Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 4 (2012), No. 2, 143–173
© 2012 Waxmann

Jenny Lenkeit

How effective are educational systems? A value-added approach to measure trends in PIRLS

Abstract

From an educational effectiveness perspective, research based on international large scale assessments has been limited as it neglects to take contextual conditions of educational systems into account. Further, methodological challenges of cross-sectional studies have yet prevented investigations from a longitudinal effectiveness perspective. The paper investigates how effectively educational systems grow, i.e. change, in their performance by applying a methodological approach known from school effectiveness research that captures changes at the country level within repeated cross-sectional data designs. Data from the Progress in International Reading Literacy Study (PIRLS) 2001 to 2006 trend systems is analyzed with hierarchical linear modeling. Effectiveness measures of achievement status in 2006 and of change from 2001 to 2006 are investigated and compared. Results suggest that there are systems which exceed their expected outcomes (status and change) as well as systems which stay below what could have been expected, changing the picture of "high" and "low" performing systems, when contextual conditions and prior performances are taken into account. The study contributes to methodological developments of educational effectiveness research in cross-national assessments. Its results provide complementary information for policymakers to further look at policies, practices, and structures that have favored effectiveness.

Keywords

Cross-national comparisons; Educational effectiveness; Repeated cross-sectional design

Jenny Lenkeit, M.A., Department of Education, University of Hamburg, Binderstraße 34, 20146 Hamburg, Germany
e-mail: j.lenkeit@uva.nl

Jenny Lenkeit

Wie effektiv sind Bildungssysteme? Zur Untersuchung von Entwicklungen in PIRLS mit value-added-Modellen

Zusammenfassung

Aus dem Blickwinkel der Effektivitätsforschung sind bisherige Forschungsansätze mit Daten aus international vergleichenden Studien unbefriedigend, da sie die kontextuellen Bedingungen in einzelnen Bildungssystemen vernachlässigen. Weiterhin fehlen Ansätze längsschnittlicher Betrachtungen, die über deskriptive Analysen hinausgehen. Der Beitrag untersucht, wie effektiv sich Bildungssysteme hinsichtlich ihrer durchschnittlichen Performanz verändern. Hierfür werden methodische Ansätze aus der Schuleffektivitätsforschung herangezogen, welche Veränderungen von Institutionen mit unterschiedlichen Kohorten erfassen können. Trendländer der Progress in International Reading Literacy Study (PIRLS) 2001–2006 werden mit hierarchisch linearen Modellen diesbezüglich untersucht. Effektivitätsmaße für den Leistungsstatus in 2006 und den Leistungszuwachs von 2001 zu 2006 werden analysiert. Die Ergebnisse lassen sowohl Länder, die wider Erwarten hohe Performanz zeigen, als auch solche mit erwartungswidrig niedriger Performanz erkennen und korrigieren das Bild „guter" und „schlechter" Bildungssysteme, wenn Kontextbedingungen und Ausgangslagen berücksichtigt werden. Die Untersuchung trägt methodisch zur Etablierung der Effektivitätsforschung im Rahmen international vergleichender Studien bei. Die Ergebnisse stellen komplementäre Informationen für politische Entscheidungsträger bereit und regen zu weiteren Betrachtungen der Steuerungsmechanismen, Reformlinien und Strukturen an, welche die Qualität und Effektivität von Bildungssystemen bedingen.

Schlagworte

International vergleichende Studien; Effektivitätsforschung; Kohortendesign

1. Introduction

Recent decades have seen a trend towards evaluating and comparing educational systems around the world with large scale assessments (LSAs) of student outcomes in different academic domains and school stages. In 1959 the International Association for the Evaluation of Educational Achievement (IEA) started its first international comparative study with 12 participating educational systems (The Pilot Twelve Country Study; Foshay, Thorndike, Hotyat, Pidgeon, & Walker, 1962). Since then, several new LSAs have emerged and the number and variety of participating educational systems has increased remarkably. In 1995 IEA's first Trends in International Mathematics and Science Study (TIMSS) assessed the achievement of students in 40 participating educational systems (at third, fourth, seventh, eighth, and the final grade of secondary school), followed by the Progress in International

Reading Literacy Study (PIRLS) in 2001 with 35 participating educational systems. TIMSS and PIRLS have repeatedly assessed student performance across educational systems in 4 and 5 year intervals respectively. Sixty-three educational systems and 14 benchmarking entities participated in the latest TIMSS 2011 cycle and 49 educational systems and 7 benchmarking entities in the latest PIRLS 2011 cycle. Further, the OECD (Organisation for Economic Co-operation and Development) launched the first PISA survey (Programme for International Student Assessment) in 2000 with 43 participating educational systems. In 2009 the fourth cycle already included 65 educational systems. Additionally, assessments with a more regional focus such as PASEC (Programme d'Analyse des Systèmes Educatifs de la CONFEMEN) for Francophone Africa, SACMEQ (Southern and Eastern Africa Consortium for Monitoring Educational Quality) for Anglophone Africa, and SERCE (Second Regional Comparative and Explanatory Study) for Latin America have emerged. Schwippert and Lenkeit (2012a) provide a recent overview of studies and participating educational systems.

Generally, the goal of international LSAs is to produce a description of academic outcomes, overall structures, and significant features of educational systems (Mullis, Martin, Kennedy, & Foy, 2007; OECD, 2009; Watermann & Klieme, 2002). International LSAs provide information for policymakers and administrators in order to form decisions concerned with their educational institutions or systems. By revealing deficiencies (as well as strengths) international LSAs often act as initiators of reforms and educational programs within the national systems. For example, Liegmann and van Ackeren (2012) and van Ackeren (2007) showed that a number of reforms aimed at improving schools' context and input quality (e.g. curriculum reforms, teacher qualification) as well as process and output centered strategies (e.g. development of national standards, monitoring systems) emerged as a direct and indirect consequence of PIRLS. Likewise TIMSS (see Howie & Hughes (2000) for the example of South Africa) and PISA (Grek, 2009; Ringarp & Rothland, 2010) have had an influential role in educational policy worldwide.

But there are also limitations related to the information international LSAs can provide. These are related, for example, to differences across educational systems in the school grade or age of the target population, construct equivalence, and scale and measurement equivalence that could potentially introduce a bias in international comparisons (Bechger, van den Wittenboer, Hox, & De Glopper, 1999; Byrne & van de Vijver, 2010; Mislevy, 1995). Furthermore, even if technical aspects of comparative validity are met, cultural (Bank & Heidecke, 2009; Bempechat, Jimenez, & Boulay, 2002; Solano-Flores & Nelson-Barber, 2001) and structural economic (Baker, Goesling, & Letendre, 2002; Caro & Lenkeit, 2012; Chiu, 2007; Chudgar & Luschei, 2009) differences between educational systems preclude researchers from extrapolating international results on the relationship between structures, school processes, and average performance to national contexts. And these differences often impede researchers to make inferences about overall quality of national educational systems.

The field of Educational Effectiveness Research (EER) (Creemers & Kyriakides, 2008; Stevens, 2005; Teddlie & Reynolds, 2000) has established methodological approaches to evaluate quality on school and classroom levels independent of structural economic differences. EER is guided by the conviction that there are factors influencing academic achievement that educators and institutions should not be held accountable for, because they are not amenable to education policy (Ballou, Sanders, & Wright, 2004; Martineau, 2006; Thomas, 1998). These non-malleable factors include individual and compositional socioeconomic and sociocultural characteristics of the student body (Coe & Fitz-Gibbon, 1998; Newton, Darling-Hammond, Haertel, & Thomas, 2010; OECD, 2008). Statistical models should control for these factors in order to produce adjusted measures of school performance and thus provide a "fair comparison" of schools (Nachtigall, Kröhne, Enders, & Steyer, 2008; OECD, 2008). The models also yield a measure of expected performance which is contrasted with the observed performance to produce an indicator of school effectiveness. This approach to identify effective schools and classes independent of their students' characteristics builds the basis for researchers to investigate effectiveness enhancing factors and for policymakers to initiate school developing processes.

This paper attempts to establish a link between the fields of international LSAs and educational effectiveness research by developing effectiveness indicators for educational systems that represent performance that is adjusted for relevant macro-level differences between those systems. With that, the paper seeks to contribute to the analytical approaches for reporting results of international LSA studies. The proposed procedure to measure effectiveness of educational systems is illustrated by examining achievement status and trends with data from PIRLS 2001 and 2006. The results could offer educational stakeholders valuable information about the effectiveness of educational systems irrespective of the socioeconomic conditions in which they operate. Importantly, although measures of effectiveness take economic and developmental differences between educational systems into account, they still are limited by comparability issues that originate from cultural aspects.

2. Educational effectiveness research: The notion of quality and empirical approximations

EER (Creemers & Kyriakides, 2008; Scheerens, 1997) represents an integration of the fields of school effectiveness (i.e. school organization and education policy) (Teddlie & Reynolds, 2000) and research aiming at the classroom level (i.e. teacher behavior, instruction methods, and curriculum analyses) (Campbell, Kyriakides, Muijs, & Robinson, 2003; Opdenakker & van Damme, 2006; Stronge, Ward, & Grant, 2011). With a proceeding awareness of contextual impacts on learning processes, approaches were elaborated that regarded effectiveness as a multilevel phenomenon. These approaches integrated cross-level relationships in the theoretical models of educational effectiveness.

Investigations of educational effectiveness follow a distinctive notion of the quality of classes and schools. This notion rests on evidence that the student intake (reflected by socioeconomic and cognitive characteristics of students) is strongly associated with processes that take place within schools and classrooms (Opdenakker & van Damme, 2007; Stevens, 2005) and thereby with the educational outcome of classes and schools. EER advocates that educators and institutions should not be held accountable for the effect of the student intake, that is, statistical models should control for the student intake in order to evaluate effectiveness (Ballou et al., 2004; Martineau, 2006; OECD, 2008; Thomas, 1998). The identification of effective schools and classrooms is the prerequisite to implement research concerned with effectiveness enhancing factors and to carry out in-depth investigations on their specific structural and process characteristics (Bonsen, Bos, & Rolff, 2008; Mintrop & Trujillo, 2007). The dynamic model of educational effectiveness by Creemers and Kyriakides (2008) guides the identification of effectiveness enhancing factors and provides an understanding of the mechanisms at work. In terms of policy, the empirical evidence provided by EER lays the basis for the design and implementation of educational interventions (Lind, 2004; Mintrop & Trujillo, 2007).

Methodologically, different approaches exist from which to derive effectiveness measures, depending on the study design. Models for cross-sectional data control for the student intake by including family background characteristics (OECD, 2008) such as social and economic status indicators, which usually are strong predictors of educational outcomes (Nachtigall et al., 2008; Sirin, 2005). Researchers that fall back on data designs with at least two measurement points consider student intake by means of controlling for prior attainment. Measures of prior attainment are considered to be the most important and accurate factor that affects subsequent achievement (Thomas & Mortimore, 1996; Sammons, 1996). When more measurement points are available it is possible to estimate achievement growth of students. The growth approach is regarded by educational researchers as most appropriate to assess effectiveness and has been extensively applied in EER (Goldschmidt, Choi, Martinez, & Novak, 2010; Teddlie, Reynolds, & Sammons, 2000; Zvoch & Stevens, 2008). Achievement growth rates, though, not only result from school effects, but they are also a function of family background (Alexander, Entwisle, & Olsen, 2001; Caro & Lehmann, 2009; Cortina, Carlisle, & Zeng, 2008; Hecht, Burgess, Torgesen, Wagener, & Rashotte, 2000). But unlike cross-sectional approaches, the growth model reflects the fact that learning itself is a cumulative process (Kennedy & Mandeville, 2000; Willet, 1988).

In both cross-sectional and longitudinal approaches, the observed outcome of units is evaluated against the expected outcome for the characteristics of the student intake. The model's error term captures the difference between the observed and expected outcome and, given that the model is reasonably specified, directly provides a measure of effectiveness (Raudenbush, 2004). The specific understanding of effectiveness is thereby determined by the choice of student intake variables

in that different model specifications would lead to different effectiveness measures (Coe & Fitz-Gibbon, 1998).

3. Measures of educational system performance in international LSAs

International LSAs provide information about the performance of educational systems and the student, family, and school factors related to the performance results. International reports of different studies (Mullis, Martin, & Foy, 2008; Mullis et al., 2007; OECD, 2010a) list average achievement scores and their distribution for the participating educational systems. Typically, results are broken down to subgroups along key characteristics (e.g. gender, social background, and individual dispositions). The reports further provide information about macro-level indices such as GDP (Gross Domestic Product), HDI (Human Development Index) and educational system indicators (e.g. school entrance age, average class size). International reports thus provide policymakers with information about the position of their educational system in an international context.

According to Postlethwaite (1999) international LSAs also intend to distinguish characteristics and policies that are capable of explaining differences in average achievement across educational systems. However, insufficient recommendation is provided about how knowledge of other systems' characteristics can be utilized to remedy own weaknesses (Jaworski & Phillips, 1999; Mislevy, 1995; Shorrocks-Taylor, 2010). For example, the high performance of Finish students has raised great interest in the characteristics and structures of the Finish educational system. But it is questionable whether lessons from the Finish case can be extrapolated to other national contexts (Beese & Liang, 2010; Kobarg & Prenzel, 2009; Waldow, 2010). In general, it seems difficult to explain, conclude and predict achievement differences between educational systems with data from international comparative assessments.

The reasons are manifold. For example, critics caution against cross-cultural validity issues such as language, task contents and formulations (Bank & Heidecke, 2009; Solano-Flores & Nelson-Barber, 2001). Leung and van de Vijver (2008) and others further discuss threats to construct invariance of self-reported beliefs, attitudes and practices in cross-national comparative studies that arise, e.g. from differences in construct conceptualization and the way these are operationalized (Artelt, 2005; Bempechat et al., 2002; OECD, 2010b; Tan & Yates, 2007). Ercikan (2002) as well as Grisay, Gonzalez, and Monseur (2009) further identify and discuss differential item functioning as a threat to cross-cultural validity in multi-language assessments. Also, the repeated cross-sectional design of international LSAs and the intention to observe trends within and between educational systems has evoked discussions about scaling methods for repeated measurements and the validity of reported trends (Gebhardt & Adams, 2007; Robitzsch, 2010).

Furthermore, cultural, developmental, and economic differences between educational systems make it difficult for researchers to detect and generalize effective structures and processes across educational systems (Postlethwaite, 1999). Several studies have shown the association of economic and developmental factors with the performance of educational systems (Baker et al., 2002; Caro & Lenkeit, 2012; Chiu, 2007; Chudgar & Luschei, 2009). As for differences in culture, societies differ in their historical development, their institutional and systemic structures. Accordingly, societies differ in the functions they attribute to education and academic domains. This is reflected in their societal and political-ideological appreciation (Bempechat et al., 2002). Solano-Flores and Nelson-Barber (2001) state that the functioning and structures of knowledge are acquired and expressed according to cultural patterns and notions. However, for the Reading Literacy Study, Postlethwaite (1999) notes that while controlling for all other variables of the study, the inclusion of country IDs independently accounted for only 4 % of the explained variance between the schools of all educational systems. "If the ID reflected aspects of being a German or a Finn or a Briton (...) then the school systems of the world are not much affected by national culture" (Postlethwaite, 1999, p. 52). The relevance of notational cultural characteristics thus appears to be limited.

In sum, the descriptive information provided in international assessment reports is useful to compare absolute performance levels between educational systems and to position them in an international context. But, this information seems to be of less use for policymakers, who demand policy-relevant information about the effectiveness of systems. The informational gaps in the reporting of international LSA results can be somewhat addressed with the theoretical and methodological accomplishments of EER.

4. From educational effectiveness to educational system effectiveness

4.1 Past studies and their limitations

In the literature we find approximations to link effectiveness research and cross-national comparative studies. Scheerens (2006) has discussed the potential of international LSAs for conducting effectiveness research that would originate from developing and assessing indicators of accountability and evaluation arrangements and infrastructure at national levels. One of the earlier empirical investigations on educational effectiveness in the contexts of cross-national research was conducted by Postlethwaite and Ross (1992) using IEA's Study of Reading Literacy. However, rather than effectiveness of the systems themselves, they examined characteristics of effective schools across different educational systems. Also, they did not consider the hierarchical nature of the data in multilevel models. A major finding was

nevertheless that indicators which distinguish some schools as more effective than others differ across educational systems.

Few studies such as the International System for Teacher Observation and Feedback (ISTOF; Sammons, 2006) and the International School Effectiveness Research Project (ISERP; Reynolds, 2006) have explicitly implemented a study design for investigating educational effectiveness across educational systems. Both focus on the new insights into educational effectiveness from comparative research as well as the possible validation and transfer of theoretically developed factors of school and teacher effectiveness to other systems. While a shortcoming of the ISTOF study is its cross-sectional design, ISERP follows the same students of nine educational systems over two years. This research design is however difficult to implement in international LSA studies including many participating educational systems.

Acknowledging the strong association of socioeconomic characteristics with average achievement, PISA (OECD, 2010c) adjusts achievement scores for the effect of students' family and home background (as represented by the composite of the PISA social, economic and cultural status) and compares predicted average achievement scores across educational systems. Although, this adjustment represents essentially the idea of operationalizing effectiveness measures, the approach misses to include macro-level factors that also determine cross-national differences in average achievement. Moreover, PISA considers cross-sectional data only.

Research conducted by van Damme, Liu, Vanhee, and Putjens (2010) essentially takes up the idea of addressing educational effectiveness at the level of educational systems in a longitudinal perspective by asking whether changes in age, socioeconomic status, and class size explain changes in average reading achievement from PIRLS 2001 to 2006. They miss however, to investigate differences between educational systems that remain despite removing differences within educational systems over time and their analytic strategy is restricted to a separate model for each system.

The use of longitudinal data to measure educational system effectiveness is important, because it allows controlling for prior performance and educational systems' economic characteristics. In the same way as student intake is associated with achievement growth in school effectiveness models, it is assumed that the systems' economic and developmental characteristics are related to their potential to change, meaning for example, implementing reforms or increasing educational spending. However, studies interested in educational system effectiveness have not been concerned with the operationalization of effectiveness measures obtained from longitudinal data.

4.2 A model for educational system effectiveness

Willms and Raudenbush (1989) have proposed a statistical model that adapts well to the study of effectiveness with the longitudinal data from international LSAs. Concerned with the stability of school effects on levels of attainment they exam-

ined different cohorts of students in a particular grade in consecutive years (see also Kelly & Monczunski, 2007; Luyten, 1994). The multilevel models nested students into cohorts and cohorts into schools. Likewise, international LSAs have a multilevel design which nests students into schools, schools into cohorts (i.e. survey cycles), and cohorts into educational systems. While the multilevel structure is not the same, the model by Willms and Raudenbush (1989) can be adapted to evaluate educational system effectiveness for different cohorts of students across systems and over time.

Empirical specifications of models need to control for variables at the different levels (Kelly & Monczunski, 2007; Willms & Raudenbush, 1989). Apart from socioeconomic status (SES) controls at student and school level, models should account for sociodemographic characteristics of educational systems. For example, differences in the average age of students in educational systems need to be controlled. It has been shown that younger students obtain on average lower achievement than older students despite equal years of schooling (Breznitz & Teltsch, 1989; Cliffordson & Gustafsson, 2007; Jones & Mandeville, 1990) and the average age of students can change between cycles due to grade entrance policies. Further, economic and developmental status of educational systems are viewed as non-malleable factors that are associated with average achievement (Baker et al., 2002; Caro & Lenkeit, 2012; Chiu, 2007; Chudgar & Luschei, 2009) and should be controlled, too. Characteristics such as educational expenditure or central educational governance are viewed to be malleable and are therefore not categorized as control variables.

Ultimately, the proposed model yields a single effectiveness measure for each educational system. Unlike the cross-sectional approach where educational system effectiveness is measured in relation to performance at a certain point in time, the model conceives effectiveness as a cumulative process related to performance change.

5. Aim of the paper

The aim of this paper is to demonstrate a methodological approach that purges the effect of economic and developmental differences of educational systems and introduces a longitudinal perspective to study their effectiveness. It moreover introduces a notion of quality that is widely accepted in effectiveness research. By defining effectiveness as the relation of the observed and expected outcome, it moves beyond the comparison of unadjusted achievement scores.

The approach is demonstrated with data from PIRLS 2001 and 2006. Although other international LSAs provide data sets with more measurement points, PIRLS was chosen for the following reason. A recent project on the impact of PIRLS showed that reform measures undertaken in educational systems are accompanied by insecurities of policymakers regarding the evidence on which their deci-

sions were based (Schwippert & Lenkeit, 2012b). The application of effectiveness approaches to international LSAs adds relevant information for policymakers to this evidence base. For example, international reports of PIRLS show that South Africa's average performance is well behind that of Germany or Hong Kong, SAR. While that is relevant information in itself, policymakers would benefit from information which indicated, for example, that despite its contextual conditions South Africa was very effective, i.e. it exceeded its expected outcome, whereas Germany may lack behind of what could have been expected, considering its economic and developmental status. Identifying effective systems is the basis for further in-depth investigations about the structures and processes that lead to better performance. Further, the project revealed that reform measures and programs could be categorized as direct and/or indirect effects of PIRLS. However, no empirical evaluation of their impact had taken place. The analytical approach provides a complementary evidence source for policymakers to specify the impact of reform measures and programs in further investigations.

6. Methodological approach

6.1 Data and measures

Data stem from the educational systems which participated in both 2001 and 2006 cycles of PIRLS. The United States was excluded from the 28 trend participants as it did not assess all of the necessary background data. Morocco was excluded because background data was available only in the cycle of 2006. The two Canadian provinces Ontario and Quebec were excluded from the analyses as their inclusion would have overrepresented Canada as a country in the sample while at the same time not being representative for Canada as a whole.

The overall analytic sample was organized in two data sets for reasons of model specification. First, considering only the cross-sectional data of the 2006 cohorts, effectiveness measures were obtained that relate to the educational system's average achievement status in 2006 (24 educational systems, 4.073 schools, 110.974 students). Secondly, to estimate the effectiveness of change rates of achievement from 2001 to 2006 a pooled data set with cohorts of both assessment cycles was created (24 educational systems, 7.850 schools, 210.187 students).

Socioeconomic status (SES, SESSM at school level) is a weighted composite of parents' highest education level, parents' highest occupation status, parents' highest employment status, number of books at home and four variables of home possessions answered by students across all educational systems (personal computer, study desk, own books, daily newspaper). Missing rates on these variables considerably varied between educational systems (see Table 1).

Table 1: Missing rates of constitutive SES index variables per educational system and cohort, in percent

Educational System	Cohort	Number of books at home	Parents' highest education level	Parents' highest employment status	Parents' highest occupation status	Home possessions: Personal computer	Home possessions: Study desk	Home possessions: Own books	Home possessions: Daily newspaper
Bulgaria	2001	4.8	6.1	15.5	27.3	2.1	2.0	1.7	2.2
	2006	4.2	5.0	10.3	6.5	2.6	2.4	2.3	2.5
England	2001	45.2	49.1	52.8	55.7	2.2	0.8	0.8	0.8
	2006	53.2	56.1	56.0	54.4	1.4	1.6	1.4	1.4
France	2001	11.0	21.7	23.9	34.5	4.3	3.7	3.8	4.2
	2006	8.3	13.5	15.8	12.4	3.6	3.3	3.3	4.0
Germany	2001	13.3	35.9	29.1	34.6	6.9	5.8	5.6	5.9
	2006	13.6	25.2	23.0	16.9	10.1	9.4	9.2	9.8
Hong Kong, SAR	2001	5.9	8.6	25.7	29.0	3.1	3.2	3.3	3.5
	2006	3.8	3.7	9.5	4.0	2.9	2.9	2.9	2.9
Hungary	2001	4.7	9.3	17.7	28.5	2.4	2.0	2.3	2.5
	2006	9.4	12.8	15.2	11.5	1.8	1.2	1.3	1.4
Iceland	2001	16.9	17.1	20.1	28.9	5.4	4.6	4.2	4.5
	2006	24.4	24.8	26.5	24.8	2.0	2.0	2.0	2.2
Iran	2001	16.9	17.1	20.1	28.9	5.4	4.6	4.2	4.5
	2006	3.0	3.9	28.9	8.2	3.9	2.8	3.4	3.7
Israel	2001	53.4	57.3	70.0	68.1	7.7	7.2	7.6	7.8
	2006	38.3	43.3	46.8	44.7	8.5	8.0	8.3	8.2
Italy	2001	3.6	4.2	23.2	16.2	1.6	1.0	1.3	1.6
	2006	4.8	6.9	14.6	8.2	1.8	1.6	1.6	2.0
Latvia	2001	4.6	11.7	22.3	31.3	3.3	1.8	1.5	2.0
	2006	5.9	9.4	11.4	7.6	1.3	1.2	1.1	1.3
Lithuania	2001	2.4	4.2	23.6	22.1	2.5	1.5	1.4	1.9
	2006	2.7	4.6	9.8	5.1	1.0	1.0	0.8	1.0
Macedonia	2001	23.0	36.6	50.2	42.9	10.3	7.2	7.4	7.2
	2006	4.6	14.6	25.8	15.0	8.3	7.1	7.7	7.8
Moldova	2001	2.4	7.2	29.2	26.6	3.0	1.7	1.8	2.2
	2006	4.5	5.6	16.1	5.2	3.2	1.7	2.0	2.5
Netherlands	2001	35.4	37.2	43.1	44.1	1.6	1.4	1.5	1.5
	2006	31.7	36.8	35.8	32.3	0.8	0.9	1.0	1.3
New Zealand	2001	16.3	18.7	28.3	31.7	5.7	3.2	3.1	3.3
	2006	36.7	38.3	42.1	40.1	3.9	3.9	3.8	4.3
Norway	2001	9.4	10.4	17.2	18.8	2.1	2.0	1.9	2.3
	2006	8.3	11.1	11.5	9.9	5.6	5.8	5.8	6.4
Romania	2001	3.0	11.0	25.8	9.5	3.8	1.9	2.0	2.1
	2006	2.8	5.3	10.2	5.1	2.0	1.7	1.7	2.2

Educational System	Cohort	Number of books at home	Parents' highest education level	Parents' highest employment status	Parents' highest occupation status	Home possessions: Personal computer	Home possessions: Study desk	Home possessions: Own books	Home possessions: Daily newspaper
Russian Federation	2001	1.4	1.6	13.1	15.5	2.5	1.3	1.4	1.5
	2006	1.1	3.0	6.3	1.4	1.2	0.7	0.7	0.9
Scotland	2001	37.5	39.2	45.3	46.9	3.0	2.0	1.8	2.0
	2006	48.3	53.4	51.8	49.2	1.0	1.0	1.0	1.3
Singapore	2001	2.2	9.4	15.8	24.8	1.1	1.1	1.1	1.1
	2006	2.4	4.6	8.4	4.9	1.0	1.1	1.0	1.0
Slovak Republic	2001	3.2	6.4	13.7	21.1	2.4	1.5	1.4	1.7
	2006	3.4	6.0	9.1	5.7	1.3	0.9	1.0	1.2
Slovenia	2001	3.3	5.3	7.9	20.9	1.7	0.6	0.9	1.1
	2006	5.5	6.7	7.9	9.5	0.9	0.7	0.8	0.9
Sweden	2001	9.3	9.7	16.3	15.7	3.4	2.6	2.4	4.2
	2006	7.2	15.6	10.6	9.3	1.6	1.5	1.7	2.1

Multiple imputation methods were used to account for missing data uncertainty (Rubin, 1987). Five imputed data sets were created using data augmentation (DA) (Schafer & Olsen, 1998). DA is an iterative simulation technique, a special kind of Markov Chain Monte Carlo (MCMC) that has a strong resemblance to the Expected Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). The imputation technique draws on information from the observed part of the data set to create plausible versions of the complete data set (Schafer & Olsen, 1998). Data on reading performance and other educational home activities were included in the imputation model.[1] Data was imputed separately for each country in a pooled data set including both cohorts, in order to take account of the specific relationships of the variables with each other and the achievement variable. The SES index was created jointly for all educational systems applying factor analyses to each imputed dataset. Point averages from the five imputed data sets yielded a reliability of $\alpha = .674$ and indicated that constituent items explained 32.2 % of the latent SES construct. Both imputed and non-imputed data showed a very similar reliability ($\alpha = .657$ and 30.6 % explained variance for non-imputed data). The final SES index has a mean of 0 and a standard deviation of 1 for the overall analytic sample. In the course of five years the average SES had increased for all educational systems. The increase between the two cohorts ranged from a minimum of 0.03 % for Germany to a maximum of 26.1 % of the SES scale for Hong Kong, SAR.

1 The following items were considered for the imputation model: Student questionnaire: How often do you talk with your family about what you are reading?, About how many children's books are there in your home?; Home questionnaire: About how many books are there in your home?, Before your child began <ISCED Level 1>, how often did you or someone else in your home read books with him or her?, How often do you or someone else in your home discuss your child's classroom reading work with him/her?, How often do you or someone else in your home go to the library or a bookstore with your child?, How often do you or someone else in your home help your child with reading for school?

Table 2: Descriptives of average achievement, SES, age, and HDI by educational system and cohort

Educational system	Achievement 2001			SES 2001		Age 2001		Achievement 2006			SES 2006		Age 2006		Age difference	HDI 2006
	M	(*SE*)	*SD*	*M*	(*SE*)*	*M*	*SD*	*M*	(*SE*)	*SD*	*M*	(*SE*)*	*M*	*SD*	*M*	M
Bulgaria	550	(3.80)	82.5	-0.44	(0.02)	11.0	0.6	547	(4.40)	82.7	-0.21	(0.02)	11.0	0.5	-0.01	0.729
England	553	(3.40)	86.5	0.12	(0.02)	10.3	0.3	539	(2.60)	86.9	0.31	(0.02)	10.3	0.3	0.00	0.860
France	525	(2.40)	70.5	-0.10	(0.02)	10.1	0.5	522	(2.10)	66.6	0.19	(0.01)	10.1	0.5	-0.02	0.842
Germany	539	(1.90)	67.3	0.18	(0.01)	10.6	0.5	548	(2.20)	67.0	0.21	(0.01)	10.6	0.5	-0.07	0.881
Hong Kong. SAR	528	(3.10)	62.8	-0.63	(0.01)	10.3	0.8	564	(2.40)	59.3	-0.03	(0.01)	10.1	0.5	-0.17	0.849
Hungary	543	(2.20)	65.8	0.19	(0.01)	10.8	0.5	551	(3.00)	70.2	0.34	(0.01)	10.8	0.5	0.08	0.802
Iceland	512	(1.20)	74.7	0.47	(0.01)	10.0	0.3	511	(1.30)	68.1	0.81	(0.01)	10.0	0.3	-0.01	0.883
Iran	414	(4.20)	92.2	-1.49	(0.01)	10.5	0.8	421	(3.10)	94.7	-1.24	(0.01)	10.5	0.7	-0.08	0.674
Israel	509	(2.80)	93.7	-0.02	(0.02)	10.1	0.4	512	(3.30)	98.8	0.22	(0.02)	10.1	0.4	-0.02	0.864
Italy	541	(2.40)	71.1	-0.24	(0.01)	9.9	0.4	551	(2.90)	67.9	-0.07	(0.02)	9.9	0.3	0.01	0.844
Latvia	545	(2.30)	61.5	0.04	(0.01)	11.1	0.5	541	(2.30)	62.6	0.46	(0.01)	11.1	0.5	-0.03	0.771
Lithuania	543	(2.60)	64.3	-0.07	(0.02)	11.0	0.5	537	(1.60)	56.9	0.30	(0.01)	11.0	0.4	-0.04	0.780
Macedonia	442	(4.60)	103.1	-0.84	(0.02)	10.7	0.4	442	(4.10)	101.3	-0.49	(0.02)	10.7	0.4	-0.03	0.684
Moldova	492	(4.00)	75.2	-0.89	(0.02)	11.0	0.5	500	(3.00)	69.0	-0.69	(0.02)	11.0	0.5	0.04	0.613
Netherlands	554	(2.50)	57.3	0.05	(0.01)	10.4	0.5	547	(1.50)	53.0	0.38	(0.01)	10.4	0.5	0.03	0.882
New Zealand	529	(3.60)	93.4	0.23	(0.02)	9.6	0.4	532	(2.00)	87.0	0.45	(0.01)	10.6	0.3	0.98	0.898
Norway	499	(2.90)	81.1	0.57	(0.01)	10.0	0.3	498	(2.60)	66.6	0.78	(0.01)	10.0	0.3	0.01	0.934
Romania	512	(4.60)	89.8	-0.77	(0.02)	11.1	0.5	489	(5.00)	91.5	-0.59	(0.02)	11.1	0.5	-0.07	0.743
Russian Federation	528	(4.40)	66.4	-0.12	(0.01)	10.4	0.6	565	(3.40)	68.8	0.24	(0.01)	10.9	0.5	0.49	0.700
Scotland	528	(3.60)	84.2	0.00	(0.02)	9.8	0.3	527	(2.80)	79.9	0.31	(0.01)	9.9	0.3	0.03	0.842
Singapore	528	(5.20)	91.8	0.02	(0.01)	10.0	0.4	558	(2.90)	76.7	0.27	(0.01)	11.0	0.4	1.02	0.832
Slovak Republic	518	(2.80)	70.2	-0.11	(0.01)	10.4	0.5	531	(2.80)	74.2	0.14	(0.01)	10.5	0.5	0.02	0.803
Slovenia	502	(2.00)	71.7	0.10	(0.02)	9.9	0.4	522	(2.10)	70.7	0.27	(0.01)	9.9	0.3	0.01	0.819
Sweden	561	(2.20)	65.8	0.59	(0.01)	11.0	0.3	549	(2.30)	63.6	0.80	(0.01)	11.0	0.3	0.02	0.885

Note. * = Estimation errors for the SES-index were calculated by integrating estimates from imputation sets based on Rubin's formulas.

Reading Achievement (READ) is the dependent variable represented by five plausible scores calculated using Item Response Theory (Martin, Mullis, & Kennedy, 2007). To accurately measure trends, the means and standard deviations of the link scores (i.e. plausible values for trend systems) for all five PIRLS scales were made to match the means and standard deviations of the scores reported in the 2001 assessment (Martin, Mullis, & Kennedy, 2007). The plausible values have a mean of 500 and a standard deviation of 100 in PIRLS 2001. Average reading scores of included educational systems are presented in Table 2. Each of the five plausible values was allocated to one of the five data sets that were created through the multiple imputation procedure described above.

Cohort (COHORT) is a dichotomous variable that differentiates between students assessed in PIRLS 2001 (-1) and those assessed in the PIRLS 2006 (0).

Age (AGE) is the combination of students' year and month of birth and represents students' age at the measurement point.

Age difference (AGED) represents differences in age of student cohorts at the system level.

Human Development Index (HDI) is a composite of three dimensions and four indicators (Health: life expectancy at birth; Education: mean years of schooling, expected years of schooling; Living standards: gross national income per capita) for the year 2006. Information has been retrieved from the website of the United Nations Development Report Programme (HDRO, n.d.).

6.2 Models

Models were estimated by means of hierarchical linear modeling accounting for the multilevel structure of the data (Bryk & Raudenbush, 2002). As the interest of investigation is related to effects on the educational system level, covariates at student and school level were grand mean centered to control for student and school effects in the results on educational system level effects (Bryk & Raudenbush, 2002; Enders & Tofighi, 2007). Data was also weighted at student level with the "student senate weight". The student senate weight is a linear transformation of the total student weight, which comprises the selection probability of students in classrooms and classrooms in schools (Martin et al., 2007). It thus takes into account the two-stage probability-proportional-to-size (PPS) sampling design applied in PIRLS (ibid.). Additionally, student senate weight adjusts for different population sizes of educational systems in cross-country analysis (Rutkowski, Gonzalez, Joncas, & von Davier, 2010). Measures of effectiveness were adjusted by reliability with the Empirical Bayes estimator (Bryk & Raudenbush, 2002; Lindley & Smith, 1972).[2]

2 The Empirical Bayes estimator corrects unreliable estimates by pulling them closer to the average estimate. Unreliable estimates might occur, e.g. when sample sizes for schools (or more general units) are small and extreme values for these schools are more likely to occur by chance (Bryk & Raudenbush, 2002).

The specification for the unconditional model is:

$$READ_{ijk} = \pi_{0jk} + e_{ijk} \quad (1)$$

$$\pi_{0jk} = \beta_{00k} + u_{0jk} \quad (2)$$

$$\beta_{00k} = \gamma_{000} + u_{00k} \quad (3)$$

where $READ_{ijk}$ is the reading performance of student i in school j in educational system k and e_{ijk} is the error term (1). Parameter π_{ojk} is the mean achievement of school j in system k. At the school level π_{ojk} is a function of average achievement in system k (β_{ook}) and the error term that represents the schools deviation from the expected average achievement (u_{ojk}) (2). γ_{ooo} represents the average achievement across educational systems and u_{ook} represents the system's deviation from the expected average achievement across systems (3).

To obtain effectiveness measures of educational systems for the 2006 cohort of PIRLS SES is controlled for at individual and school level. An index of SES at the level of educational systems is conceptually not meaningful; instead differences between the participants' developmental status at the system level were taken into account by controlling for HDI. While GDP would indicate purely economic status at the system level, HDI also indicates the social-developmental status and can thus be viewed as an approximation to SES. Further, age differences of students between educational systems were controlled for. The unconditional model is respecified as follows to represent the conditional model for the 2006 cohort:

$$READ_{ijk} = \pi_{0jk} + \pi_{1jk}SES_i + e_{ijk} \quad (4)$$

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}SESSM_j + u_{0jk} \quad (5)$$

$$\pi_{1jk} = \beta_{10k} \quad (6)$$

$$\beta_{00k} = \gamma_{000} + \gamma_{001}AGE_k + \gamma_{002}HDI_k + u_{00k} \quad (7)$$

$$\beta_{01k} = \gamma_{010} \quad (8)$$

$$\beta_{10k} = \gamma_{100} \quad (9)$$

Parameter π_{ojk} is the mean achievement of school j in system k and parameter π_{1jk} is the expected increase of the reading score for a one unit increment in SES (1 SD) and represents the degree of relationship between the individual SES and achievement (4). e_{ijk} is the error term that represents a student's deviation from the expected average achievement. The relationship is fixed across schools (6) and systems (9). Similarly, at the school level β_{ook} is the mean achievement of system k (5).

In the same equation parameter β_{01k} is the expected increase of the school reading score for a one unit increment in school mean SES (1 SD) and represents the degree of relationship between school mean SES and achievement. The relationship is fixed across systems (8). u_{0jk} is the error term that represents the schools deviation from the expected average achievement. γ_{000} represents the average achievement across educational systems with an age cohort and HDI equal to the grand mean (7). γ_{001} and γ_{002} represent the degree of relationship of the average age of the student cohort and HDI, respectively, with the average achievement. u_{00k} represents the system's deviation from the expected average achievement across systems, taken the included covariates of the model into account (7). It is the effectiveness measure in 2006 based on the cross-sectional data.

To obtain effectiveness measures for change scores of the educational systems the model of Willms and Raudenbush (1989) was adapted to the PIRLS 2001 and 2006 data set. Theoretically, if we had data for several cohorts, then the model would include four levels: students nested in schools, schools in cohorts, and cohorts in educational systems. But the two cohorts (i.e. PIRLS 2001 and 2006) provide insufficient variation to create a new level and cohort differences were controlled with a dummy indicator. First a cohort-only model is specified by altering equation (2) of the unconditional model as follows:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(COHORT)_j + u_{0jk} \qquad (10)$$

Where β_{01k} is the average change in performance from 2001 to 2006 (10). β_{01k} varies between the systems as indicated by u_{01k} (11).

$$\beta_{01k} = \gamma_{010} + u_{01k} \qquad (11)$$

The conditional model for change between 2001 and 2006 thus consists of three levels, similar to equations (4)–(9), but additionally includes a cohort covariate on school level.

Equations (5) and (8) are respecified as:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(COHORT)_j + \beta_{02k}(SESSM)_j + u_{0jk} \qquad (12)$$

$$\beta_{02k} = \gamma_{020} \qquad (13)$$

with cohort effects on the school reading average, $\beta_{01k,}$ varying between systems as a result of age differences, HDI, and random differences (14).

$$\beta_{01k} = \gamma_{010} + \gamma_{011}(AGED)_k + \gamma_{012}(HDI)_k + u_{01k} \qquad (14)$$

u_{01k} is then the system's deviation from expected cohort effect, that is the grand mean cohort effect (14). This deviation essentially represents the measure of effectiveness for the achievement change between 2001 and 2006.

7. Results

Table 3 gives an overview of model results for effectiveness of systems for the 2006 cohort for the unconditional and the conditional model as described in equations (1) to (6). The overall mean achievement across educational systems is 521.3 score points. Twenty-four percent of the overall achievement variance is attributed to differences between schools and 16.3 % to differences between educational systems. The conditional model shows that SES is positively related to reading achievement at both individual and school level. Students score on average 27.6 points higher on the achievement scale if their SES index exceeds the average SES index across educational systems by 1 SD. They additionally score on average 28.1 points higher if their average school SES exceeds the grand mean school SES by 1 SD. After controlling for SES at the student and school level differences in average student age and HDI are, however, not significantly related to average reading achievement across systems in the 2006 cohort. Predictors explain 9.8 % of the student level variance, 47.6 % of school level variance and 35.4 % of system level variance. 64.6 % of the overall system level variance thus remain unexplained and may be subject to other (potentially malleable) system level factors.

Table 3: Three-level regression estimates for reading achievement across educational systems in 2006

Fixed effects		Unconditional model		Conditional model	
		Coefficient	*SE*	Coefficient	*SE*
Intercept		521.3*	6.7	525.3*	5.6
Student level					
	SES			27.6*	1.6
School level					
	SESSM			28.1*	6.9
System level					
	AGE			16.8	13.8
	HDI			-4.7	6.6
Random effects (in %)					
Student level variance		59.7			
School level variance		24.0			
System level variance		16.3			
Explained student level variance				9.8	
Explained school level variance				47.6	
Explained system level variance				35.4	

*$p < .05$.

The educational system level residuals of the conditional model indicate the system's effectiveness. In particular, the residuals represent the deviation of the expected achievement score based on the systems' characteristics on SES, SESSM, AGE and HDI from the predicted score based on the model specifications (u_{ook} in equation 7). Systems with positive residuals exceed their expected outcome. Those with negative residuals stay behind their expected outcome. Figure 1 illustrates the distribution of residuals by educational system. It can be seen that Italy has the highest residual score. Its predicted score exceeds its expected score by 56.4 scale points, and it is therefore the most effective system in the analytic sample. Likewise, the educational systems of Hong Kong, SAR and Bulgaria exceed their expected outcome by 50.2 and 39.1 score points respectively. The systems of Romania, Israel, Lithuania, Slovenia, Moldova, France, and Scotland perform close within the range of their expected outcome. Least successful systems are those of Macedonia (-54.5 score points), Norway (-54.0 score points), Iceland (-38.6 score points), and Iran (-32.2 score points).

Figure 1: Residuals of adjusted achievement scores (i.e. effectiveness measures) in 2006 by educational system

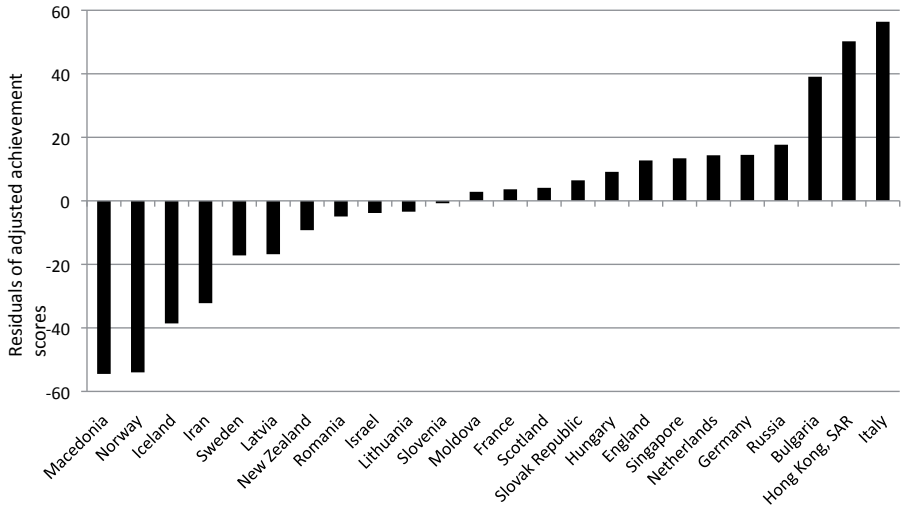


Figure 2 compares the rank order of effectiveness measures (i.e. residuals) with the ones based on observed unadjusted performance. High ranks indicate effective systems and high unadjusted achievement scores respectively. Educational systems have been sorted by observed performance. Sweden's educational system for example is ranked place 20 for its average observed achievement. In terms of its effectiveness, however, it is ranked in place 5 out of 24 educational systems, indicating, that given its contextual conditions Sweden's educational system has more potential than it is able to demonstrate. Considering their socioeconomic conditions the educational systems of Latvia and Hungary would also be ranked 9 and 5 po-

sitions lower, respectively. In contrast, Moldova, Germany, and Romania would be ranked in higher positions (7, 6, and 5 respectively) for their effectiveness than for their unadjusted achievement scores. 8 of the 12 lower achieving educational systems would be assigned to higher ranks and 6 of the 12 higher achieving educational systems would be assigned to lower ranks. Overall, no clear pattern is evident that higher achieving systems systematically underperform or lower achieving systems systematically outperform their expected outcome (and vice versa).

Figure 2: Differences in ranks of unadjusted achievement scores and residuals of adjusted achievement scores (i.e. effectiveness measures) by educational system

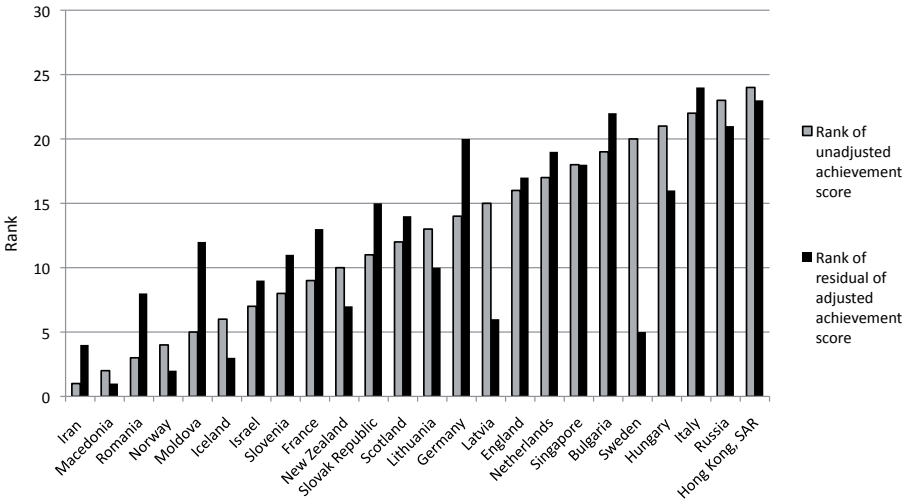


Table 4 gives an overview of model results for the investigation of effectiveness of change from 2001 to 2006. The overall mean achievement of students from both cohorts is 520.1 score points (unconditional model). 25.1 % of the overall variance is attributed to differences between schools and 14.7 % to differences between educational systems. The cohort-only model indicates that the 2006 cohort exceeds the 2001 cohort by an average of 2.2 scale points when no other control variables are included. The average difference between cohorts of 2.2 points is not statistically significant but the random effects indicate that the variation across educational systems is significant. And when SES, SESSM, AGE and HDI are controlled (conditional model), the average performance of the 2006 cohort is significantly lower by 11.6 points. The average characteristics on these variables thus appear to have positively affected the average achievement of the 2006 cohort. The model further shows that across both cohorts, students score on average 26.8 points higher on the achievement scale if their SES index score exceeds the average SES index score across educational systems by 1 SD and they additionally score on average 28.8 points higher if their average school SES exceeds the grand mean school SES by 1 SD. Average differences between systems across cohorts in AGE and HDI

are not significantly associated with differences in average achievement. Variation in the cohort effect across educational systems is partially explained by age differences between cohorts of the educational systems. Considering that students, e.g. in Singapore and New Zealand are older by one year and by half a year in Russia in the 2006 cohort (Mullis et al., 2007) this result is not surprising. HDI does not predict differences in average achievement or achievement change across educational systems, though. The conditional model explains 9.1 % of the student level variance, 46 % of school level variance and 41.1 % of system level variance. The included predictor variables moreover explain 27 % of the cohort parameter variance. Seventy-three percent of the variance in change scores across educational systems thus remains unexplained. This suggests that other factors are associated with differences between average achievement change scores.

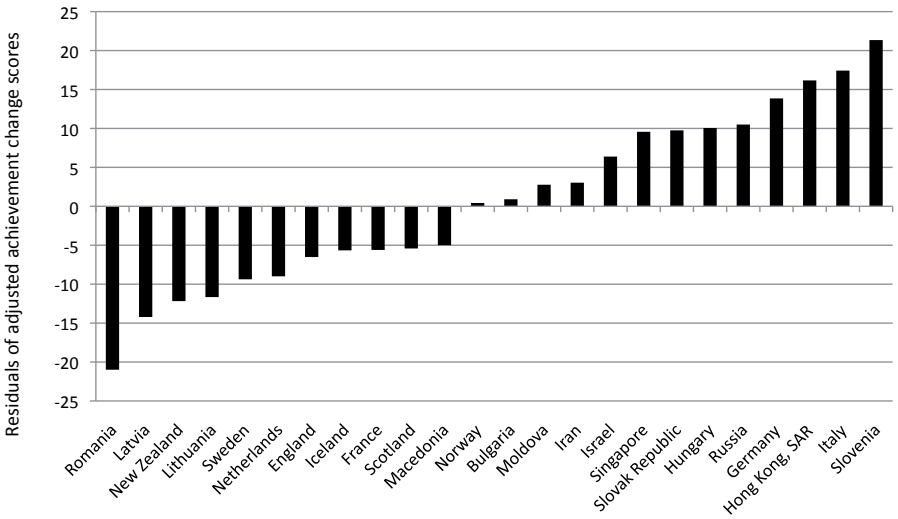
Table 4: Three-level regression estimates for change in reading achievement across educational systems from 2001 to 2006

Fixed effects		Unconditional model		Cohort-only model		Conditional model	
		Coefficient	*SE*	Coefficient	*SE*	Coefficient	*SE*
Intercept		520.1*	6.3	520.1*	6.3	523.4*	10.3
Student level							
	SES					26.8*	1.5
School level							
	SESSM					28.8*	5.7
	COHORT			2.2	2.9	-11.6*	3.6
System level							
	AGE					35.3	37.7
	HDI					1.3	14.6
Cohort (b01)							
	AGED					19.3*	8.1
	HDI					-2.6	2.5
Random effects (in %)							
Student level variance		60.2					
School level variance		25.1					
System level variance		14.7					
Explained student level variance				0.0		9.1	
Explained school level variance				0.7		46.0	
Explained system level variance				-0.1		41.1	
Cohort parameter variance (b02), SD				13.7 *			
Explained parameter variance						27.0	

*$p < .05$

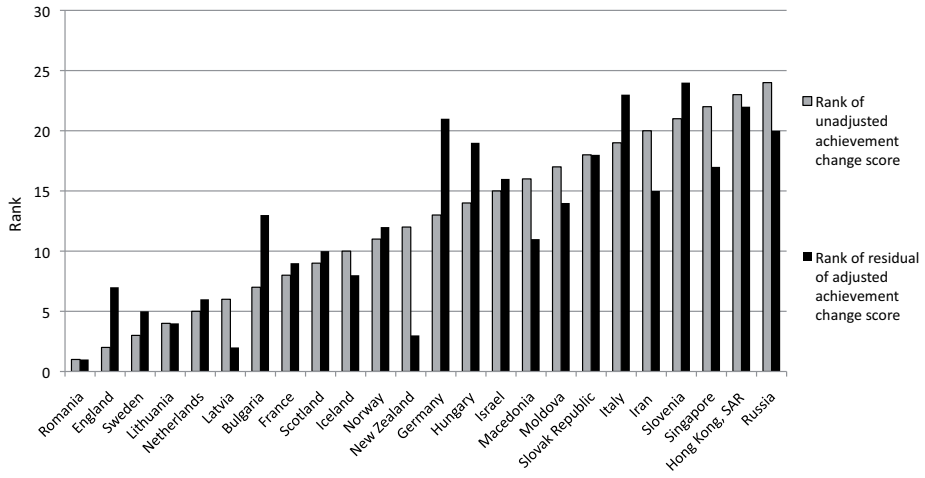
Figure 3 illustrates the distribution of change score residuals. It can be seen that, taking SES, SESSM, AGE and HDI between systems as well as AGED and HDI between cohorts into account, Slovenia's residual change score amounts to 21.3 scale score points. Italy (17.4), Hong Kong, SAR (16.2), and Germany (13.9) for example have also exceeded their expected change score and effectively improved their average performance. The systems of Iran, Moldova, Bulgaria, and Norway perform close within the range of their expected outcome. Romania (-21.0), Latvia (-14.2), New Zealand (-12.2), and Lithuania (-12.7) for example are less effective and have not reached what could have been expected given their contextual conditions.

Figure 3: Residuals of adjusted achievement change scores (i.e. effectiveness measures) from 2001 to 2006 by educational system



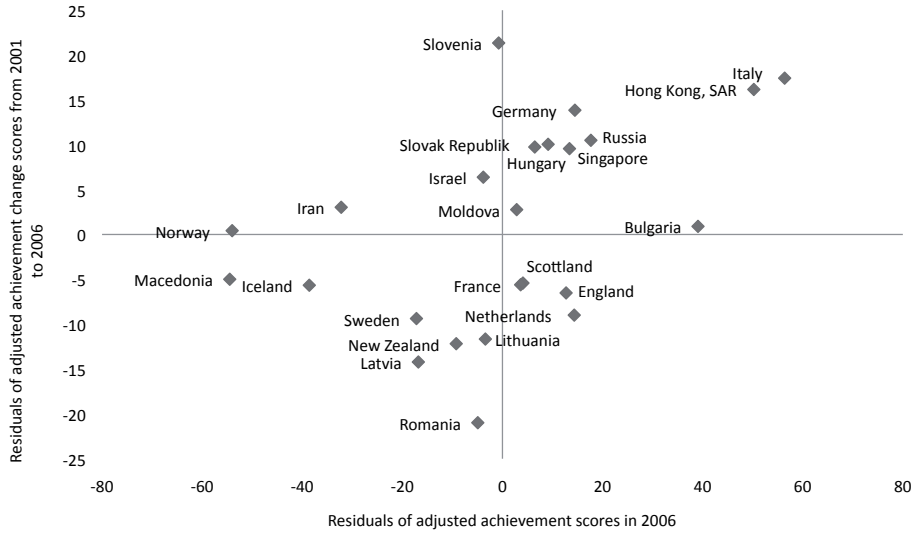
Equivalent to Figure 2, in Figure 4 effectiveness measures of educational systems have been ordered and contrasted with the ranks unadjusted achievement change scores of the respective systems. It can for example be seen that Russia, Hong Kong, SAR, and Singapore, the educational systems with the highest ranks for the unadjusted achievement change score obtain lower ranks for the effectiveness to change. Additionally New Zealand, Macedonia, and Iran would also be placed 9 and 5 (both Macedonia and Iran) positions lower when evaluated by their effectiveness. In contrast, England, Bulgaria, and Germany attain higher ranks if ordered by their effectiveness to change, with rank differences of 5 (England), 6 (Bulgaria), and 8 (Germany) positions. Overall, 7 of the 12 lower ranking educational systems would be assigned to higher ranks and 7 of the 12 higher ranking educational systems would be assigned to lower ranks.

Figure 4: Differences in ranks of unadjusted achievement change scores and residuals of adjusted achievement change scores (i.e. effectiveness measures) by educational system



Additionally, Figure 5 illustrates the correlation of effectiveness measures in 2006 and effectiveness measures for change from 2001 to 2006 to investigate if educational systems are equally effective for their average achievement in 2006 and their change score. The correlation is moderate but significant (r = .451). The upper right corner contains educational systems that have successfully managed to enhance their average performance from 2001 to 2006 and exceed their expected performance in 2006. Specifically, Italy and Hong Kong, SAR stand out. It is reasonable to assume that effectiveness in 2006 is at least partially a consequence of effective change from 2001 to 2006. Systems located in the upper left corner may have effectively improved their average achievement in the course of five years, however, this improvement has been insufficient (Iran) or just sufficient enough (Israel) to achieve their expected performance. In this group, Slovenia's educational system stands out with the highest change score and it is now near the average performance that could have been expected. Seven systems are positioned in the lower left corner which indicates ineffective performance in 2006 as well as ineffectiveness regarding change scores. Here, e.g. located is Romania's educational system that has been less successful in enhancing the average achievement score from the first assessment in 2001 to the second in 2006 and has fallen behind of what average performance could have been expected. In the lower right corner we find educational systems that are ineffective regarding their change score, but overall still effective with regard to their average performance in 2006.

Figure 5: Correlation of residuals of adjusted achievement scores in 2006 and residuals of adjusted achievement change scores from 2001 to 2006



8. Discussion

The paper demonstrates a methodological approach that can broaden the way results of international LSAs are reported. The approach moves beyond the comparison of unadjusted achievement scores by taking the effect of economic and developmental differences between educational systems into account. It further introduces a longitudinal perspective to study the effectiveness of educational systems based on their change in performance over time. The approach helps distinguishing high and low achieving systems from effective and ineffective systems.

The results have shown that educational systems can be categorized differently depending on the applied criterion: international standards or expected outcomes. Both are valuable information for policymakers, firstly to position oneself internationally and secondly to estimate the effectiveness of educational systems. Identifying effective systems presents the basis for further investigations about the structures and processes that favor effectiveness. For example, Sweden has one of the highest observed scores in 2006 (548 points). But once economic and developmental status is taken into account, it stays behind the performance that would have been expected. Hence, it seems questionable whether other educational systems should consider Sweden as an example for designing educational reforms. Slovenia, in contrast, has lower observed performance but may function as an example for many Eastern-European educational systems. Its effectiveness to change its average performance between 2001 and 2006 may provide a case for investi-

gating the characteristics, structures and reform measures of Slovenia's educational system.

The applied approach has however limitations. So far, it can only be the basis for further in-depth investigations into effectiveness enhancing factors. Certainly complementary information is needed to understand the reasons behind effectiveness. The analysis of process variables at the educational system level would contribute in this direction, but most international achievement studies still lack this information (Reynolds, 2006; Scheerens, 2006). Likewise, information and analysis of implemented educational reforms is important to understand their impact on the average performance in a longitudinal perspective.

There are further limitations of the paper itself that should not be neglected. With the PIRLS data progress in average reading achievement could only be modeled over two measurement points. But with more measurement points the model described by Willms and Raudenbush (1989) could include an additional level for the cohort units and provide more reliable results for the anticipated change measures.

Analysis is also limited by the measurement and validity of the included constructs. In general international LSAs of academic achievement are restricted by the cultural biases in cognitive assessments and their results (Solano-Flores & Nelson-Barber, 2001). Measurement and validity are furthermore an issue for the SES construct as it has been operationalized in this study. As Chudgar, Luschei, Fagioli, and Lee (2012) have shown, a different choice of constitutive variables would alter the association of the SES construct with achievement. The inaccurate measurement of SES could thus lead to biased estimates. Also, the comparability of the SES construct is limited by the different structures of social stratification across educational systems (Buchmann, 2002) and the fact that constitutive items are not equally indicative of SES across educational systems (Caro & Cortés, 2012). This limitation of comparability was accepted over the possibility to analyze cross-national data at all. Possible improvements to the SES index have been discussed, e.g. by May (2002) and should be taken into account in future analysis. Caro, Sandoval-Hernández, and Lüdtke (2012) have shown, though, that measurement invariance for the SES construct could not be supported for their sample of participating educational systems in PISA 2009 and PIRLS 2006. In fact even support of weak invariance for combinations of two educational systems was scarce. Measurement invariance for an index of socioeconomic status thus remains a major challenge for studies concerned with the analysis of cross-national data.

Shin and Raudenbush (2010) have, moreover, discussed the potential bias introduced by unreliable measures of compositional variables, such as the school mean of SES, which may occur when cluster sizes are insufficiently large in multilevel models. They propose a model that operationalizes the unit's mean on the covariate as a latent variable. In future analyses on effectiveness enhancing factors a more reliable latent compositional control variable may also yield more reliable associations of other higher level variables with the outcome variable.

Further, HDI as a macro level indicator of the economic and developmental status does not predict differences in average achievement or change in achievement. It is reasonable to assume, though, that HDI is a relevant predictor when a wider range of educational systems from more disadvantaged regions of the world are included in the analyses.

Another limitation is that the suggested models control for a restricted set of variables to evaluate effectiveness. Generally, effectiveness studies only control for variables that are associated with achievement *and* can be viewed as non-malleable by educators (Creemers & Kyriakides, 2008) such as socioeconomic and sociocultural background characteristics. In international LSAs the choice of control variables is restricted because the association of variables such as migration background of students with achievement is not stable across educational systems (Mullis et al., 2007; Mullis et al., 2008; OECD, 2010a). SES is the only factor that has been shown to be associated strongly with achievement across educational systems *and* is viewed as non-malleable in EER (ibid; Nachtigall et al., 2008; Raudenbush, 2004). Further theoretical and empirical investigations may yield a more complete selection of relevant predictors of achievement that can simultaneously be categorized as non-malleable across educational systems and at their different levels.

Although frequently used in the paper because of simplicity, it should be emphasized that the analyses yielded no evaluations of entire educational systems, but merely with regard to reading literacy at the end of fourth grade. Another academic domain would likely have produced very different results. Additionally, the concept of effectiveness as it has been operationalized here can only be measured against the included educational systems and positions are dependent on them respectively. Consequently, results are expected to change with the inclusion or exclusion of further educational systems.

Acknowledgement

The author is grateful to Daniel H. Caro and two anonymous reviewers for their helpful comments regarding this paper.

References

Alexander, K. L., Entwisle, D. R., & Olsen, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis*, *23* (2), 171–191.

Artelt, C. (2005). Cross-cultural approaches to measuring motivation. *Educational Assessment*, *10* (3), 231–255.

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.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal for Educational and Behavioral Statistics*, *29* (1), 37–65.

Bank, V., & Heidecke, B. (2009). Gegenwind für PISA. Ein systematisierender Überblick über kritische Schriften zur internationalen Vergleichsmessung [Head wind for PISA. A systemizing overview of critical papers on international comparative assessments]. *Vierteljahresschrift für Wissenschaftliche Pädagogik*, *85*, 361–372.

Bechger, T. M., van den Wittenboer, G., Hox, J. J., & De Glopper, C. (1999). The validity of comparative educational studies. *Educational Measurement: Issues and Practice*, *18* (3), 18–26.

Beese, J., & Liang, X. (2010). Do resources matter? PISA science achievement comparisons between students in the United States, Canada, and Finland. *Improving Schools*, *13* (3), 266–279.

Bempechat, J., Jimenez, N. V., & Boulay, B. A. (2002). Cultural-cognitive issues in academic achievement: New directions for cross-national research. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 117–149). Washington, DC: National Academy Press.

Bonsen, M., Bos, W., & Rolff, H.-G. (2008). Zur Fusion von Schuleffektivitäts- und Schulentwicklungsforschung [The fusion of school effectiveness and school improvement research]. In W. Bos, H. G. Holtappels, H. Pfeiffer, H.-G. Rolff, & R. Schulz-Zander (Eds.), *Jahrbuch der Schulentwicklung* (pp. 11–39). Weinheim: Juventa.

Breznitz, Z., & Teltsch, T. (1989). The effect of school entrance age on academic achievement and social-emotional adjustment of children: Follow-up study of fourth graders. *Psychology in the Schools*, *26*, 62–68.

Bryk, A. S., & Raudenbush, S. W. (2002). *Hierarchical linear models. Applications and data analysis methods*. London: Sage Publications.

Buchmann, C. (2002). Measuring family background in international studies of education: Conceptual issues and methodological challenges. In National Research Council (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 150–197). Washington, DC: National Academy Press.

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issues of nonequivalence. *International Journal of Testing, 10* (2), 107–132.

Campbell, R. J., Kyriakides, L., Muijs, R. D., & Robinson, W. (2003). Differential teacher effectiveness: Towards a model for research and teacher appraisal. *Oxford Review of Education*, *29* (3), 347–362.

Caro, D. H., & Cortés, D. (2012). Measuring family socioeconomic status: An illustration using data from PIRLS 2006. *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments, 5,* 9–33.

Caro, D. H., & Lehmann, R. (2009). Achievement inequalities in Hamburg schools: How do they change as students get older? *School Effectiveness and School Improvement*, *20* (4), 407–431.

Caro, D. H., & Lenkeit, J. (2012). An analytical approach to study educational inequalities: 10 hypothesis tests in PIRLS 2006. *International Journal of Research and Method in Education, 35* (1), 3–30.

Caro, D. H., Sandoval-Hernández, A., & Lüdtke, O. (2012, August). *An application of exploratory structural equation modeling to evaluate sociological theories in international large scale assessments*. Paper presented at the Sixth Biennial Meeting of EARLI SIG 1 (Assessment and Evaluation), Brussels, Belgium.

Chiu, M. M. (2007). Families, economies, cultures, and science achievement in 41 countries: Country-, school-, and student-level analyses. *Journal of Family Psychology, 21* (3), 510–519.

Chudgar, A., & Luschei, T. F. (2009). National income, income inequality, and the importance of schools: A hierarchical cross-national comparison. *American Educational Research Journal, 46* (3), 626–658.

Chudgar, A., Luschei, T. F., Fagioli, L. P., & Lee, C. (2012, April). *Socio-economic status (SES) measures using the Trends in International Mathematics and Science Study data*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.

Cliffordson, C., & Gustafsson, J.-E. (2007). Effects of age and schooling on intellectual performance: Estimates obtained from analysis of continuous variation in age and length of schooling. *Intelligence, 36* (2), 143–152.

Coe, R., & Fitz-Gibbon, C. T. (1998). School effectiveness research: Criticisms and recommendations. *Oxford Review of Education*, *24* (4), 420–438.

Cortina, K., Carlisle, J. F., & Zeng, J. (2008). Context effects on students' gains in reading comprehension in Reading First Schools in Michigan. *Zeitschrift für Erziehungswissenschaft*, *11* (1), 47–66.

Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness. A contribution to policy, practice and theory in contemporary schools*. London: Routledge.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistics Society. Series B: Statistical Methodology, 39* (1), 1–38.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12* (2), 121–138.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2* (3–4), 199–215.

Foshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievements of thirteen-year-olds in twelve countries: Results of an international research project, 1959–1961*. Hamburg: UNESCO Institute for Education.

Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement*, *8* (3), 305–322.

Goldschmidt, P., Choi, K., Martinez, F., & Novak, J. (2010). Using growth models to monitor school performance: Comparing the effect of the metric and the assessment. *School Effectiveness and School Improvement*, *21* (3), 337–357.

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

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments, 2*, 63–83.

HDRO – Human Development Report Office. (n.d.). *Indices & Data. Human Development Index. Human Development Reports (HDR). United Nations Development Programme (UNDP)*. Retrieved from http://hdr.undp.org/en/statistics/hdi/

Hecht, S. A., Burgess, S. R., Torgesen, J. K., Wagener, R. K., & Rashotte, C. A. (2000). Explaining social class differences in growth of reading skills from beginning kindergarten through fourth-grade: The role of phonological awareness, rate of access, and print knowledge. *Reading and Writing: An Interdisciplinary Journal*, *12*, 99–127.

Howie, S., & Hughes, C. (2000). South Africa. In D. F. Robitaille, A. E. Beaton, & T. Plomp (Eds.), *The impact of TIMSS on the teaching and learning of mathematics and science* (pp. 139–145). Vancouver: Pacific Educational Press.

Jaworski, B., & Phillips, D. (1999). Looking abroad: International comparison and the teaching of mathematics in Britain. In B. Jaworski & D. Phillips (Eds.), *Comparing standards internationally. Research and practice in mathematics and beyond* (pp. 7–22). Oxford: Symposium Books.

Jones, M. M., & Mandeville, G. K. (1990). The effect of age at school entry on reading achievement scores among South Carolina Students. *Remedial and Special Education*, *11* (2), 56–62.

Kelly, S., & Monczunski, L. (2007). Overcoming the volatility in school-level gain scores: A new approach to identifying value added with cross sectional data. *Educational Reseacher*, *36* (5), 279–287.

Kennedy, E., & Mandeville, G. (2000). Some methodological issues in school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp. 189–205). London: Routlegde.

Kobarg, M., & Prenzel, M. (2009). Stichwort: Mythos der nordischen Bildungssysteme [Keyword: The myth of the Nordic educational systems]. *Zeitschrift für Erziehungswissenschaft*, *12* (4), 597–615.

Leung, K., & van de Vijver, F. J. R. (2008). Strategies for strengthening causal inferences in cross cultural research: The consilience approach. *International Journal of Cross Cultural Management, 8* (2), 145–168.

Liegmann, A. B., & van Ackeren, I. (2012). The impact of PIRLS in 12 countries: A comparative summary. In K. Schwippert & J. Lenkeit (Eds.), *Progress in reading literacy in national and international context. The impact of PIRLS 2006 in 12 countries* (pp. 228–252). Münster: Waxmann.

Lind, G. (2004). Erfahrungen mit Standards in den USA – eine Übersicht [Experiences with standards in the USA – an overview]. *Journal für Schulentwicklung*, *8* (4), 55–60.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B: Statistical Methodology, 34*, 1–41.

Luyten, H. (1994). Stability of school effects in Dutch secondary education: The impact of variance across subjects and years. *International Journal of Educational Research*, *21* (2), 197–216.

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

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal for Educational and Behavioral Statistics*, *31* (1), 35–62.

May, H. (2002). *Development and evaluation of an internationally comparable scale of student socioeconomic status using survey data from TIMSS* (Doctoral Dissertation). Retrieved from ProQuest. Paper AAI3073031. Retrieved from http://repository.upenn.edu/do/search/?q=author_lname%3A%22May%22%20author_fname%3A%22Henry%22&start=0&context=19929

Mintrop, H., & Trujillo, T. (2007). The practical relevance of accountability systems for school improvement: A descriptive analysis of California schools. *Educational Evaluation and Policy Analysis*, *29* (4), 319–352.

Mislevy, R. J. (1995). What can we learn from international assessments? *Educational Evaluation and Policy Analysis*, *17* (4), 419–437.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary school in 40 countries*. Chestnut Hill, MA: Boston College.

Nachtigall, C., Kröhne, U., Enders, U., & Steyer, R. (2008). Causal effects and fair comparison: Considering the influence of context variables on student competencies. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 315–335). Göttingen: Hogrefe & Huber.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modelling of teacher effectiveness: An exploration of stability across models and context. *Education Policy Analysis Archives, 18* (23), 1–27.

OECD – Organisation for Economic Co-operation and Development. (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools*. Paris: OECD.

OECD – Organisation for Economic Co-operation and Development. (2009). *PISA 2009: Assessment framework: Key competencies in reading, mathematics and science*. Paris: OECD.

OECD – Organisation for Economic Co-operation and Development. (2010a). *PISA 2009 results: Learning trends: Changes in student performance since 2000 (Volume V)*. Paris: OECD.

OECD – Organisation for Economic Co-operation and Development. (2010b). *TALIS 2008: Technical report*. Paris: OECD.

OECD – Organisation for Economic Co-operation and Development. (2010c). *PISA 2009 Results: Overcoming social background: Equity in learning opportunities and outcomes (Volume II)*. Paris: OECD.

Opdenakker, M.-C., & Van Damme, J. (2006). Teacher characteristics and teaching styles as effectiveness enhancing factors of classroom practice. *Teaching and Teacher Education*, *22*, 1–21.

Opdenakker, M.-C., & Van Damme, J. (2007). Do school context, student composition and school leadership affect school practice and outcomes in secondary education? *British Educational Research Journal*, *33* (2), 179–206.

Postlethwaite, T. N. (1999). Overview of issues in international achievement studies. In B. Jaworski & D. Phillips (Eds.), *Comparing standards internationally. Research and practice in mathematics and beyond* (pp. 23–60). Oxford: Symposium Books.

Postlethwaite, T. N., & Ross, K. N. (1992). *Effective schools in reading. Implications for educational planners*. Hamburg: International Association for the Evaluation of Educational Achievement.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29* (1), 121–129.

Reynolds, D. (2006). World class schools: Some methodological and substantive findings and implications of International School Effectiveness Research Project (ISERP). *Educational Research and Evaluation*, *12* (6), 535–560.

Ringarp, J., & Rothland, M. (2010). Is the grass always greener? The effect of the PISA results on the education debates in Sweden and Germany. *European Educational Research Journal*, *9* (3), 422–430.

Robitzsch, A. (2010). TIMSS 1995 und 2007: Trend der mathematischen Kompetenzen in Österreich [TIMSS 1995 and 2007: Trends of mathematic competences in Austria]. In B. Suchań, C. Wallner-Paschon, & C. Schreiner (Eds.), *TIMSS 2007. Österreichischer Expertenbericht* (pp. 56–63). Graz: Leykam.

Rubin, D. B. (1987). *Multiple imputation for non-response in surveys*. New York, NY: Wiley.

Rutkowski, D., Gonzalez, E. J., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39* (2), 142–151.

Sammons, P. (1996). Complexities in the judgement of school effectiveness. *Educational Research and Evaluation, 2* (2), 113–149.

Sammons, P. (2006). The contribution of international studies on educational effectiveness: Current and future directions. *Educational Research and Evaluation*, *12* (6), 583–593.

Schafer, J. L., & Olsen, M. K. (1998). *Multiple Imputation for multivariate missing-data problems: A data analyst's perspective*. Retrieved from: http://www.stat.psu.edu/~jls/reprints/schafer_olsen_1998_mbr.pdf

Scheerens, J. (1997). Conceptual models and theory-embedded principles on effective schooling. *School Effectiveness and School Improvement*, *8* (3), 269–310.

Scheerens, J. (2006). The case of evaluation and accountability provisions in education as an area for the development of policy malleable system level indicators. *Zeitschrift für Erziehungswissenschaft*, *9* (6), 207–224.

Schwippert, K., & Lenkeit, J. (2012a). Introduction. In K. Schwippert & J. Lenkeit (Eds.), *Progress in reading literacy in national and international context. The impact of PIRLS 2006 in 12 countries* (pp. 9–21). Münster: Waxmann.

Schwippert, K., & Lenkeit, J. (Eds.). (2012b). *Progress in reading literacy in national and international context. The impact of PIRLS 2006 in 12 countries*. Münster: Waxmann.

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics, 35* (1), 26–53.

Shorrocks-Taylor, D. (2010). International comparisons of student achievement: An introduction and discussion. In D. Shorrocks-Taylor & E. W. Jenkins (Eds.), *Learning from others* (pp. 13–27). Dordrecht: Kluwer Academic Publishers.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: a meta-analytic review of research. *Review of Educational Research*, *75* (3), 417–453.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38* (5), 553–573.

Stevens, J. (2005). The study of school effectiveness as a problem of research design. In R. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 166–208). Maple Grove: JAM Press.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education, 62* (4), 339–355.

Tan, J. B. Y., & Yates, S. M. (2007). A Rasch analysis of the Academic Self-Concept Questionnaire. *International Education Journal*, *8* (2), 470–484.

Teddlie, C., & Reynolds, D. (2000). *The International Handbook of School Effectiveness Research*. London: Routledge.

Teddlie, C., Reynolds, D., & Sammons, P. (2000). The methodology and scientific properties of school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *The International Handbook of School Effectiveness Research* (pp. 55–133). London: Routledge.

Thomas, S. (1998). Value-added measures of school effectiveness in the United Kingdom. *Prospects*, *28* (1), 91–108.

Thomas, S., & Mortimore, P. (1996). Comparison of value-added models for secondary-school effectiveness. *Research Papers in Education, 11* (1), 5–33.

van Ackeren, I. (2007). Comparative synthesis. In K. Schwippert (Ed.), *Progress in reading literacy. The impact of PIRLS 2001 in 13 countries* (pp. 243–264). Münster: Waxmann.

van Damme, J., Liu, H., Vanhee, L., & Putjens, H. (2010). Longitudinal studies at the country level as a new approach to educational effectiveness: Explaining change in reading achievement (PIRLS) by change in age, socio-economic status and class size. *Effective Education*, *2* (1), 53–84.

Waldow, F. (2010). Der Traum vom „skandinavisch schlau werden“. Drei Thesen zur Rolle Finnlands als Projektionsfläche in der gegenwärtigen Bildungsdebatte [The dream of “becoming clever the Scandinavian way”. Three propositions on the role of Finland as ‘projection screen’ in the present educational debate]. *Zeitschrift für Pädagogik*, *56* (4), 497–511.

Watermann, R., & Klieme, E. (2002). Reporting results of large-scale assessments in psychologically and educationally meaningful terms: Construct validation and proficiency scaling in TIMSS. *European Journal of Psychological Assessment*, *18* (3), 190–203.

Willet, J. B. (1988). Chapter 9: Questions and answers in the measurement of change. *Review of Research in Education*, *15*, 345–422.

Willms, D. J., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, *26* (3), 209–232.

Zvoch, K., & Stevens, J. J. (2008). Measuring and evaluating school performance. An investigation of status and growth-based achievement indicators. *Evaluation Review*, *32* (6), 569–595.