to monitor educational outcomes and study the effectiveness of educational systems. We consider different purposes of linking data, namely, extending outcomes measures, analyzing differences over time or across cohorts, and supplementing context information. These linking strategies are illustrated by a non-exhaustive selection of studies that exploited ILSAs to investigate a wide range of educational topics. We conclude that the main contribution of ILSA to educational research lies in the ways they facilitate analyses of educational policy and policy-related issues at the institutional level by means of cross-country analyses. However, the scope of these studies also covers high-quality data on lower levels of the educational system.

ARTICLE HISTORY
Received 24 April 2016
Accepted 30 October 2016

KEYWORDS
Comparative research; international large-scale assessments; methodology

International large-scale assessments (ILSAs) generate data that can be used to make *generalized descriptions* of educational outcomes across countries. They also collect context information at different stages and levels of educational systems. In this context, "stages" refer to educational stages such as early childhood education, primary education, lower-secondary education and so forth, and "levels" describe the multilevel structure of educational systems where, for example, children are nested within classrooms that are in turn nested within schools and so forth. This data can be used to investigate the *determinants* (e.g. class size effects on educational outcomes) and *consequences* (e.g. economic returns to educational outcomes) for the observed differences in educational outcomes. Against this backdrop, the purpose of the present paper is to review how different ILSA data sources can be combined to investigate educational research questions. Notice that we will not discuss ILSA's impact on policy-making in this paper (e.g. Breakspear, 2012; Grek, 2009; Meyer & Benavot, 2013), but the kinds of research questions that can be addressed based on the combination of different ILSA data sources. The combination of data may not only be based on data from within ILSAs (i.e., test scores and survey data), but may involve external data sources that provide additional information on educational systems and other societal issues (e.g., gross domestic product [GDP]). In particular, we argue that combinations of the data from different sources provide opportunities to address research questions that cannot be assessed without the data from ILSAs. To illustrate such studies, we will present a selection of studies and describe how they exploited the potential of ILSAs to address questions in the context of education research.

CONTACT Rolf Strietholt rolf.strietholt@tu-dortmund.de Institute for School Development Research, Technische Universität Dortmund, Vogelpothsweg 78, Dortmund 44227, Germany

The Origin of ILSAs

To understand how ILSAs contribute to educational research, it is worthwhile remembering the origin of the first international assessments. Over the last century, international comparative research has become increasingly popular in social policy analysis. Although education has not been central to this field, its role for economic development and social mobility was studied extensively (e.g., Barro, 1991; Shavit & Blossfeld, 1993). For this purpose, scholars relied on measures that quantified participation in education, such as the number of years of schooling or the proportion of an age group enrolled at various levels of the educational system. Internationally comparable data on these measures were already available through policy documents or national statistical institutes in many countries. Over the years, however, it has increasingly been recognized that what students learn *within* countries might be quite different *across* countries even though the amount of schooling is the same (e.g., Anderson, 1961; Postlethwaite, 1995). In 1958, a group of researchers met at the UNESCO Institute for Education in Hamburg to discuss the possibility of extending the existing comparative research on education by measures of learning outcomes. The basic ideas were to conduct a study with some "objective indicators" of educational outcomes measured by means of internationally comparable tests and questionnaires, and to use these quality indicators to investigate which institutional features of educational systems facilitate student learning. The researchers decided to undertake an international pilot study in different regions and languages around the world in order to explore if an ILSA is feasible. As the international administration proved to be practicable and the study produced meaningful results, the same group of researchers conducted a full-scale study with rigorous sampling criteria and refined procedures for the development of the survey and test instruments (Foshay, Thorndike, Hotyat, Pidgeon, & Walker, 1962; Husén, 1967). Design and procedures of ILSAs have been further developed since then; more than 50 studies with up to 70 participating countries followed the pilot (Hanushek & Woessmann, 2011; Papanastasiou, Plomp, & Papanastasiou, 2011). The largest assessments cover mathematics, reading, and science; fewer and smaller assessments focus on civic and citizenship education, computer and information literacy, and foreign language education; whilst some areas of history, literature education, arts, religious studies, and social science are (currently) not well represented (see Meyer, Strietholt, & Hal-Levi, in press).

Comparability is the Heart of the Studies

The core feature of ILSAs is that they are designed to establish comparability across countries. Comparability refers to the population under investigation and the cross-cultural validity of the instruments. International comparisons are biased if the respective samples are not balanced in terms of both age and schooling (Strietholt, Rosén, & Bos, 2013). For this reason, these studies define the target population precisely, apply a rigorous sampling strategy, and synchronize the months of test administration to ensure a high degree of comparability. Further, ILSAs have comprehensive quality monitoring, carefully developed assessment frameworks that are meaningful in a wide range of countries, translation verifications, and standardized administration procedures to ensure a cross-cultural validity of the instruments and measurement procedures.

Before we elaborate on the potential ILSAs provide for educational research, we briefly bring to attention the different sets of measures these studies collect (see also McDonnell, 1995; Strietholt, Gustafsson, Rosén, & Bos, 2014; Travers & Westbury, 1989). Information on the different levels of the educational systems is collected from students, their parents, teachers, and principals, as well as from the national research coordinators by means of standardized achievement tests, questionnaires, and reports. National research coordinators prepare country reports and complete contextual questionnaires on the contents of test items included in national curricula, the structure and organization of the school system, national examinations, teacher certification, and so on. Further, aggregated data from lower levels can be used to describe, for example, the social structure in a

country. Principals and teachers complete questionnaires on educational settings to survey information on school and classroom conditions and processes including school features such as leadership, public or privately operated, what is actually taught in individual classrooms, who teaches it, and how it is taught. Aggregated data from students or their parents can be used to describe, for example, the social composition of schools. Students and their parents are surveyed about the socioeconomic status, migration background, parental support, and other home factors that affect student learning. Lastly, students are tested to gather information on what they have actually learned and what they think about it.

The development of the achievement tests and scores has received special attention to ensure high psychometric quality standards for valid comparisons of educational systems worldwide (e.g., Davier, Gonzales, & Mislevy, 2009). However, educational outcomes and contexts for learning are not only measured by achievement tests, they are also measured by background questionnaires. Given that researchers have pointed to the lack of standardization in the assessment instruments used in different studies in educational research (e.g., Breen, Luijkx, Müller, & Pollak, 2009; Klieme, 2013), ILSAs use the same scales or items across countries. This is in fact an attractive feature because they facilitate comparable measures of complex constructs such as socioeconomic status, instructional quality, or school autonomy. International assessments have traditionally administered paper-based assessments; however, since recently, the Organisation for Economic Co-operation and Development's (OECD) Program for International Student Assessment (PISA) has provided opportunities to take computer-based tests. These tests enable the collection of process (or log file) data and therefore provide another rich data source for educational studies, for example, on response times or processes (Greiff, Niepel, Scherer, & Martin, 2016).

ILSAs as Sources of Research Data

The ILSAs assemble a wide range of information on educational outcomes and contexts for learning, yet no single study captures all educational measures. For this reason, the combination of the data from different sources promises interesting research designs that allow researchers to approach educational issues that cannot be addressed within a single study. To explore the scope of research questions that can be studied based on the data from ILSAs, we first propose that it is possible to identify at least three different purposes of combining different data sources. Thereafter, we use Bray and Thomas' (1995) framework of research in comparative education analyses to evaluate what specific topics can be addressed with ILSA data. On the basis of a non-exhaustive selection of studies, we will then showcase secondary analyses of ILSA that combined different kinds of ILSA data to investigate various research topics.

Purposes of Linking

Study Differences Over Time or Across Cohorts

Some studies focus on multiple grades and/or are conducted repeatedly, such that is it possible to combine data from the *same* "source study." For example, the Trends in International Mathematics and Science Study (TIMSS) was introduced in 1995 to assess educational outcomes in primary, secondary, and upper-secondary school, and it is repeated in a four-year cycle. With such complex study designs, researchers can exploit variation across multiple educational stages (e.g., Hanushek & Woessmann, 2006) and over time (Rosén & Gustafsson, 2016). Moreover, they allow researchers to apply so-called "fixed-effects models" to make causal claims about the effects of certain features of the educational system on educational outcomes (see Schlotter, Schwerdt, & Woessmann [2011] for a discussion of the methodological advantages of such research designs). It is also possible to combine the data from *different* ILSAs. A recent study linked the achievement tests of the Progress in International Reading Literacy Study (PIRLS) (2011/2006/2001), the Reading Literacy Study (2001/1991), and the Reading Comprehension Study (1971) in order to measure international trends

in students' reading achievement at the end of primary school over 40 years (Strietholt & Rosén, 2016).

Extend outcome measures

Information from ILSAs on different domains can be combined to capture a more comprehensive picture of educational outcomes. For example, Martin, Foy, Mullis, and O'Dwyer (2013) (see below) combined TIMSS and PIRLS data to compare the relationships between student achievement and a set of covariates in reading, mathematics, and science. On the basis of these data, they were able to disentangle the degree of domain specificity in these relationships (see Strietholt, Manitius, Berkemeyer, & Bos, 2015 for another example). Besides this, the availability of multiple outcome domains is also attractive from a methodological standpoint, because it is possible to relate within-student between-outcome differences to instructional practices in order to overcome bias from unobserved student variables. Schwerdt and Wuppermann (2011) provided an example for this analytical strategy. They compared how well students perform in mathematics and science and combined this information with survey data on the instructional practices of their teachers in both domains.

Supplement context information

Another purpose of combining different data sources is to supplement outcome measures from ILSAs with context information from other sources such as international databases (e.g., United Nations Educational, Scientific, and Cultural Organization, World Bank) or surveys. For instance, some ILSAs primarily focus on the assessment of student achievement without administering questionnaires to teachers; in the cases where they administer them, these questionnaires may not necessarily capture teacher constructs that are highly relevant for student learning. In this regard, further international studies on teachers and teacher education such as the Teaching and Learning International Survey (TALIS) and the Teacher Education and Development Study in Mathematics come into play because they facilitate such information, and – combined with student data – extend the scope of ILSAs. Hanushek and Woessmann (2012) provided another example on the association between student achievement and economic growth. The authors combined various ILSAs with external data on growth in GDP.

These and other combinations of individual and institutional data from various sources will be illustrated in a collection of example studies in the following section.

Topics of Investigation

The vast amount of data that has been amassed in ILSAs makes it possible to address a wide range of educational issues including descriptions of differences in educational outcomes within and between countries, as well as studies on the determinants and consequences of the observed differences. Bray and Thomas (1995) provide a useful framework to organize this research in comparative education analyses. They propose a three-dimensional classification of comparative, educational research with regard to (1) geographical/locational levels, (2) nonlocational demographic groups, and (3) aspects of education and society. Figure 1 shows an adapted version of the cube that was originally presented by Bray and Thomas (1995). Comparative research projects that make use of ILSA data concern all three dimensions, and can therefore be located in at least one cell of the cube. As an example of how a research project can be located in this cube, the shaded cells refer to a study that compares how instructional practices in Norway and Sweden impact student achievement.

It is worth noting that educational systems have a multilevel structure – students are nested within classrooms, classrooms are nested within schools, and so forth. Although educational policies are typically located at higher levels, they also manifest on lower levels. In other words, educational outcomes result from complex interactions between different educational actors such as policy makers, principals, teachers, parents, and students on the one hand, and different contexts such

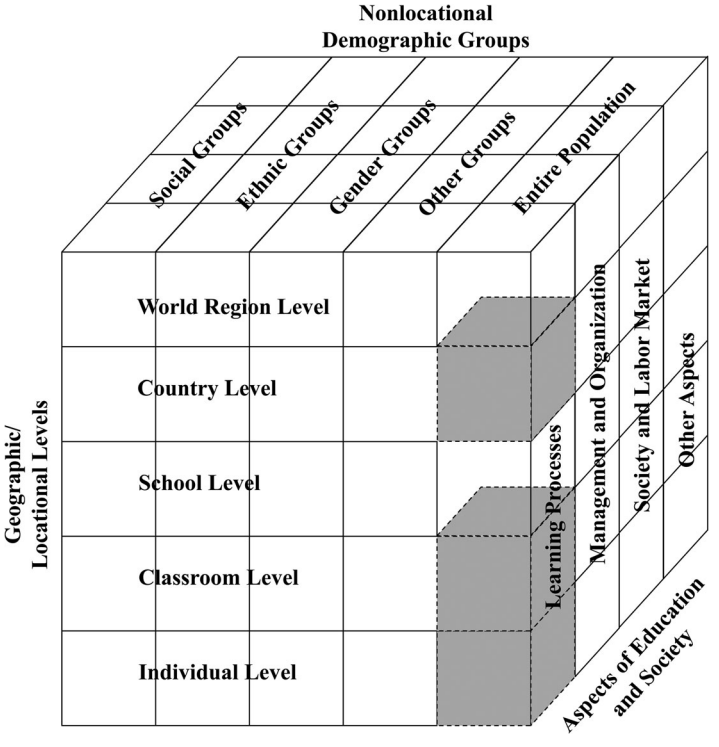

Figure 1. Bray and Thomas's (1995, p. 475) framework for comparative education analyses.

as families, classrooms, schools, and governments on the other hand (Klieme, 2013). For this reason, multilevel analyses are required to study direct, mediating, and moderating effects at the various levels to understand the complex mechanisms in effect between educational policies, learning processes, and outcomes, or, in other words, to understand what works for whom in what circumstances (see Pawson & Tilley, 1997). In this context, it is clear that any comparative study has to limit their scope, but the distinction between geographical/locational levels, nonlocational demographic groups, and different aspects of education and of society provides at least a framework to recognize limitations of the phenomena a study addresses and to identify the mutual influences of other layers and levels (Bray & Thomas, 1995; Bray, Adamson, & Mason, 2007).

Examples of Studies Based on ILSA Data

In this section, we illustrate the potential of combining data that were obtained from different large-scale studies or different cycles of the same studies. We will showcase a selection of studies and describe how they exploited the potential of ILSAs to address relevant research questions in the contexts of school effectiveness, students' performance and behavior in achievement tests, and effects of home background variables on achievement. Specifically, we classify the data sources, purposes of linking, and topics of investigation according to Bray and Thomas' framework.

Bibliographic analyses of the representation of PISA (Domínguez, Vieira, & Vidal, 2012) and PIRLS (Lenkeit, Chan, Hopfenbeck, & Baird, 2015) in scientific journals indicate that ILSA data are increasingly utilized by researchers from different countries and diverse disciplinary backgrounds for methodologically as well as substantively oriented studies. Our research adds to this research in that we emphasize the combination of data source and classify the topics based on a predefined framework. In this respect, we chose seven recent studies – published between 2012 to 2016 –

that represent different linking strategies and diverse research topics. It is important to note that our selection of studies is by no means exhaustive. Rather, we aim to identify a set of studies that we consider suitable to exemplify the range of linking strategies and areas of investigation that can be addressed with ILSA data. Besides, this section is aimed at encouraging researchers to combine different data sources on the one hand, and different data levels of educational systems on the other hand. To this end, we provide a somewhat more extended description of the respective studies.

Table 1 shows that secondary analyses of ILSAs made use of the data to address a quite diverse range of research topics. The studies utilized not only data from recent assessments such as PIRLS, PISA, and TIMSS, but also data from previous studies that were conducted in the 1960s, 1970s, 1980s, and 1990s. With regards to the different purposes of linking, our selection of studies covers examples that combine data from different sources to extend outcome measures, study differences over time and across cohorts, and supplement contextual information. Furthermore, our overview shows that ILSA data have been used to study learning processes, management and organization, and society- and labor market-related issues for different social, ethnic, and gender groups as well as for entire populations, and are located at the individual, classroom, school, country, and world region levels. In other words, the few studies listed in Table 1 demonstrate that ILSA data have the potential to study all geographical/locational levels, nonlocational demographic groups, and aspects of education and society, as proposed by Bray and Thomas (1995). In the remaining part of this section, we will encapsulate the research goals, processes of data combination, methodological approaches, and main findings of the seven studies.

Exploring Generalizability Across Countries and Domains (Martin et al. 2013)

There is no ILSA that captures the three core-subjects mathematics, reading, and science simultaneously in primary school. The combination of TIMSS and PIRLS data, however, provides a unique opportunity to extent the outcome measures of the respective studies to not only examine the relations among achievement in three domains but also to study whether or not differential effects of school-related learning processes can be identified across domains. Information on learning processes was surveyed from principals, teachers, and students; student and teacher data were aggregated at the school level. The study shows how ILSA data can be used to examine the whether research findings can be replicated across countries and domains.

Research goals

The goals of this study were to examine how school variables relate to student achievement across the three domains for all countries participating in PIRLS and TIMSS jointly.

Process of data combination

In 2011, 34 countries and three of the sub-national entities agreed on participating in both TIMSS and PIRLS jointly and to administer the corresponding test material to the same students (Mullis & Martin, 2013). As a consequence, for these fourth-grade students, data on their achievement in three domains – mathematics, science, and reading – and information on their motivation, self-perceptions, and perceptions of classroom instruction were available (Scherer & Gustafsson, 2015). The fact that the same students participated in both studies ruled out major alternative explanations for differences across outcome domains that were related to person characteristics or sampling error. Since both large-scale studies were administered to the same students, the process of data combination was based on the student level; data linking was finally conducted with the help of unique student identification numbers that reoccurred in both TIMSS and PIRLS 2011.

Methodological approach

In order to achieve the above mentioned research goal, multilevel modeling was conducted to account for the hierarchical data structures in TIMSS and PIRLS with students as individual units

Table 1. Classification of the studies.

Study	Data sources	Purpose of linking	Topic of investigation research question	Geographical/ locational levels	Aspect of education and society	Nonlocational demographic groups
Martin et al. (2013)	TIMSS 2011, PIRLS 2011	Extending outcome measures	Compare the impact of school variables on learning processes across different outcome domains	Country Classroom Individual	Learning processes	Entire population
Nilsen and Gustafsson (2014)	TIMSS 2007, TIMSS 2011	Study differences over time or across cohorts	Test if change in the performance level is mediated through changes instructional practices in Norway	Classroom Individual	Learning processes	Entire Population
Greiff et al. (2015)	PISA 2012, Log file data	Supplement context information	Explore sequences of actions on a computerizes tests and their relation with achievement across countries	Country Individual	Learning processes	Entire Population
Rosén and Gustafsson (2016)	SIRS 1991/2001, PIRLS 2001/2006	Study differences over time or across cohorts	Investigate if a countries' reading performance dropped as a function of spread of personal computers	Countries Individual	Society and labor market	Social Groups Ethnic Groups Gender Groups Entire Population
Austin et al. (2015)	TALIS 2013, PISA 2012	Supplement context information	Study international variation in how school composition and teachers' self-efficacy are associated	Country School	Other aspects	Entire Population
Hanushek and Woessmann (2012)	FIMS, FIRS, FISS, SIMS, SIRS, SISS, TIMSS, PIRLS, PISA (various cycles), Penn World Table	Supplement context information Extending outcome measures	Investigate the relationship between average performance and GDP growth across countries	World Region Countries	Society and labor market	Entire Population
Ruhose and Schwerdt (2016)	TIMSS, PIRLS, PISA (various cycles), UNESCO	Study differences over time or across cohorts Supplement context information Extending outcome measures	Estimate the causal effect of tracking on the migrant-native achievement gap	Countries Individual	Management and organization	Ethnic groups

Note: FIMS = First International Mathematics Study (1964); FIRS = First International Reading Study (aka. Six Subject Survey: Reading Comprehension, 1971); FISS = First International Science Study (1971); SIMS = Second International Mathematics Study (1982); SIRS = Second International Reading Study (aka. Reading Literacy Study 1991/2001); SISS = Second International Science Study (1984).

of analysis and schools as aggregated units of analysis. Martin et al. (2013) chose factors describing the school environment (e.g., safety in schools, emphasis on academic success), school instruction (e.g., teacher support), and students' home background (e.g., home resources for learning). These country-by-country analyses resulted in contextual effects. Achievement scores in the three domains served as outcome variables.

Main findings

Exploiting the potential of the TIMSS & PIRLS 2011 link, the authors found large country differences in the relations among home and school contextual variables and achievement. For example, the home background explained between less than 10% and more than 40% of the total variation of the achievement scores. Interestingly, there was a considerable degree of stability across the domains of mathematics, science, and reading within countries.

Linking Different Study Cycles and Levels of Analysis to Study Change: Relations Between Schools' Emphasis On Academic Success and Changes in Science Performance in Norway (Nilsen & Gustafsson, 2014)

Particularly for countries that do not conduct national assessments, ILSAs sometimes provide the only databases to trace developments in educational outcomes and learning contexts (Johansson, Strietholt, Rosén, & Myrberg, 2013). Although this is not necessarily the case for Norway, this study investigated whether the change in schools' emphasis on academic skills had caused the recent rise in science performance of Norwegian students, and illustrates how different cycles of the same study can be combined at the item level, thereby treating the cycles as groups. This approach allows researchers to investigate differences in relations among constructs or their means within a country. In this context, it is worth mentioning that this study ensured the comparability of the measures used in these cycles by examining specific levels of measurement invariance (see also Millsap, 2011).

Research goals

This study fed into research on school effectiveness by examining: (1) the differences in aspects of teachers' perceived school emphasis on academic success (SEAS) between the TIMSS cycles that were conducted in 2007 and 2011; (2) the relation between these SEAS and differences in students' science achievement; (3) the relations described under (2) after controlling for students' background variables.

Process of data combination

In order to address the aforementioned research agenda, Nilsen and Gustafsson (2014) focused on the Norwegian TIMSS 2007 and 2011 samples of eight graders and their science teachers. The authors combined the two data sources, the TIMSS 2007 and TIMSS 2011 student and teacher/classroom data, on the basis of anchor items. Specifically, since SEAS was measured in both TIMSS cycles by the same items, the data could be combined in a trend design framework. In fact, the two TIMSS samples represented different groups. Regarding the achievement data in the domain of science, the plausible values obtained from the two studies were treated as comparable between TIMSS 2007 and 2011. This treatment was justified in light of the common-item design of the TIMSS achievement tests that allowed for a concurrent calibration between the study cycles (Mullis, Martin, Foy, & Arora, 2012; Olson, Martin, & Mullis, 2008).

Methodological approach

On the basis of the combined TIMSS 2007 and 2011 science data, the authors specified a series of multilevel structural equation models (Rabe-Hesketh, Skrondal, & Pickles, 2004), directly addressing the hierarchical data structure (within-level: students, between-level: classrooms/teachers). Differences between the TIMSS cycles were quantified by the effects of a dummy variable, which indicated

the year of the cycle, on the variables under investigation. This "multiple indicators multiple causes" (Brown, 2015) approach considers the cycles to be a grouping variable in the combined data set.

Main findings

The authors found that differences in students' science performance were fully mediated by SEAS, thus suggesting that SEAS and changes in science performance were associated.

Analyzing Log Files to Understand How Children Solve Complex Problems (Greiff, Wüstenberg, & Avvisati, 2015)

Greiff et al.'s (2015) study points to the potential of log file data, made available by computer-based assessments, that can be exploited to inform analysts about the *processes* of solving a task beyond students' performance. Their study showcases how different kinds of individual, computer-derived assessment data can be combined within a single study. Given the richness of such log-file data, future large-scale assessments may benefit from the additional information on students' behavior to better understand and describe students' performance (Quellmalz et al., 2012).

Research goals

This study demonstrates how data on students' performance can be enhanced by behavioral data obtained from log files. Specifically, the authors sought to describe the processes and strategies of solving complex problems in a computer-based assessment in addition to the mere performance measures such as overall proficiency scores and single item scores. In this respect, Greiff et al. (2015) focused on one of the most prominent strategies of vary-one-thing-at-a-time (VOTAT), with which the specific relations among variables could be disentangled systematically (Schwichow, Croker, Zimmerman, Höffler, & Härtig, 2016; Tschirgi, 1980).

Process of data combination

The OECD's PISA 2012 administered computer-based assessments of creative problem solving to about 85,000 15-year-old students in 44 countries. Students' individual performance data on a specific problem solving item, *Climate control*, and the corresponding log files on their behavior within the computer-based assessment were used in this study. The latter contained all information on observable behavior, such as the VOTAT strategy, the number of clicks, response times, and the sequence of actions. Greiff et al. (2015) examined these data at both the individual and the country level. Data from the different sources were directly combined at the individual (student) level. The OECD provided separate data files on students' performance and their log file data, which were linked via students' unique identification numbers in the data sets.

Methodological approach

The authors tracked the application of the VOTAT strategy for each student and generated further dichotomous indicators to categorize whether or not students mastered the *Climate control* task. The correlations between these indicators to students' performance on the task (0 = incorrect, 1 = correct) and the overall, creative problem solving performance, as indicated by plausible values, were estimated. Frequencies and percentages were transformed and subsequently aggregated to the country level.

Main findings

Greiff et al. (2015) found that the individual students' performance of the VOTAT strategy was significantly and positively related to performance on the *Climate control* item and the overall problem solving score. Moreover, the authors were able to identify different mastery levels that described students' performance on the selected item on the one hand, and the successful or unsuccessful application of VOTAT in order to solve this item on the other hand. At the country level, high correlations between the frequency of VOTAT and the overall problem solving performance were found.

Consequences of Technological Change for Reading Achievement: Country-level Longitudinal Evidence (Rosén & Gustafsson, 2016)

This study shows how the trend design within a study can be exploited in order to draw causal inferences on differences in variables over time (Gustafsson, 2007). Using a difference-in-differences approach and controlling for cofounding variables at the same time provided strong support for the causal claim. Besides the opportunity to study causality with observational data to some extent, the repeated and linked design of PIRLS allows researchers to investigate research questions longitudinally at the country level. Nevertheless, this design does not permit inferences about intra-individual change, but rather changes over time within a country. The latter has the potential to inform researchers and policy-makers about changes in student achievement and possible explanatory variables thereof.

Research goals

The goal of this study was to examine the causal effect of computer availability at home and reading achievement for fourth grade students. Specifically, Rosén and Gustafsson (2016) investigated the relation between changes in computer availability and changes in students' reading achievement between 1991 and 2006 in order to disentangle causal effects and interaction effects with students' background.

Process of data combination

Rosén and Gustafsson (2016) used data from two sources: the International Association for the Evaluation of Educational Achievement's (IEA) Reading Literacy Study that was conducted in 1991 and 2001, and PIRLS 2001 and 2006. Separate sets of results are reported for these two data sources. Students' computer availability at home, information on their socioeconomic background, and their reading achievement fed into the analyses. Rosén and Gustafsson pointed out that the Reading Literacy Studies did not contain any information on students' computer availability at home; as a consequence, the authors restricted the 1991 score of this variable to zero, assuming that most of the participating countries did not provide conditions such that computers were widely available at that time. For the 2001 study, the computer availability information was obtained from PIRLS 2001. In total, the data obtained from nine countries for the period between 1991 and 2001, and 15 countries and 3 sub-national entities for the period between 2001 and 2006, were analyzed. Similar to the Nilsen and Gustafsson (2014) study, the data obtained from the different studies were combined at the item and construct level, resulting in datasets that considered the different studies as grouping variables. This process of data combination was conducted for each country that participated in either of the two study cycles. Each of the achievement scores was considered to be comparable between the cycles of the Reading Literacy Study and PIRLS. In fact, similarly to TIMSS, the latter followed a common item design and allowed for a concurrent calibration of items between PIRLS 2001 and PIRLS 2006 (Martin, Mullis, & Kennedy, 2007).

Methodological approach

The analyses mainly focused on the country level. Specifically, Rosén and Gustafsson (2016) aimed at relating changes in computer availability at home to changes in students' reading achievement across all participating countries. Hence, they took a difference-in-differences approach, and specified it as a regression model with fixed country effects.

Main findings

The authors found a negative effect of home computer availability on student achievement in reading for the two types of studies across countries. They concluded that this effect could be, at least to some extent, interpreted causally.

Linking Different Studies and Levels of Analysis: Examining the School Context and its Influence On Teachers (Austin et al., 2015)

Austin et al.'s (2015) study shows how two large-scale assessments that provided information on different actors and levels of education (i.e., students and teachers) can be combined in order to disentangle factors determining teacher variables. The major strength of this link is that both the teacher and the student samples were representative for the participating schools. This study further shows how contextual and climate effects can be studied with teachers as the main focus. Nevertheless, although an almost direct link between the two samples could be established at the school level, findings relevant for educational effectiveness – for example, with respect to how teaching practices and beliefs relate to student achievement and motivation – must be interpreted with caution, given that teachers and students did not necessarily belong to the same classrooms (Klieme, 2013). But clearly, the Teaching and Learning International Survey (TALIS)-PISA link allows researchers, school developers, and other relevant stakeholders to examine school contexts from the perspective of both teachers and students.

Research goals

The authors of this OECD working paper combined data from the two international large-scale assessments, the TALIS 2013 and PISA 2012. Their main goal was to explore the extent to which school context variables relate to various teacher variables such as their professional development, teaching practices and beliefs, teacher cooperation, teacher self-efficacy, and job satisfaction. Moreover, Austin et al. (2015) intended to describe profiles of schools and teachers with respect to the above mentioned variables.

Process of data combination

In total, the teacher sample comprised about 26,600 teachers who participated in TALIS 2013 and about 103,000 students who participated in PISA 2012. Teachers and students belonged to the same schools in nine countries. It is noteworthy that this linking study focused on mathematics teachers only, as PISA 2012 chose mathematics as the major domain. Both the teacher and student samples were representative of teachers and students within schools. The TALIS-PISA linking was conducted already in the design and administration of the study; teachers and students in the same schools were assigned to the same school identification number, such that a direct link at the school level could be established.[1] As a consequence, Austin et al. (2015) catered their analytic approach to the school level.

Methodological approach

Student data obtained from PISA 2012 were aggregated to the school level; hence, all analyses were conducted for schools within a country. The authors specified multilevel regression models (individual level: teachers, aggregated level: schools) and quantified cross-country variation in the resultant regression coefficients.

Main findings

Austin et al. (2015) identified considerable variation in the extent to which teacher and student variables were related. For instance, they found a significant relation between students' average economic, social, and cultural status and teachers' engagement in professional development activities with varying directions of this relation. Nevertheless, a number of relations and profiles could be identified that were by and large consistent across countries. For instance, teachers' domain-specific self-efficacy in teaching mathematics was lower than their domain-general self-efficacy. Austin et al.

[1]We notice that previous attempts to link TALIS 2008 and PISA 2009 data for the domain of reading were based on indirect linking strategies that required methods of data fusion. This was partly due to the missing identification of schools that participated in both studies. For a more detailed discussion on data fusion approaches, we refer the reader to Kaplan and McCarty (2013).

(2015) concluded their presentation of results with the finding that teachers in schools in which a high variability of PISA scores existed showed higher domain-general self-efficacy in teaching than those in schools with less variability. Further results were presented and discussed in the working paper.

Education and Economic Growth (Hanushek & Woessmann, 2012)

How can ILSAs be combined with other (non-educational) international databases? The example presented here goes back to a classic issue in comparative research, that is, the role of education for economic development. Economists observed a considerable international variation in economic growth during the second half of the last century. Most previous cross-country research regressed economic growth on the number of years of schooling. Unlike this research, Hanushek and Woessmann's (2012) study demonstrated how ILSA data supplement information from other data sources on economic development.

Research goals

The study aimed at estimating the effect of education on economic growth. It recaptured the idea that students within countries might be quite different across countries, and replaced the number of years of schooling by how well countries perform in various ILSAs.

Process of data combination

The Penn World Table (PWT) is a database that has provided information on annual growth in real per capita GDP since the 1950s for more than 150 countries (Feenstra, Inklaar, & Timmer, 2015). Information on student performance was taken from not less than 12 ILSAs that were conducted between 1964 and 2002 (see the overview in Table 1). A very interesting feature of Hanushek and Woessmann's (2012) study is that the achievement scales of the respective ILSA were linked through a linear equating approach, which was based on the National Assessment of Educational Progress in the USA. Data on performance and growth was combined at the country level using the country names for the linking. The authors identified 50 countries with information on both student achievement (ILSAs) and economic growth (PWT). In further analyses, the authors aggregated the country-level data on world region level such as Central Europe, Latin America, the Middle East, and North Africa.

Methodological approach

The basic analysis model was a regression model at the country level with a sample size of 50. Specifically, the authors regressed the average annual growth rate in GDP on the average performance level across the different ILSAs. Extensive sensitivity analyses included the number of years of schooling and other covariates.

Main findings

The study showed that student achievement was a strong predictor of economic growth at the country level and an even stronger predictor at the level of world regions. The effect of years of schooling became insignificant after accounting for student achievement. This finding indicated that a mere quantitative extension of schooling is ineffective for boosting economic growth.

Early Tracking and the Migrant-Native Achievement Gap (Ruhose & Schwerdt, 2016)

Our final example refers to a study that supplements ILSA data on inequality in educational outcomes by UNESCO data on school structure, with the aim of examining whether early educational tracking affects the emergence of the migrant-native achievement gap. Furthermore, the study highlights an interesting methodology, because it combines the data from various ILSAs to capture

educational outcomes at different educational stages. The basic idea was to observe two samples of the same cohort of students in primary and secondary school in order to analyze differences in the achievement gap across educational stages.

Research goals

After primary education, some countries select students into schools according to their abilities (i.e., ability grouping), others have a comprehensive secondary school system. This study aimed to identify the causal effect of tracking on the achievement gap between migrant and native students.

Process of data combination

Achievement data from all previous PIRLS, PISA, and TIMSS cycles were used to trace changes in the migrant-native achievement gap at different educational stages in 45 countries. These studies were linked in such a way that they covered the same outcomes measure in primary and secondary education. For example, the combined ILSA data from TIMSS 1995 (4th grade) and PISA 2000 (15-year-olds) enabled the authors to investigate how the achievement gap evolved for (approximately) the same cohort of students in mathematics and science. Data from the UNESCO International Bureau of Education (2013) supplemented the achievement data by information on the age of first tracking. The authors distinguished between countries that select students into different school tracks before age 15 and countries that have a comprehensive system by then.

Methodological approach

Because the size of the migrant-native achievement gap may vary across countries for multiple reasons, Ruhose and Schwerdt (2016) employed a differences-in-differences strategy to identify the causal effect of early tracking. Their analytical approach was based on the observation that no country tracks students in primary school. Thus, information on the size of the achievement gap prior to and after tracking was available for all countries. A differences-in-differences model addressed whether change in the achievement gap was related to different tracking policies, while eliminating bias from time-invariant confounding variables.

Main findings

The study suggested that early tracking does not significantly affect migrant-native achievement gaps. Several robustness checks confirmed this overall pattern. The authors concluded that the differences-in-differences analyses that were based on different educational stages allowed them to draw causal inferences.

Conclusion

Exploit Variation in Institutional and Societal Features

Our reflections on ILSAs and the studies that utilize their data illustrate how international assessments contribute to educational research. It is worthwhile emphasizing that the contributions of ILSA to educational research are related to the research opportunities of international comparative research in general, and that is that such studies facilitate the analyses of variation as well as similarities across countries. Henceforth, an important contribution of international comparative studies to the research on education is the possibility to observe and exploit variation in educational and societal features at the level of countries or world regions. The analyses of the effect of educational outcomes on economic growth, the consequences of technological change and reading performance, and the impact of tracking on educational inequality are three examples for such studies.

Investigate Generalizability Across Different Countries

A second contribution of ILSA is the opportunity to test hypotheses about the relations between variables observed at lower levels of the educational system for different countries. The set of home background variables Martin et al. (2013) considered in their study explains between <10% and >40% of the variance in student performance in the respective countries. This huge variation points to the limited generalizability of the findings from individual countries. At the same time, this variation raises new questions about the role of variables on higher levels of the educational systems, that is, for multilevel analyses on moderation and cross-level interaction effects. These are important pieces of information in the process of theory development. For example, Coleman (1968) argued that the effect of schooling on academic achievement is less than that of family background. Heyneman and Loxley (1983) replicated these analyses for a diverse set of countries with ILSA data from outside the USA. They found that schooling outweighs the impact of family background in low-income countries. Subsequent research hypothesized that increasing access to school though mass schooling has diminished the impact of schooling on achievement because there is little variation in schooling within countries (e.g., Baker, Goesling, & LeTendre, 2002).

Wide Range of Research Topics

The value of international assessment for educational research may also be evaluated with respect to the educational issues addressed in studies based upon data from ILSAs. In this regard, we presented seven illustrative examples; interestingly, however, even these few studies cover all the geographical/locational levels, nonlocational demographic groups, and aspects of education and society proposed by Bray and Thomas (1995). As mentioned in the previous section, we believe that the true value of international studies is that they enable comparative research that incorporates the country level to understand effective learning conditions and educational policies. The wide coverage of educational issues addressed in secondary analyses of their data, however, further underlines the capacity of ILSA.

Combined Data From Different Sources Facilitate Innovative Research Designs

The combined data from various ILSAs as well as combinations with other existing data sources facilitate particularly interesting research designs. A first area of innovation concerns the opportunity to draw causal inferences about the causal determinants of educational outcomes. Traditionally, some educationalists were nervous about the bias from unobserved confounding variables in ILSA because "TIMSS, PIRLS, and PISA are observational studies … language that implies causality should generally not be used when discussing the results of analyses of these studies" (Rutkowski, Gonzales, Joncas, & Davier, 2010, p. 148). The combination of ILSA data for different outcome domains, educational stages, and time points, however, allows interesting analyses strategies that can eliminate the bias from unobserved confounding variables (see the examples on early tracking and technological change; see also a recent special issue edited by Rutkowski [2016]). A second area of innovation concerns combinations of ILSA data with information on the determinants and consequences of educational outcomes. The example of the effect of education on economic growth is just one example for such a study. In this regard, the rich data from international organizations such as UNESCO, the World Bank, or the World Health Organization, and other international studies may be combined with data form ILSAs to provide a deeper understanding of the role of education for individuals and societies around the world.

Addressing Methodological Limitations

The described potential of ILSA as research data may create the impression that sophisticated and meaningful analyses are unproblematic and firmly established. Yet it is important to highlight the

limitations of the usage of ILSA. International comparisons and trend studies derive from the assumption that there are comparable groups of students in the respective samples. However, due to differences in the school entry ages across countries, the samples in ILSA typically cover students that differ either in terms of age, grade, or both (e.g., Strietholt et al., 2013). Such differences can be even larger with the combined data from different studies (e.g., age-based PISA and grade-based PIRLS samples). Participation and exclusion rates represent further issues (Rutkowski & Rutkowski, 2016). With regard to the measures used in ILSAs, meaningful comparisons across countries also presume that the measurement of the tertium comparationis – that is the quality that countries that are being compared have in common – is reliable and valid across countries. Critics have argued that the social and cultural pluralism in the construction of education is compressed into a single point in ILSA (e.g., Meyer & Benavot, 2013; Torrance, Lauder, Brown, Dilabough, & Halsey, 2006) and that these scores are subject to methodological limitations (e.g., Rutkowski & Rutkowski, 2010). These limitations are contentious issues in the literature based on ILSA (Domínguez et al., 2012; Johansson, 2016; Lenkeit et al., 2015) and we recommend that researchers recognize them when analyzing ISLA data and conduct, if possible, robustness checks.

Scope of the Present Paper and Outlook

A limitation of the present paper concerns the selection of studies we presented here to illustrate how ILSAs are used as research data. Our selection of studies may to demonstrate that meaningful combinations of data from ILSA and other sources are possible and that a wide range of topics can be addressed based on this data. However, this selection is not representative for all secondary analyses of ILSA. Although existing research in this area already provides detailed bibliographic information such as publication year, journal, authors' affiliation, the classification of specific topics is rather broad and based on ad hoc developed classification schemes (Domínguez et al., 2012; Lenkeit et al., 2015). In addition to this research, it would be interesting to classify a representative set of studies based on an existing classification scheme – such as the one presented by Bray and Thomas (1995) –not only to understand how frequent specific topics have been addressed, but also to identify areas that are not well represented.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Swedish Research Council (Vetenskapsrådet) [grant number 2015-01080].

References

Anderson, C. A. (1961). Methodology of comparative education. *International Review of Education*, *7*(1), 1–23.

Austin, B., Adesope, O. O., French, B. F., Gotch, C., Bélanger, J., & Kubacka, K. (2015). *Examining school context and its influence on teachers: Linking TALIS 2013 with PISA 2012 student data*. Paris: OECD.

Baker, D. P., Goesling, B., & LeTendre, G. K. (2002). Socioeconomic status, school quality, and national economic development: A cross-national analysis of the "heyneman-loxley effect" on mathematics and science achievement. *Comparative Education Review*, *46*(3), 291–312.

Barro, R. J. (1991). Economic growth in a cross section of countries. *Quarterly Journal of Economics, 106*(2), 407–443.

Bray, M., Adamson, B., & Mason, M. (2007). *Comparative education research: Approaches and methods*. Dordrecht: Springer.

Bray, M., & Thomas, R. M. (1995) Levels of comparison in educational studies: Different insights from different literatures and the value of multilevel analyses. *Harvard Educational Review*, *65*(3), 472–491.

Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance*. Paris: OECD.

Breen, R., Luijkx, R., Müller, W., & Pollak, R. (2009). Nonpersistent inequality in educational attainment: Evidence from eight European countries. *American Journal of Sociology*, *114*(5), 1475–1521. doi:10.1086/595951

Brown, T. A. (2015). *Confirmatory factor analysis for applied research (2nd ed.)*. New York, NY: Guilford Press.

Coleman, J. S. (1968). The concept of equality of educational opportunity. *Harvard Educational Review*, *38*(1), 7–22.

Davier, M. v., Gonzales, E. J., & Mislevy, R. J. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. *2*, pp. 9–36). Hamburg, Germany: IEA-ETS Research Institute.

Domínguez, M., Vieira, M.-J., & Vidal, J (2012). The impact of the programme for international student assessment on academic journals. *Assessment in Education: Principles, Policy & Practice*, *19*(4), 393–409. doi:10.1080/0969594X.2012.659175

Feenstra, R. C., Inklaar, R., & Timmer, M. P. (2015). The next generation of the Penn World Table. *American Economic Review, 105*(10), 3150–3182. doi:10.1257/aer.20130954

Foshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievements of thirteen-year-olds in twelve countries*. Hamburg, Germany: UNESCO Institute for Education.

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, *61*, 36–46. doi:10.1016/j.chb.2016.02.095

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105. doi:10.1016/j.compedu.2015.10.018

Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy*, *24*(1), 23–37. doi:10.1080/02680930802412669

Gustafsson, J. E. (2007). Understanding causal influences on educational achievement through analysis of differences over time within countries. In T. Loveless (Ed.), *Lessons learned: What international assessments tell us about math achievement* (pp. 37–63). Washington, DC: Brookings.

Hanushek, E. A., & Woessmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal*, *116*(510), C63–C76. doi:10.1111/j.468-0297.2006.01076.x

Hanushek, E. A., & Woessmann, L. (2011). The economics of international differences in educational achievement. In E. A. Hanushek, S. Machin & L. Wößmann (Eds.), *Handbook of the economics of education* (Vol. *3*, pp. 89–200). Amsterdam: Elsevier.

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth*, *17*, 267–321. doi:10.1007/s10887-012-9081-x

Heyneman, S. P., & Loxley, W. A. (1983). The effect of primary-school quality on academic achievement across twenty-nine high-and low-income countries. *American Journal of Sociology*, *88*(6), 1162–1194.

Husén, T. (Ed.). (1967). *International study of achievement in mathematics: A comparison of twelve countries* (Vols. 1–2). Stockholm: Almqvist & Wiksell.

Johansson, S. (2016). International large-scale assessments: What uses, what consequences? *Educational Research*, *58* (2), 139–148. doi:10.1080/00131881.2016.1165559

Johansson, S., Strietholt, R., Rosén, M., & Myrberg, E. (2013). Valid inferences of teachers' judgements of pupils' reading literacy: Does formal teacher competence matter? *School Effectiveness and School Improvement*, *25*(3), 394–407. doi:10.1080/09243453.2013.809774

Kaplan, D., & McCarty, A. T. (2013). Data fusion with international large scale assessments: A case study using the OECD PISA and TALIS surveys. *Large*-Scale *Assessments in Education*, *1*(1), 1–26. doi:10.1186/2196-0739-1-6

Klieme, E. (2013). The role of large-scale assessments in research on educational effectiveness and school development. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 115–147). Dordrecht: Springer Netherlands.

Lenkeit, J., Chan, J., Hopfenbeck, T. N, Baird, J.-A. (2015). A review of the representation of PIRLS related research in scientific journals. *Educational Research Review*, *16*, 102–115. doi:10.1016/j.edurev.2015.10.002

Martin, M. O., Foy, P., Mullis, I. V., & O'Dwyer, L. M. (2013). Effective schools in reading, mathematics, and science at fourth grade. In M. O. Martin & I. V. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade - implications for early learning* (pp. 109–180). Amsterdam & Boston, MA: International Association for the Evaluation of Educational Achievement and Lynch School of Education, Boston College.

Martin, M. O., Mullis, I. V., & Kennedy, A. M. (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

McDonnell, L.M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis*, *17*(3), 305–322.

Meyer, H.-D. & Benavot, A. (2013). *PISA, power, and policy: The emergence of global educational governance*. Oxford studies in comparative education. Oxford: Symposium.

Meyer, H. D., Strietholt, R., & Hal-Levi, D. (in press). Three models of global education quality and the emerging democratic deficit in global education governance. In M. Akiba & G. LeTendre (Eds.), *Routledge international handbook of teacher quality and policy*. New York: Routledge.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Mullis, I. V., Martin, M. O., & Foy, P. (2013). The impact of reading ability on TIMSS mathematics and science achievement at the fourth grade: An analysis by item reading demands. In M. O. Martin & I. V. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—implications for early learning* (pp. 67–108). Amsterdam & Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College; International Association for the Evaluation of Educational Achievement.

Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement.

Nilsen, T., & Gustafsson, J.-E. (2014). School emphasis on academic success: Exploring changes in science performance in Norway between 2007 and 2011 employing two-level SEM. *Educational Research and Evaluation*, *20*(4), 308–327. doi:10.1080/13803611.2014.941371

Olson, J. F., Martin, M. O., & Mullis, I. V. (2008). *TIMSS 2007 technical report*. Amsterdam: International Association of Educational Achievement & TIMSS & PIRLS International Study Center; Lynch School of Education, Boston College.

Papanastasiou, C., Plomp, T., & Papanastasiou, E.C. (Eds.) (2011). *IEA 1958-2008: 50 years of experiences and memories*. The Hague, the Netherlands: The International Association for the Evaluation of Educational Achievement.

Pawson, R. & Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks, CA: Sage.

Postlethwaite, T. N. (1995). International empirical research in comparative education. An example of the studies of the International Association for the Evaluation of Educational Achievement (IEA). *Tertium Comparationis*, *1*, 1–19.

Quellmalz, E. S., Timms, M. J., Buckley, B. C., Davenport, J. L., Loveland, M., & Silberglitt, M. D. (2012). 21st century dynamic assessment. In M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st century skills* (pp. 55–89). Charlotte, NC: Information Age Publishing.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167–190. doi:10.1007/bf02295939

Rosén, M., & Gustafsson, J.-E. (2016). Is computer availability at home causally related to reading achievement in grade 4? A longitudinal difference in differences approach to IEA data from 1991 to 2006. *Large-scale Assessments in Education*, *4*(1), 1–19. doi:10.1186/s40536-016-0020-8

Ruhose, J., & Schwerdt, G. (2016). Does early educational tracking increase migrant-native achievement gaps? Differences-in-differences evidence across countries. *Economics of Education Review*, *52*, 134–154. doi:10.1016/j.econedurev.2016.02.004

Rutkowski, D., & Rutkowski, L. (2010). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative & International Education*, *8*(3), 259–278. doi:10.2304/rcie.2013.8.3.259

Rutkowski, L. (2016). Introduction to special issue on quasi-causal methods. *Large-scale Assessments in Education*, *4*(8). doi 10.1186/s40536-016-0023-5

Rutkowski, L., Gonzales, E. J., Joncas, M., & Davier, M. v. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151. doi:10.3102/0013189X10363170

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measures approach to reporting and interpreting PISA results. *Educational Researcher*, *45*(4), 252–257. doi:10.3102/0013189X16649961

Scherer, R., & Gustafsson, J.-E. (2015). Student assessment of teaching as a source of information about aspects of teaching quality in multiple subject domains: An application of multilevel bifactor structural equation modeling. *Frontiers in Psychology*, *6*(1550). doi:10.3389/fpsyg.2015.01550

Schlotter, M., Schwerdt, G., & Woessmann, L. (2011). Econometric methods for causal evaluation of education policies and practices: A non-technical guide. *Education Economics, 19*(2): 109–137. doi:10.1080/09645292.2010.511821

Schwerdt, G., & Wuppermann, A.C. (2011). Is traditional teaching really all that bad? A within-student between-subject approach. *Economics of Education Review*, *30*(2), 365–379. doi:10.1016/j.econedurev.2010.11.005

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, *39*, 37–63. doi:10.1016/j.dr.2015.12.001

Shavit, Y., & Blossfeld, H.-P. (1993). *Persistent inequality: Changing educational attainment in thirteen countries*. Boulder, CO: Westview Press.

Strietholt, R., Gustafsson, J.-E., Rosén, M., & Bos, W. (2014). Outcomes and causal inference in international comparative assessments. In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 9–18). Münster, Germany & New York, NY: Waxmann.

Strietholt, R., Manitius, V., Berkemeyer, N., & Bos, W. (2015). Bildung und Bildungsungleichheit an Halb- und Ganztagsschulen. *Zeitschrift für Erziehungswissenschaft*, *18*(4), 737–761. doi:10.1007/s11618-015-0634-6

Strietholt, R., & Rosén, M. (2016). Linking large-scale reading assessments: Measuring international trends over 40 years. *Measurement: Interdisciplinary Research and Perspectives*, *14*(1), 1–26. doi:10.1080/15366367.2015.1112711

Strietholt, R., Rosén, M., & Bos, W. (2013). A correction model for differences in the sample compositions: The degree of comparability as a function of age and schooling? *Large*-Scale *Assessments in Education*, *1*(1), 1–20. doi:10.1186/2196- 0739-1-1

Torrance, H., Lauder, H., Brown, P., Dilabough, J.-A., & Halsey, A. H. (2006). Globalizing empiricism: What, if anything, can be learned from international comparisons of educational achievement. In H. Lauder, P. Brown, J.-A. Dillabough, & A. H. Halsey (Eds.), *Education, globalization and social change* (pp. 824–834). Oxford: Oxford University Press.

Travers, K., & Westbury, I. (1989). *The IEA study of mathematics I: Analysis of mathematics curricula*. Oxford: Pergamon Press.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, *51*(1), 1–10. doi:10.2307/1129583

UNESCO International Bureau of Education. (2013). *Country dossiers*. Retrieved from http: www.ibe.unesco.org/en/worldwide.html

Scandinavian Journal of Educational Research

ISSN: 0031-3831 (Print) 1470-1170 (Online) Journal homepage: https://www.tandfonline.com/loi/csje20

The Contribution of International Large-Scale Assessments to Educational Research: Combining Individual and Institutional Data Sources

Rolf Strietholt & Ronny Scherer

To cite this article: Rolf Strietholt & Ronny Scherer (2018) The Contribution of International Large-Scale Assessments to Educational Research: Combining Individual and Institutional Data Sources, Scandinavian Journal of Educational Research, 62:3, 368-385, DOI: 10.1080/00313831.2016.1258729

To link to this article: https://doi.org/10.1080/00313831.2016.1258729

Published online: 10 Jan 2017.