

Achievement status and growth as predictors of educational outcomes and effectiveness

Dissertation

zur Erlangung des akademischen Grades eines Doktors der
Philosophie (Dr. phil)

an der Fakultät für Erziehungswissenschaft, Psychologie und
Bewegungswissenschaft der Universität Hamburg

vorgelegt von

Jenny Lenkeit

bei Prof. Dr. Knut Schwippert

und Prof. Dr. Wilfried Bos

Hamburg, September 2012

Content		Page
CHAPTER 0	OVERVIEW	4
CHAPTER 1	INTRODUCTION	6
1.1	Assessment results as quality indicators	6
1.2	Quality in educational institutions	8
1.2.1	Quality of what?	9
1.2.2	Quality criteria	11
1.3	Educational standards, student socioeconomic status, and accountability	13
1.4	Educational Effectiveness Research	15
1.5	Achievement status and growth as indicators of educational effectiveness	18
1.5.1	Ability and accomplishment	18
1.5.2	Different measures – different quality?	20
1.6	Research questions	21
1.6.1	Achievement growth and school track recommendations	21
1.6.2	Comparability of effectiveness measures obtained from status and growth models	22
1.6.3	Effectiveness measures for educational systems based on level and change in performance	23
1.6.4	Reliability of effectiveness measures obtained from achievement growth	24
1.7	Significance	25
CHAPTER 2	STUDY 1: THE ROLE OF ACADEMIC ACHIEVEMENT GROWTH IN SCHOOL TRACK RECOMMENDATIONS	28
CHAPTER 3	STUDY 2: EFFECTIVENESS MEASURES FOR CROSS-SECTIONAL STUDIES: A COMPARISON OF VALUE-ADDED MODELS AND CONTEXTUALIZED ATTAINMENT MODELS	39
CHAPTER 4	STUDY 3: HOW EFFECTIVE ARE EDUCATIONAL SYSTEMS? A VALUE-ADDED APPROACH TO STUDY TRENDS IN PIRLS	65
CHAPTER 5	CONCLUSION AND DISCUSSION	102
5.1	Aims and objectives recapitulated	102
5.2	Summary and discussion of results	103

CHAPTER 6	SUMMARIES	111
5.1	English summary	111
5.2	German summary – Deutsche Zusammenfassung	115
REFERENCES	120
TABLE OF FIGURES	129
TABLE OF TABLES	131
ANNEX	132
A.1	Liste der Einzelarbeiten	132
A.2	Curriculum Vitae	133
A.3	Liste der Publikationen und Präsentationen	135

CHAPTER 0 OVERVIEW

Educational researchers increasingly use measures of achievement growth as quality indicators. However, neither is there consensus about the nature of quality that is captured by growth measures, nor about the advantages of the methodological approaches applied to obtain them.

Theoretically, levels of academic achievement and growth of achievement predict educational outcomes differently. Growth indicates the potential and capacity to acquire knowledge and skills (Guo, 1998) and reflects the fact that learning itself is a cumulative process (Willet, 1988). Achievement levels rather capture ability, as well as characteristics of family background that influence students' academic performance (Haveman & Wolfe, 1995). Methodologically, growth measures are believed to be unconfounded with family background characteristics, while achievement levels are usually adjusted for their influence. Both adjusted achievement levels and growth measures can be described as indicators of "contextualized quality" because they take students' characteristics into account when evaluating quality of teachers and schools. In educational research this notion of quality is also entitled *effectiveness* (Creemers & Kyriakides, 2008; Scheerens & Creemers, 1989). Still, in Educational Effectiveness Research (EER) achievement growth is treated conceptually different from adjusted achievement levels because it reflects the cumulative process of learning and therewith a different notion of educational quality. Also, the methodological approaches to arrive at effectiveness measures of achievement growth and status diverge.

In contexts of educational policies and decision making processes choices have to be made regarding the notion of quality associated with effectiveness measures of achievement status and achievement growth. The research literature has however neglected to investigate whether the theoretical distinction of achievement status and growth is of practical relevance for the evaluation of educational quality. Above the conceptual distinction, statistical properties of the respective measures are of decisive importance for these choices. Again, the research literature and policymakers have paid insufficient attention to the relevance of these statistical properties. Both, decisions about the notion of quality as well as statistical properties of measures that are supposed to indicate quality, are of growing importance in contexts where the quality of educational institutions is increasingly judged on the basis of these measures. The thesis investigates the practical consequences and statistical properties of effectiveness measures for the evaluation of educational outputs. It also contrasts theoretical and methodological arguments for the use of effectiveness measures obtained for achievement

status and achievement growth and argues how the discrepancies between those arguments can be resolved. It thereby also deals with questions of quality criteria and accountability.

The introductory part of the thesis (chapter 1) frames the importance of conceptual and methodological adequateness of quality indicators in educational contexts. It argues that the shift in steering mechanisms within and between educational systems towards accountability oriented evaluations of educational institutions has established a new relevance of educational outputs as quality indicators. It first describes how results of standardized assessments are increasingly utilized to evaluate the quality of educational institutions and educators in national and international contexts. The second part of chapter 1 deals with the definition of quality and quality criteria against which educational outputs are measured. Here, the important differentiation between quality as the accomplishment of a fixed set of standards and “contextualized” quality is reviewed. Therewith connected, the concept of educational effectiveness is introduced (Scheerens & Creemers, 1989; Teddlie & Reynolds, 2000).

The third part of chapter 1 elaborates on the restriction of standards as the basis of quality evaluations in educational contexts. It discusses the relativity of educational quality due to wide differences in students’ socioeconomic background that educators are confronted with. This discussion is further tangent to questions of accountability. The fourth part of chapter 1 briefly reviews developments in EER. It elaborates on the development and particularities of the field such as the notion of student intake inherited in different models of effectiveness. Part five focusses on the distinction between effectiveness measures obtained from achievement growth and achievement status models, outlining conceptual and methodological arguments. Further, arguments of different effectiveness measures as quality indicators are discussed. Part six of chapter 1 summarizes the outlined research questions derived from the elaborated background and emphasizes the significance of the conducted research for the educational research field.

Chapters 2 to 4 consist of three individual articles that address the research questions in particular. Chapter 5 closes the thesis by recapitulating its overall objectives and discussing the combined results of the three articles in relation to the posed research questions. Chapter 6 provides English and German summaries of the thesis.

CHAPTER 1 INTRODUCTION

1.1 Assessment results as quality indicators

Recent decades have seen a mentionable increase in assessments of educational outcomes in various school stages and domains around the world. In 1959 the International Association for the Evaluation of Educational Achievement (IEA) started its first international comparative study (The Pilot Twelve Country Study) with only 12 participating countries (Foshay, Thorndike, Hotyat, Pidgeon, & Walker, 1962). Since then several new studies have emerged, conducted not only by the IEA (Trends in International Mathematics and Science Study [TIMSS], Program in International Reading Literacy Study [PIRLS]) but other international organizations such as the Organization for Economic Cooperation and Development (OECD) (Programme for International Student Assessment [PISA]). The role of such international large scale assessments (LSA) and their perceived importance in the global educational market is moreover reflected in the number of participating educational systems. For example, 63 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. Various domains (e.g. civic education, classroom environment), academic disciplines (e.g. reading and mathematics literacy, computer literacy) as well as age cohorts (e.g. 15 year old students, fourth graders) have been the objects of these studies. A recent overview of studies and participating countries is found in Schwippert and Goy (2008) and Schwippert and Lenkeit (2012).

Next to international LSA systematic national assessments are an integrated part in many educational systems such as of the USA and England in which accountability oriented evaluations of schools are established. In the USA, for example, programs such as the National Assessment of Educational Progress (NAEP) and the more recent No Child Left Behind Act (NCLB) systematically assess students' achievement levels and learning gains in order to ensure national educational standards. Systematic national assessments are, however, a relatively new aspect of Germany's educational system. Changes in steering mechanisms were only vigorously initiated after the results of TIMSS in 1995 (Beaton et al., 1996) suggested that the exhibited academic outputs of students lacked behind expectations. This led to doubts that solely input oriented steering mechanisms could provide the anticipated outcomes. An orientation to evaluation based on academic outputs set in, resulting in a major increase in national assessments of student performances throughout grades and academic subjects (e.g. Lehmann & Peek, 1997; Lehmann et al., 2001; Bos, Bonsen, Gröhlich, Jelden,

& Rau, 2006; Lehmann & Nikolova, 2005) that inevitably culminated in the development of national educational standards (Bundesministerium für Bildung und Forschung (BMBF), 2007).

The increase of international and national educational assessments can only partly be ascribed to a scientific interest in the functioning and structures of educational institutions and systems, but is rather linked with a strong shift of steering mechanisms towards performance based evaluations (Küssau & Brüsemeister, 2007). This shift reflects an increased interest in the effects investments in educational systems and structures actually have on educational outcomes. The assessments thereby provide information for policymakers and administrators and, if performed on a regular basis, are believed to promote school improvement (Willms, 2000; Decker & Bolt, 2008) and system improvement, respectively.

Nevertheless, the increase in performance based evaluations has evoked critical discussions. Many of these criticisms aim at the assessments' focus on academic outputs. They therewith seem to define what is valued and appreciated as the aim of institutionalized education by using the assessed academic outputs as indicators of educational quality. Meanwhile other aims of education such as equity (Choi & Seltzer, 2003; Sammons, 2007), the development of meta-cognitive skills, and non-cognitive outcomes (Campbell, Kyriakides, Muijs, & Robinson, 2003) are rigorously neglected. Accordingly, Burbules (2004) points out, that the production of academic outcomes is by no means the only aim of education. Sociocultural reproduction as well as social and cultural stability and development are an important aspect of education but seem to fade into the background, presumably because of the complexity to measure those aims of education (ibid.). Consequently, what is measurable in standardized tests and through educational standards seems to define what counts as worthwhile knowledge (ibid.). Scheerens (2004), too, points out different perspectives on educational quality (e.g. equity, efficiency) and acknowledges that international LSA take on the productivity perspective that focuses on academic outputs. The productivity perspective evaluates the success of educational systems by their attainment on aspired outcomes such as the quantity of school leavers or academic achievement (ibid.). Decker and Bolt (2008) further criticize the use of standardized assessments to evaluate educational quality. They emphasize that standardized assessments cannot provide a comprehensive picture of students' competences in various academic domains and content areas. Rather they assess merely a sample of skills and knowledge in some areas (ibid.).

These criticisms can be summarized to one overarching aspect that is: The use of standardized tests and the interpretation of their results often outrun their capacity to make valid statements about the overall educational quality.

1.2 Quality in educational institutions

When discussing educational assessments and quality, a clarification of what exactly is assessed (quality of what?) and in reference to what it is evaluated (quality criterion) is necessary.

There is no widely acknowledged definition of the term quality. Against many believes that quality is an observable characteristic of an object itself, Heid (2000) points out, that quality is rather the result of an evaluation of an object's nature. This evaluation itself depends on explicit and implicit decisions about the criteria against which to evaluate an object's nature. Those decisions are made by those who claim to ensure and establish quality (ibid.). In that sense quality is described as the discrepancy of what is desired as an outcome or a characteristic by relevant actors and what can respectively be observed. In other words, an object can only be designated "good" or "bad" in reference to a normative evaluation criterion (ibid.). Against this background, quality has to be viewed as a relative parameter, which is subject to social constructions and legitimations (Kuper, 2002). Importantly, the questions arises who constructs and who legitimizes or rather whose interests are reflected in the determination of evaluation criteria?

To exemplify this, the complexity of the definition of a criterion for equity as an educational output shall be outlined. An emphasis on equity as an indicator of educational quality distinguishes whether gaps in different domains are reduced or increased (Choi & Seltzer, 2003; Creemers & Kyriakides, 2008). Nevertheless, an examination of equity in education also needs a critical view on the concept of fairness. As discussed in Schwippert and Walker (2003), educators will have to decide how they allocate limited amounts of time and attention to students of different ability and hence provide educational opportunities. Thereby they have distinct motives, beliefs, and virtues that form their allocation strategies (Heckhausen, 1981; Schwippert & Walker, 2003): the need principle; the justness principle; the equality principle. The preference of one of these principles will then specify the respective quality criterion.

Harvey and Green (2000), too, point out that quality is always relative because different actors would evaluate an object in reference to different criteria. As such students, parents,

teachers, administrators, researchers, and politicians will apply different standards for their evaluation of a specific aspect of institutionalized education. Further, the idea of evaluating educational quality is much too broad and vague. Firstly, both Fend (2008) and Kuper (2002) distinguish different levels on which quality in education can be evaluated. Objects of evaluation may be allocated to the level of the educational system, the school, the classroom, as well as to the level of the student and his or her parents. And secondly, the vast amount of facets (e.g. academic outcomes, equity, development of social skills) and components (e.g. instruction, school climate) in education prohibits a single evaluation criterion for educational quality. Thus, referring to educational quality entitles no consensus whatsoever (Heid, 2000), neither with regard to the standard applied, nor the level referred to or the specific aspects valued within the system of institutionalized education. Every critique of quality consequently requires a specification of what exactly is evaluated and based on which criteria.

1.2.1 Quality of what?

In educational research quality can refer to two broader components which are mutually dependent. Figure 1 presents the theoretical framework of the relationship between student achievement and its determinants as it has been elaborated in the PIRLS Germany reports (Bos et al., 2007) based on the work of Wang, Haertel, and Walberg (1993). Note, though, that the presented framework is not a model of educational quality itself.¹ Rather, each of the depicted components and their constituting factors are subject to quality evaluation and will be further discussed.

The first component comprises educational processes and structures (Holtappels, 2003). These are captioned with “within school conditions”, “teachers”, “classroom environment and structure”, “instruction”, and “educational policies and school system factors” in Figure 1. Especially research focused on teacher practices and instruction methods is concerned with the quality of teaching and learning processes usually in relation to a defined output (Campbell et al., 2003; Opdenakker & Van Damme, 2006). Likewise the school principle and his or her organization of the institution is gaining increased attention, strengthening the importance of the quality of leadership within schools (Rice, 2010). Further, educational policies and school system factors (e.g. tracking, retention policies, educational spending) are particularly accentuated as components of structural quality in cross national comparisons.

¹ The model has nevertheless been chosen here, because generally quality models do not take on a holistic approach of quality but focus for example on quality of classrooms and instruction (Helmke, 2007) or supply and use models (Fend, 2000).

The second component refers to the impact that structures and processes have on the educational output. The definition of this output is to a large part a conviction of what is to be considered the aim of schooling. Here, the distinction between academic and non-academic outputs is primary for the definition of the educational output. Most assessments relate to cognitive domains of schooling (De Maeyer, van den Bergh, Rymeanans, Van Petegem, & Rijlaarsdam, 2010; Postlethwaite & Ross, 1992), arguing that they best represent the school's societal assignment and the areas in which schools can make a recognizable difference (Opdenakker & Van Damme, 2000). This is also reflected in the fact that the influence of classes and schools tends to be higher on cognitive domains than on non-cognitive domains (ibid.).

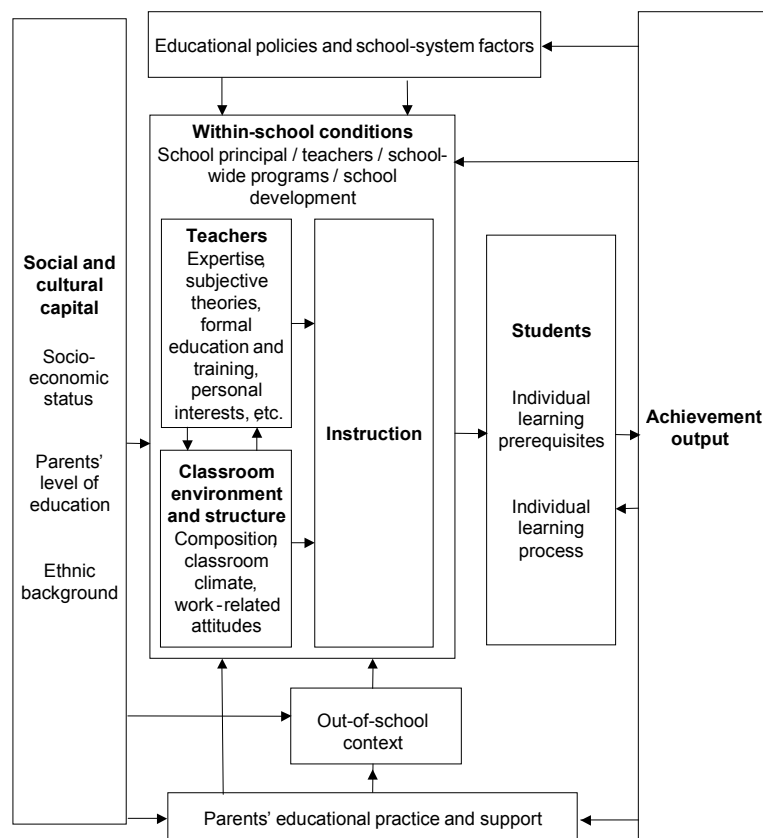


Figure 1. Theoretical framework of the relationship between student achievement and its determinants

Source: Adapted from Bos et al. (2007)

Others propose to loosen this focus and include outcomes such as equity (Choi & Seltzer, 2003; Sammons, 2007), meta-cognitive skills and non-cognitive outcomes in evaluations as

indicators of the quality of a school's comprehensive work (Campbell et al., 2003). Nevertheless, most national and international LSA assess academic outputs such as reading, mathematics, and science and consequently make statements about the value of these specific facets of education.

The third component is not evaluated in the research literature as a part of institutional quality. It comprises “social and cultural capital”, “parents’ educational practice and support” as well as “out-of-school context”. Rather, these characteristics are constituents of educational processes and outcomes within this framework. In the research literature they represent what is called the “student intake”. The student intake can be described as the compositional characteristics of the student body and represents the non-malleable part of what constitutes the “input” in common models of steering mechanisms (Fend, 2008; Köller, 2009). It follows that these factors should be considered part of assessments that evaluate aspects of educational quality.

1.2.2 Quality criteria

Derived from the different broad components for which quality can be defined (structures and processes, outputs), different quality criteria can be established. *Structures and processes* are usually measured against what is known from educational research to be conditional factors that determine outcomes of schooling. These include structural factors such as achievement requirements, transitional structures, curriculum, opportunities to learn as well as patterns and processes such as instructional profiles, interactions between students, and students and teachers (Creemers & Kyriakides, 2008; Holtappels, 2003). Quality is then measured by whether institutions exhibit certain characteristics of these factors.

However, a clearer definition of the evaluation criterion is essential for assessments focused on *outputs*, independent of whether they are cognitive or non-cognitive outputs. Here a specific reference against which the achieved output is measured is necessary. In alignment with the scope of the thesis, this section deals with criteria applied to evaluate academic outputs.

Generally, criteria against which to measure quality of academic outputs are formulated in terms of standards. A standard comprises for example certain output indicators such as the quantity of school leavers on a specific level (Scheerens, 2004) or competences students should have acquired in a certain academic domain and at a certain stage of their educational careers (BMBF, 2007). Scheerens (2004) distinguishes the following types of standards: relative, comparative standards and absolute standards. Relative, comparative standards

provide reference points and are intended to initiate learning through comparison (Scheerens, 2004, cf. EU, 2000, 3). Absolute standards are “content based standards, defined on elaborated scales of student achievement, established by means of expert consensus” (ibid., p.130). In this sense, national educational standards represent absolute standards. More generally, in Germany as well as in other countries national educational standards describe educational aims that function as an orientation for schools and educators and are the basis to assess and evaluate the results of institutionalized education (BMBF, 2007). More specifically, “the constitutive aims of the pedagogical work [own translation]“ (BMBF, 2007, p. 9) are formulated as competencies students should have acquired in certain grades and academic subjects. The standards thus determine which competences students must have acquired. Whether or not these standards are met is assessed with standardized assessments that are linked with curricular regulations. The requirements are formulated in competence models that allow for graduation and distinguish developments in academic attainment (ibid.). Therewith they aim at investigating and establishing quality within the educational system. Educational standards are the criterion on which the evaluative interpretation about quality is ultimately made (Scheerens, 2004).

International LSA of student achievement face restrictions towards the formulation of common standards. Because educational systems attribute different functions and expected academic outputs to institutionalized education (Kuper, 2000), it seems difficult to determine standards that capture the exact aims and objectives of different educational systems and are therewith capable evaluating quality across educational systems. This is specifically reflected in the extent to which formulated competence requirements and specific tasks cover the national specific curricula (Goldstein, 2004; Klieme & Baumert, 2001). Although, IEA’s studies generally claim to be oriented towards national curricula because items are developed and legitimated by representatives of the participating systems (e.g. Martin, Mullis, & Kennedy, 2007), these items rather reflect the lowest common denominator of a multitude of curricula, which are in fact predominantly of western origin. Nevertheless, international LSA apply criteria against which the success of educational systems is evaluated. Similar to national educational standards experts determine characteristics of skills at different levels and therewith classify students’ exhibited competencies in proficiency levels (Mullis, Martin, Kennedy, & Foy, 2007; OECD, 2009; Olsen, 2005). They thus formulate competences and respective competence levels. The frequently found term benchmark essentially represents the standard (Scheerens, 2004). Benchmarks “represent the range of performance shown by students internationally” (OECD, 2009, p. 67) and, as Scheerens (2004) further points out,

their “connotation is an “outward looking” approach where it is assumed that there are external standards to which one can compare oneself” (p. 129). In that sense standards in international LSA are both absolute and relative.

1.3 Educational standards, student socioeconomic status, and accountability

Many theoretical and empirical works have emphasized the limits of educational standards and their assessment through standardized tests for the purpose of educational quality evaluation. Essentially, the limits concern conceptual (see critical review in section 1.1) and methodological aspects such as the assumption that the complexity of what constitutes educational quality can be represented in a single statistical score (Ballou, 2002; Burbules, 2004; Kelly & Monczunski, 2007). Other critical arguments are issued in reference to the accountability imposed on educators and educational institutions based on results of standardized assessments. These critiques are of particular significance in high stakes educational systems, where educators and schools are faced with severe personnel decisions based on their assessment results (Kohn, 2000). In particular, critique evolves around the fact that educators can be differentially effective for different academic domains and that a school’s work cannot rightfully be judged by assessing its quality in one domain only (Hill & Rowe, 1996; Sammons, Nuttall, & Cuttance, 1993).

Another more pertinent critique refers to the fact that schools educate and teachers instruct students with different family backgrounds. Figure 1 (section 1.2.1) illustrated that there are factors outside the school influencing the educational outputs and therewith the accomplished quality. “Social and cultural capital”, “out-of-school context” as well as “parents’ educational practice and support” directly influence the processes that take place within schools and classrooms (Opdenakker & Van Damme, 2007; Stevens, 2005) and through them influence the educational output. Further, they are likely to direct school policy and practices (Raudenbush, 2004; Raudenbush & Willms, 1995). This component of the model is based on theories and empirical findings that demonstrate the relationship of students’ socioeconomic status (SES) and educational attainment.

Theoretically, SES is defined as the relative position of an individual or a family within a hierarchical social structure. This position is based on their access to or control over wealth, prestige and power (Mueller & Parcel, 1981). Bourdieu and Passeron’s (1977) theory² provides a useful framework to explain the formation of SES and its reproduction through

² Theories by Bernstein (1975) and Colman (1988) expand Bourdieu and Passeron’s (1977) framework, but are not included in this short outline of the educational reproduction theory.

forms of institutionalized education. Educational performance of students depends on the amount and composition of different capital forms and on the extent to which amount and composition serve the symbolic requisites of the dominant culture which is legitimated within the education system (Bourdieu & Passeron, 1977). The capital forms are described as economic capital, cultural capital, and social capital (Bourdieu, 1983). Economic capital can be defined as the command students have over economic resources (ibid.). It comprises e.g. income and assets that can be transformed in money (so called exchange values). According to Bourdieu (1983) cultural capital is represented by continuous dispositions embedded in the human mind and body, cultural goods and educational certifications. Bourdieu (ibid.) further describes social capital as the lasting network of institutionalized relationships (family relationships as well as formalized clubs). The members of the network profit from the capital owned by the group of people belonging to the network. Therewith relationships become beneficial as they secure material and symbolic profits.

Many national (e.g. Bos et al., 2008; Caro, McDonald, & Willms, 2009; Condrón, 2007; Lehmann & Lenkeit, 2008) and international (Mullis, Martin, Kennedy, & Foy, 2007; Mullis et al., 2008; OECD, 2010) empirical studies have provided evidence for the relationship of SES and educational achievement. They show that students from educationally and socioeconomically disadvantaged families have systematically lower educational achievements than students from educationally and socioeconomically affluent families (Alexander, Entwisle, & Olsen, 2001; Bos, Stubbe, & Buddeberg, 2010; Schnabel, Alfeld, Eccles, Köller, & Baumert, 2002; Tramonte & Willms, 2010).

Most researchers therefore agree that contextual conditions (as represented by student SES) of educational processes and outcomes are factors that teachers and schools cannot be held responsible for (Ballou, Sanders, & Wright, 2004; Martineau, 2006). These factors have to be taken into account when evaluating the quality of educators and educational institutions.

Nevertheless, absolute standards are increasingly used not only to judge the quality of entire educational systems, schools, principals, and teachers but to hold the respective actors responsible for the evaluated quality. Here, a particular question arises: Who can be responsible for what? To answer this question evidence from educational research should be taken into account. Maaz, Baumert, and Trautwein (2009) and Baumert, Stanat, and Watermann (2006) differentiate between three effects that affect students' outputs in academic achievement tests – individual effects, compositional effects and institutional effects (see also e.g. Bryk & Raudenbush, 2002). The first effect describes the specific influence of students' attributes such as motivation and self-concept as well as family background

characteristics such as socioeconomic background on academic achievement. Many empirical studies have shown that these characteristics directly affect a student's academic achievement (e.g. Sirin, 2005; Willms, 2003). Compositional effects describe the influences that originate from the composition of these characteristics within a school or class that influence students' academic achievement over and above individual effects (e.g. Bryk & Raudenbush, 2002). Students in classes with a higher advantaged background have on average higher academic achievement scores (Lehmann, 2006; Maaz et al., 2009). Institutional effects are those that a class and/or a school have on the academic achievement of students after controlling for both individual and compositional effects (Raudenbush, 1995, 2004). Therefore, independent of the professional skills of a teacher or a school principle, his or her students will perform worse when they come from disadvantaged family backgrounds and are grouped in learning environments that cumulate these disadvantages. To hold educators responsible for educational outcomes that are directly and indirectly associated with these disadvantaging contexts would be unreasonable and unfair. Taking this complexity of different effects into account, educators can therefore only be held responsible for what is defined as the institutional effect (Maaz et al., 2009), that is, for what is in their sphere of influence (Ballou, et al., 2004; Martineau, 2006, OECD, 2008).

These elaborations demonstrate that even within a set framework of the evaluation object and the respective evaluation criteria (as they are formulated by standards) a common understanding about the evaluation of the involved educators' work is far from simple. Educational Effectiveness Research (EER) offers a conceptual framework and methodological approaches to address these complexities.

1.4 Educational Effectiveness Research

The EER field has established a notion of quality that considers the contextual conditions in which teachers and schools operate. This notion of quality is based on evidence that the student intake influences the academic outcome of an educational unit, that is, usually the school or the class (e.g. Sirin, 2005). The student intake is for example reflected by socioeconomic and cognitive characteristics of students. The notion of quality is then based on the belief that the student intake is non-malleable by educators and that they should not be held responsible for what is not amenable to education policy (Ballou et al., 2004; Martineau, 2006; OECD, 2008; Thomas, 1998). With this notion of quality, EER essentially aims at isolating the institutional effect on student academic outcomes from influences that originate from individual and compositional background characteristics.

The EER field is thus primarily concerned with the effects classroom and school practices, processes and policies have on student achievement (Creemers & Kyriakides, 2008; Postlethwaite & Ross, 1992; Reynolds, Teddlie, & Townsend, 2000; Scheerens, 1997). What today is called EER captures a range of research areas from different waves and strands (Creemers & Kyriakides, 2008; Reynolds et al., 2000). It represents an integration of the fields of school effectiveness (school organization and educational policy) (e.g. Teddlie & Reynolds, 2000) and research aimed at the classroom level (teacher behavior, instruction methods, and curriculum analyses) (e.g. Campbell et al., 2003; Stronge, Ward, & Grant, 2011). With a proceeding awareness and empirical evidence of contextual impacts on learning processes, approaches were elaborated that viewed effectiveness as a multilevel phenomenon integrating cross-level relationships in the theoretical models. This development promoted the blending of the former approaches (Creemers & Kyriakides, 2008; Scheerens, 1997) to what has commonly been called educational effectiveness. It has moreover yielded in the dynamic model of educational effectiveness as elaborated by Creemers and Kyriakides (2008) (see Figure 2).

Research undergone by many educational researchers within the field of educational effectiveness is often concerned with practices and factors that enhance the impact of schools and teachers on the educational outcome. Creemers' and Kyriakides' model (2008) is a consequence and evidence of the amount of research concerned with the topic. As can be seen in Figure 2 it also includes factors such as socioeconomic status (SES), ethnicity, and personal traits that should be outside the model from an accountability perspective, in terms of controlling for them. From the theoretical perspective of the model, however, the authors integrated those factors because they mediate characteristics of teaching and school policy (Creemers and Kyriakides 2008; Opdenakker & Van Damme, 2007; Stevens, 2005). It is consequently assumed that teaching and school policy characteristics are differentially effective within certain contextual conditions (Kyriakides, 2004; Strand, 2010). From a more practical perspective this knowledge is relevant for investigations concerned within the area of "school development" and "best practice schools" (Bonsen, Bos, & Rolff, 2008; Mintrop & Trujillo, 2007).

In terms of accountability, the actual effectiveness enhancing factors within schools are, however, only of secondary importance. Temporarily prepended is the identification of effective and less effective teachers and/or schools. Only then does a closer look at school structures, processes and practices with regard to effectiveness make sense. In order to identify those teachers and schools a definition of effectiveness is required.

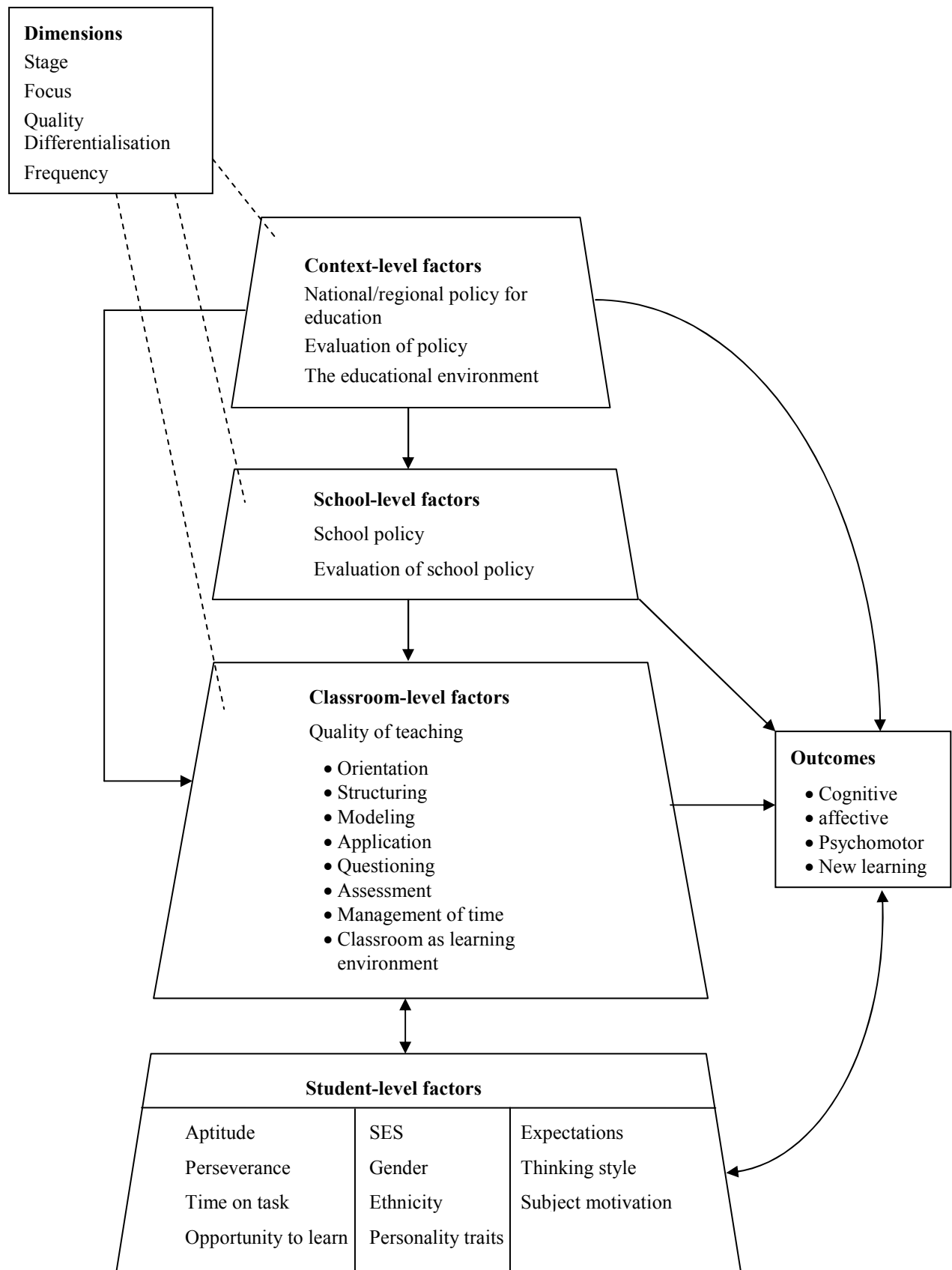


Figure 2. The dynamic model of educational effectiveness (Creemers & Kyriakides, 2008, p. 150)

In educational research emphasizing cognitive outcomes, effectiveness is often determined as a measure of achievement that is not predicted by other characteristics such as prior attainment or diverse student background variables (the so called student intake) (Hill & Rowe, 1996). Effectiveness can be defined as the relation of the observed and the expected outcome (Scheerens & Creemers, 1989; Teddlie & Reynolds, 2000). What can be expected is a function of the student intake that influences academic achievement, but is non-malleable by educators. In other words, effectiveness measures comprise institutional effects that are purged from the effects of non-malleable individual and compositional background characteristics. Consequently, what is expected functions as a standard, i.e. evaluation criterion. Therewith, EER's evaluation criteria and thus its notion of quality are neither absolute nor strictly speaking comparative in the sense that they initiate learning through comparison (Scheerens, 2004). Rather, it can be described as a "contextualized quality" because the object is not evaluated against an absolute standard but rather in reference to its own (non-malleable) contextual characteristics.

At this point, however, researchers disagree about the nature of student intake they control for and the therewith connected notion of effectiveness, i.e. quality. In cross-sectional data designs student intake is controlled for by taking students' background and compositional characteristics into account to yield adjusted achievement measures (OECD, 2008). In longitudinal data designs, it is acknowledged, that achievement gains are independent of achievement levels and therewith supposedly independent of student intake (Andrejko, 2004; Zvoch & Stevens, 2008). Hence, in longitudinal data designs prior achievement measures serve as controls for student intake (Ballou et al., 2004; Thomas & Mortimore, 1996). However, these concerns do not only pertain to methodological aspects (cross-sectional or longitudinal data designs). They are moreover connected to the distinction of ability and accomplishment. This is discussed in the section below.

1.5 Achievement status and growth as indicators of educational effectiveness

1.5.1 Ability and accomplishment

According to Guo (1998) "ability is a more stable trait than achievement and tends to be determined by both environmental and generic factors early in life" (p. 275). It is further influenced by the interactions between the environment and genetic factors (ibid.). Similarly, Haveman & Wolfe (1995) point out, that next to the genetic and cultural endowments, the availability of economic resources, the nature of these resources and their allocation

contribute to children's educational attainment. Cultural endowments are for example reflected in educational expectations and ambition conveyed within families (ibid.).

Accomplishment is more similar to learned skills (Guo, 1998) and believed to be reflected in achievement. The capability to acquire skills is naturally related to ability, but moreover to individual traits. "Whether individuals realize, fail to realize, or exceed their intellectual potential is often heavily influenced by factors such as motivation and opportunity" (ibid., p. 256). Achievement is therefore a function of ability and accomplishments. This view is also supported by Hofer, Kuhnle, Kilian, and Fries (2012) who investigate the influence of strength of self-control on various outcomes, including academic achievement. The authors also differentiate between cognitive ability and personality variables as predictors of educational attainment. Constructs such as the use of time structure, academic procrastination, and motivational interference are regarded as prerequisites of efficient learning. They reflect students' managing capabilities and are assumed to predict achievement (Hofer et al., 2012).

Conceptually, the distinction between ability and accomplishment has important implications for the research on educational effectiveness. In accordance with EER ability and its association with individual background factors would rather be reflected in achievement levels while accomplishment is reflected in achievement growth that is believed to be unconfounded with individual background factors (Andrejko, 2004; Ballou et al., 2004; Martineau, 2006; OECD, 2008; Thomas, 1998; Zvoch & Stevens, 2008).

There are of course criticisms that question whether the concepts of ability and accomplishment can be distinguished at all. The critique is mainly addressed towards the difficulty to distinctively measure those concepts and therewith the lack of valid empirical evidence (Humphreys, 1974; Lohman, 2006). There are, however, tests specifically designed to measure ability (cognitive ability tests) and those designed to test achievement and the therein reflected accomplishment. While ability tests administer broader problem solving items, achievement tests are much more tied to formal education, i.e. curricula (Guo, 1998). Since, achievement is a function of ability and accomplishment, an absolute distinction between the two concepts is not possible.

Nevertheless, there is empirical evidence that socioeconomic background factors (environmental factors or cultural endowments according to Guo (1998) and Haveman and Wolfe (1995)) are much stronger related to achievement levels than to achievement growth accomplished over several years (Caro & Lehmann, 2009; Cortina, Carlisle, & Zeng, 2008; Stevens, 2005). It is therefore reasonable to assume that achievement levels reflect to a

substantial extent innate ability, SES and other status characteristics, while achievement growth reflects better the capacity of students to acquire skills (accomplishment) over their school careers and their potential for academic success. Moreover, since the enhancement of students' achievement growth seems less influenced by SES (ibid.), it is thus the sphere in which teachers and schools can have a mentionable impact, i.e. be effective.

1.5.2 Different measures – different quality?

These conceptual considerations yielded in different methodological approaches to obtain effectiveness measures. All of them strive to control for student intake, which most reflects factors associated with ability (cognitive abilities and SES) in order to identify institutional effects on achievement. The most striking difference between the statistical models in EER is how they control for student intake. Essentially they differ with regard to whether they apply a longitudinal design, basing effectiveness on growth and change (and therewith consider prior attainment measures as student intake) or a cross-sectional design, controlling for student intake with individual and contextual background variables.

Achievement growth is often considered the most appropriate criterion to assess the effectiveness and is widely accepted by educational researchers and extensively applied in EER (Teddle, Reynolds, & Sammons, 2000). It is argued that cross-sectional designs do not reflect the fact that learning itself is a process (Willet, 1988), that educational outcomes are a consequence of this process, that schools are changing and that their respective effects are believed to be cumulative (Kennedy & Mandeville, 2000). In growth and change models, prior attainment is the most important and accurate factor that affects subsequent achievement (Thomas & Mortimore, 1996). Further, rates of change are considered unconfounded with predictors of achievement such as socioeconomic and migration background of a student (Andrejko, 2004; Zvoch & Stevens, 2008) and thus better reflect what has been entitled accomplishment.

In the shadow of these arguments it is, however, overlooked, that if achievement levels are rather a function of ability *and* accomplishment, and if we control for factors associated with ability, then theoretically, adjusted measures of achievement status would also reflect what has been entitled accomplishment.

There are further counterarguments measures of achievement growth are faced with. For example, several studies have shown that growth rates, too, although less strong, are associated with SES and other student intake factors (Caro & Lehmann, 2009; Cortina, Carlisle & Zeng, 2008; Guo, 1998). Further, it has been cautioned against the use of prior

achievement data that has been surveyed proximal to the assessment within the same school, arguing that actual school effects are thus reduced (Sammons, 1996; Teddlie et al., 2000). Consequently, as assessments in primary schools offer limited opportunities for prior achievement indicators that have been collected prior to the school entry, Teddlie et al. (2000) argue that use of background characteristics in cross-sectional study designs is more appropriate for effectiveness studies in primary school grades.

Further, although, the evaluation of effectiveness based on growth measures is intuitively sensible and more straightforward, in practice, however, researchers are often confronted with cross-sectional studies that naturally lack measures of prior attainment. As a consequence cross-sectional studies could not be used for the evaluation of effectiveness. The potential of these designs to obtain adjusted measures of school performance by controlling for student intake factors is utilized only seldom.

Moreover, studies investigating whether effectiveness measures obtained from status and growth models indeed differ with regard to their idea of effectiveness, i.e. quality, are lacking in research on educational effectiveness. If indeed effectiveness measures obtained from status and growth models are conceptually different (similarly to ability and accomplishment), it follows that a school would be attributed different ratings of effectiveness, dependent on the measure used. In this sense it is possible that schools are effective with regard to enhancing students' achievement growth, but ineffective with regard to levels of achievement. If we assume, however, that schools are equal with regard to their student intake (and the ability reflected in it), and we do so by statistically controlling for it, the possibility that a school is effective with regard to growth but ineffective with regard to achievement levels seems rather unlikely. This is mainly because achievement levels are the result of earlier learning gains if assessments are implemented in later school grades. It is thus argued that measures of adjusted achievement levels and achievement growth capture the same notion of effectiveness and therewith also quality.

1.6 Research questions

From these elaborations a specific research agenda follows that is explained in further detail.

1.6.1 Achievement growth and school track recommendations

Researchers have largely studied the importance of prior attainment for subsequent achievement levels (Andrejko, 2004; Ballou et al., 2004; Sanders, 2000; Thomas & Mortimore, 1996; Zvoch & Stevens, 2008). They have, however, not investigated the

importance achievement growth may have on further educational outcomes, like school track placements. In Germany as well as in several other educational systems, students are assigned into different school tracks by means of teacher recommendations. Tracking decisions have profound and long lasting consequences for future educational and professional careers of students. School track recommendations are essentially a result of teachers' evaluations of a student's achievement and his or her potential to pursue a certain educational level. In line with the distinction of ability and accomplishment it seems sensible to include a measure of achievement growth in the decision making process for track recommendation, rather than relying on levels of achievement only. Taking students' achievement growth into account would indicate that teachers' also differentiate between ability and accomplishment and reward them respectively.

Hence, the following research question surfaces: *Do teachers base their decisions solely on achievement levels when giving track recommendations or do they additionally consider students' achievement growth over a period of three years in the decision making process?*

This question will be addressed in the first article (Caro, Lenkeit, Lehmann, & Schwippert, 2009) that investigates the role of academic achievement growth in school track recommendations.

1.6.2 Comparability of effectiveness measures obtained from status and growth models

Further, it has been discussed, that ability and accomplishment can conceptually be distinguished. It was argued that ability is not only the results of genetic inheritance but moreover confounded with SES of students (environment). It has moreover been argued that ability is stronger reflected in achievement levels and accomplishment in achievement growth, respectively, and that educational institutional effects are most distinct for accomplishment rather than for ability. Controlling for SES, as is applied in EER's notion of quality, would presumably lever the conceptual distinction between ability and accomplishment and yield comparable effectiveness measures.

Hence, the following questions surface: *Do the different assumptions of effectiveness underlying the status and growth approaches also result in different effectiveness estimates? Are differences in effectiveness estimates of practical relevance, so that e.g. a school is effective with regard to growth but ineffective with regard to status?*

These questions are addressed in Lenkeit (2012). The article explicates comparisons of effectiveness measures obtained from growth and status models.

1.6.3 Effectiveness measures for educational systems based on levels and change in performance

It has been discussed that international LSA evaluate educational systems based on absolute (and relative) standards, too, making implications about their respective quality. The research community has however neglected to establish a notion of “contextualized quality” (effectiveness) that takes differences in contextual conditions between educational systems into account. This neglect is particularly obvious in the presentation of unadjusted achievement scores in country league tables. Although the respective tables provide information about the absolute performances of educational systems and position them in an international context (Mullis et al., 2007, 2008; OECD, 2010), the information seems to be of restricted use for policymakers, who demand policy relevant information about the effectiveness of systems independent of socioeconomic and developmental factors (intake). In this context, several studies have shown the influence of non-malleable economic and developmental factors on the performance of education systems (Baker, Goesling, & Letendre, 2002; Caro & Lenkeit, 2012; Chiu, 2007; Chudgar & Luschei, 2009). In line with the notion of quality applied in EER, that is taking socioeconomic background characteristics on the individual and school level into account, it is argued that a consideration of macro level economic and developmental factors yields effectiveness measures on which basis an evaluation of systems’ “contextualized quality” is possible.

Thus, the following research question is addressed: *Can a measure of effectiveness for educational systems be established by taking their contextual conditions into account? How does this change the picture of high or low performing educational systems?*

The third article (Lenkeit, in press) addresses this question by investigating possible statistical adjustment techniques for data from international LSA. Further, most international comparative studies follow a repeated cross-sectional design, assessing the same age and respectively grade cohorts in e.g. cycles of three (PISA) or five (PIRLS) years. In this context arguments about the distinction of ability and accomplishment are pursued, transferring them to the level of educational systems. It is argued that performance levels of educational system are to a great extent a reflection of differences between their socioeconomic and developmental status (Baker, Goesling, & Letendre, 2002; Caro & Lenkeit, 2012; Chiu, 2007; Chudgar & Luschei, 2009). Changes in performance levels across assessment cycles, however, would reflect much more their capacity to improve less effective systemic structures and processes. Similar to the argument elaborated in section 1.5.2, controlling for socioeconomic and developmental status would lever this conceptual distinction. It can thus

be assumed that effectiveness measures obtained for performance levels and those that capture change between assessments cycles represent similar concepts of effectiveness.

Hence, the following questions arise: *Are effectiveness measures of performance levels and changes in performance levels of educational systems comparable, in the sense that they capture the same notion of effectiveness? Are differences in effectiveness measures obtained for performance levels and change in performance of practical relevance, so that e.g. an educational system is effective with regard to change but ineffective with regard to level of performance?*

The research question is addressed in Lenkeit (in press) where approaches to obtain effectiveness measures from international LSA data are investigated.

1.6.4 Reliability of effectiveness measures obtained from achievement growth

The research literature has provided evidence that measures of achievement growth are often less reliable than measures for achievement levels (Raudenbush, 1995; Stevens, 2005). Poor reliability poses strong restrictions on the accuracy of the predicted progress made (Singer & Willet, 2003). These restrictions are often neglected against the background of conceptual arguments concerned with the distinction of ability and accomplishment, achievement levels and growth and the therewith implied notion of quality. This neglect is, however, fatal especially in high stakes systems and at least distorting for school development research. Considering the fact that evidence from investigations of educational effectiveness result in relevant policy decisions, the evaluation of the statistical properties of the obtained effectiveness measures has been alarmingly ignored. Despite the conceptual arguments in favor of growth measures, construct heterogeneity of the measured outcome, the choice of the metric as well as intense mobility of students across schools (especially in urban areas) are threats to unbiased measures of effectiveness in longitudinal data designs (Doran & Cohen, 2005; Goldschmidt, Choi, Martinez, & Novak, 2010).

Therefore, the following question is posed: *What characteristics have the empirical estimates of achievement growth and status and how reliable are they?*

This question is particularly addressed in (Lenkeit, 2012), and further discussed (Caro et al., 2009).

1.7 Significance

The thesis examines whether the conceptual distinction between ability and accomplishment as reflected in achievement levels and achievement growth, can be empirically observed. In particular, it investigates if and how this distinction is reflected in the specific notion of educational quality EER refers to. The significance of this investigation reveals itself along the following argumentation.

In EER growth measures are regarded most appropriate to assess educational effectiveness (Teddle et al., 2000). This view is strongly related to the conceptual distinction between ability and accomplishment (Guo, 1998) and the belief that the impact of educators' and schools' work is best reflected in measures unconfounded with non-malleable student background characteristics (Ballou et al, 2004). This conviction often excludes (if nothing else criticizes) cross-sectional studies for investigations of effectiveness. Nevertheless, EER's specific notion of quality implies the control of student SES because it is determined as a non-malleable background characteristic (Thomas, 1998). But, as has been argued, it is also a determinant of ability (reflected in achievement levels) (Haveman & Wolfe, 1995). Hence, while controlling for the influence of SES, the question arises, whether achievement levels and achievement growth yield distinguishable effectiveness measures? Therewith the somewhat impeachable standing of growth measures is confronted with the argument that effectiveness measures obtained from cross-sectional data are just as appropriate for studying effectiveness.

And further, faced with evidence of the often poor statistical reliability of growth measures (Stevens, 2005), it is argued that they are in fact potentially less suitable to evaluate quality than effectiveness measures obtained from cross-sectional studies.

Investigations made in this thesis regarding the conceptual comparability of effectiveness measures obtained from cross-sectional and longitudinal data of national and international assessments are significant for researchers and policymakers. Researchers are restricted in their use of cross-sectional data, but lack sufficient comparative evidence that measures of achievement growth are indeed better predictors of effectiveness. Therewith the investigations make a contribution to advance the EER field in its discussion about the appropriateness of achievement levels or achievement growth to obtain effectiveness measures. From the results of the investigations implications can be derived that are especially important to educational policymakers who base their decisions on the seemingly established research results of the EER field.

In particular, the three successive and interrelated articles are of relevance for more specific aspects within the educational research field.

Primary, the first article (Caro et al., 2009) broadens our knowledge about the underlying mechanisms of school track recommendations and the relevance of achievement growth for these recommendations. Although research has provided evidence for the relevance of SES (e.g. Baumert & Schümer, 2001), class and school composition characteristics (e.g. Tiedemann & Billmann-Mahecha, 2007), and gender (Updegraff, Eccles, Barber, & O'Brien, 1996), it has neglected the role of achievement growth for track recommendations. By investigating the role of achievement growth the study provides evidence whether teachers distinguish the concepts of ability and accomplishment and reward them in their evaluation and prediction of students' academic potential.

Following, based on this empirical reflection of the distinction between the concepts of ability and accomplishment, the second article (Lenkeit, 2012) investigates if schools are differently effective for achievement levels and achievement growth. It thereby also investigates if the distinction between ability and accomplishment is still reflected in SES-adjusted achievement scores, i.e. effectiveness measures. Although, EER has been concerned with comparisons of effectiveness measures predicted with and without prior attainment estimates, (Sammons, et al., 1993), only Zvoch and Stevens (2008) have compared effectiveness measures for achievement levels and achievement growth. Their investigation did however not surpass a comparison of results from inferential analyses with descriptive statistics of student achievement status and growth. Hence, inquiries made in Lenkeit (2012) contribute to the EER field by investigating whether effectiveness measures obtained from status and growth models differ with regard to their conceptual idea of the quality of a school's work.

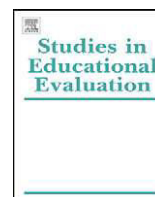
Finally, the third article (Lenkeit, in press) transfers the accomplishments of the EER field to the level of educational systems. So far cultural (e.g. Bank & Heidecke, 2009) and structural economic (e.g. Caro & Lenkeit, 2012) differences between educational systems precluded researchers from extrapolating 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 the overall quality of national educational systems. In this context the article applies established methodological approaches to evaluate quality on school and classroom levels independent of structural economic differences to cross-cultural comparisons. Thus, the article attempts to establish a link between the fields of international LSAs and EER by developing effectiveness indicators for

educational systems. With that, the article seeks to contribute to the analytical approaches for reporting results of international LSA studies.

Furthermore, and stronger aligned to the overall significance of the thesis, the third article, too, compares effectiveness measures obtained for performance levels with those obtained for change in performance. Thereby it investigates whether the distinction between the concepts of ability and accomplishment can roughly be transferred to the level of educational systems and if this distinction can be empirically observed. One would, however, rather distinguish performance that is to a great extent confounded with socioeconomic and developmental status of the systems and efforts undertaken to change and improve less effective structures and processes of educational systems.

CHAPTER 2 STUDY 1

Daniel H. Caro, Jenny Lenkeit, Rainer Lehmann, & Knut Schwippert (2009). The role of academic achievement growth in school track recommendations. *Studies in Educational Evaluation*, 35, 183-192.



The role of academic achievement growth in school track recommendations

Daniel H. Caro^{a,*}, Jenny Lenkeit^b, Rainer Lehmann^c, Knut Schwippert^b

^a IEA Data Processing and Research Center, Mexikoring 37, 22297 Hamburg, Germany

^b Universität Hamburg, Fachbereich Erziehungswissenschaften, Sektion I: Allgemeine, Interkulturelle und International vergleichende Erziehungswissenschaft, Binderstrasse 34, D-20146 Hamburg, Germany

^c Humboldt-Universität zu Berlin, Philosophische Fakultät IV, Institut für Erziehungswissenschaften, Abteilung Empirische Bildungsforschung und Methodenlehre, Unter den Linden 6 (Sitz: Geschwister-Scholl-Str.7), 10099 Berlin, Germany

ARTICLE INFO

Article history:

Received 3 July 2009

Received in revised form 6 November 2009

Accepted 11 December 2009

Keywords:

School tracking

Achievement growth

Empirical Bayes estimator

Student evaluation

ABSTRACT

Students in Germany are tracked into different forms of secondary schooling based on teachers' recommendations. The literature shows that school tracking is largely affected by academic achievement levels, but neglects the influence of individual achievement growth. The authors used data from the Berlin study ELEMENT ($N = 2242$) to characterize math growth trajectories, obtain reliability-adjusted measures of individual growth, and evaluate their effect on teacher's recommendations. The findings suggest that teachers reward math growth while issuing track recommendations. Females, immigrants, and higher SES students are more likely to obtain a college track recommendation other things being equal. And, the probability of a college track recommendation decreases in classes with higher achievement levels and smaller proportion of immigrants.

© 2009 Elsevier Ltd. All rights reserved.

Several European countries track students into different school types in the transition from primary to secondary schooling. Tracking decisions have profound and lasting consequences for future educational and professional careers of students. Their underlying mechanisms are therefore of great interest to education researchers and policy makers. Extensive research has documented that school track placements are largely influenced by prior academic performance of students and that family SES plays an additional role in that parents with high levels of education or employed in high-prestige occupations are more likely to enroll their children in the academic track (i.e., the track leading to college education) than those from low SES families even when their children have comparable levels of academic performance (Baumert & Schümer, 2001; Bos et al., 2004; Ditton, 2007; Ditton & Krüsken, 2006; Lehmann & Peek, 1997; Merckens & Wessels, 2002; Schnabel, Alfeld, Eccles, Köller, & Baumert, 2002). Boudon's theoretical model is often invoked to frame family SES and academic performance influences at this transitional point (Boudon, 1974; Maaz, Trautwein, Lüdtke, & Baumert, 2008).

Next to academic achievement levels and family SES, recent research has shown the influence of aspects such as cultural capital (Condrón, 2007), class and school composition characteristics

(Tiedemann & Billmann-Mahecha, 2007; Trautwein & Baeriswyl, 2007), and gender (Updegraff, Eccles, Barber, & O'Brien, 1996) on the tracking decision. The literature, however, has neglected the role of academic achievement growth in track recommendations, in spite of increased attention of educational researchers in growth rather than status in learning (Willet, 1988). In words of Willet (1988) "The very notion of learning implies growth and change" (p. 346). Conceptually, the distinction between achievement levels and growth has important implications for the research on school tracking. Whereas achievement levels reflect to a substantial extent ability levels, family SES, and other status characteristics, achievement growth reflects better the capacity of students to acquire skills over their school careers and their potential for academic success.

The German decree for primary school level establishes that irrespective of a child's origin, he/she shall enter a path of the education system in accordance with his/her capacity to acquire skills, aptitude, disposition, and its will to intellectual work (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2006). Thus, from a theoretical and policy perspective, the question of whether the capacity to acquire skills is valued for school track placements surfaces. The current analysis adds to the research on school tracking by evaluating whether teachers reward the achievement growth of students while issuing track recommendations. The guiding question is: Are students who have grown more rapidly in their skills more likely to be recommended to the academic track and therefore benefit from further educational opportunities irrespective of their background and initial achievement levels?

* Corresponding author. Tel.: +49 40 48 500 672; fax: +49 40 48 500 501.

E-mail addresses: daniel.caro@iea-dpc.de (D.H. Caro),

jenny.lenkeit@uni-hamburg.de (J. Lenkeit), rlehmann@educat.hu-berlin.de (R. Lehmann), knut.schwippert@uni-hamburg.de (K. Schwippert).

Secondary research questions are: How is achievement growth affected by socioeconomic and demographic characteristics? To what extent does family SES affect school track recommendations indirectly via academic achievement and directly when achievement is controlled? And what are the specific gateways for the effect of SES? Are students with immigrant background less likely to obtain a recommendation for the academic track once family SES and academic achievement are controlled? What is the effect of reference group characteristics on school track recommendations? They are addressed using three measurement points from the Berlin study ELEMENT (Grades 4–6). ELEMENT reports can be consulted in [Lehmann and Nikolova \(2005\)](#) and [Lehmann and Lenkeit \(2008\)](#).

The analysis proceeds in two stages. First, predictors of achievement growth are evaluated and reliability-adjusted measures of individual growth are estimated. Secondly, antecedents of school track recommendations are identified, placing emphasis on the role of achievement growth. The analyses help clarify how school track recommendations are affected by individual, family, and school factors. Methodological examinations are innovative and advance prior research in various respects. Data are more recent and include a greater source of intra-individual variability than in past studies (i.e., three measurement points as opposed to two measurement points or cross-sectional designs). Estimates of achievement growth are adjusted for reliability with the Bayes estimator.

The remainder of this article is organized as follows. Firstly, it describes school tracking policies and regulations in Berlin. Secondly, it discusses theories regarding the antecedents of school track placements. Thirdly, it describes the data, dependent variables, independent variables, and the analytical strategy. Fourthly, it reports the results of longitudinal and multilevel models of achievement growth and school track recommendations, respectively. Finally, the last section discusses main findings, limitations, and recommendations for further research.

1. School tracking in Berlin

Students in Germany are tracked into – traditionally three – different forms of secondary schooling by the end of fourth or sixth grade: the lowest track (*Hauptschule*), the intermediate track (*Realschule*), or the academic track (*Gymnasium*). While the academic track covers both lower and upper secondary levels and provides qualifications and certificates to enter higher education, the intermediate track gives students a more general education and the opportunity to enter more cognitive demanding apprenticeships as full-time vocational training courses at upper secondary level. The lowest track provides students with a basic general education in lower secondary level only often followed by less cognitive demanding handcrafts. Also, some states offer secondary schools that integrate the traditional tracks in a single comprehensive school (*Gesamtschule*).

In Berlin the majority of children start secondary schooling after Grade 6, while less than approximately 10% of the age cohort enter the academic track (*Grundständiges Gymnasium*) after Grade 4 already. For those starting secondary school in Grade 7, the initial decision as to which track a child enters lies within the hands of the parents. Nevertheless, teachers have to articulate a school track recommendation for each child at the end of Grade 6. In most cases they are not legally binding, but parents decide accordingly ([Arnold, Bos, Richert, & Stubbe, 2007](#); [Lehmann & Lenkeit, 2008](#)). The receiving secondary school, however, is allowed to reject a child if discrepancies with the given recommendation occur ([Grundschulverordnung, 2005](#); [Maaz, Neumann, Trautwein, Wendt, Lehmann, & Baumert, 2008](#)).

Specific regulations for this recommendation are issued by the Berlin Senate. The school track recommendation is intended to be given on the basis of the child's performance and the observed competencies. Since 2002 the estimation of a child's suitability for one or the other track is determined by the grades in core subjects from school years 5 and 6, which are averaged to an overall grade. Thus, grades from the subjects German, first foreign language, mathematics, and science are accounted for twice. Up to an overall grade of 2.2,¹ recommendation for the academic track is granted. Students within the range of 2.8–3.2 receive a recommendation for the intermediate track, while in case of an overall grade equal to or lower than 3.8, a recommendation for the lowest track is mandatory. In cases where overall grades fall in between these ranges, teacher's judgments of the student's learning skills is decisive for the tracking recommendation ([Grundschulverordnung, 2005](#)). According to this scheme, school track recommendations are essentially linked to students' academic performance as reflected in school grades ([Thiel, 2008](#)).

2. Factors affecting school track recommendations

When discussing social inequality in educational careers, different stages of transition are viewed as an important source of this inequality. In any form of tracking, there are normally higher percentages of students from advantaged backgrounds attending the more demanding and qualifying tracks. Social selectivity in the transition process has been well documented in the research literature, not only in the German context, with its explicit between-school tracking (e.g. [Arnold et al., 2007](#); [Ditton & Krüsken, 2006](#); [Lehmann & Peek, 1997](#)), but also in other national contexts, where tracking is established more implicitly (e.g. [Caro, McDonald, Willms, 2009](#); [Condon, 2007](#); [Dauber, Alexander, & Entwistle, 1996](#); [Schnabel, Alfeld, Eccles, Köller, & Baumert, 2002](#)).

[Boudon \(1974\)](#) distinguished primary and secondary effects of family SES to characterize the diverse mechanisms by which inequalities are amplified at points of transition. For example, in the transition from primary to secondary school, primary effects are all those expressed via the impact of family SES on academic achievement which, in turn, affects school track recommendations issued by teachers. And secondary effects are those expressed via disparate educational choices among students of comparable achievement levels but of differing family SES. While primary effects largely explain the influence of family SES on school track placements in Germany, secondary effects are also significant ([Arnold et al., 2007](#); [Ditton & Krüsken, 2006](#); [Ditton, Krüsken, & Schauenburg, 2005](#); [Tiedemann & Billmann-Mahecha, 2007](#)). They appear to be less critical for teacher's school track recommendations, though. [Baumert and Schümer \(2001\)](#), [Ditton et al. \(2005\)](#) and [Ditton & Krüsken \(2006\)](#) showed that parents' aspirations for their children's careers depend less on academic achievement than do teachers' recommendations, leaving secondary effects of social reproduction decreased by the latter.

In a German census in the city of Hamburg, [Lehmann and Peek \(1997\)](#) found that teachers generated different *critical values* for the academic track recommendation for different groups of students. Students from lower SES backgrounds on average had to reach higher levels of achievement than those from more advantaged background to obtain a recommendation for the academic track. These effects were also found by [Bos and Pietsch \(2005\)](#) in the census study KESS (*Kompetenzen und Einstellungen von Schülerinnen und Schülern*). They are more than often mediated by the school grades given by teachers and upon which the track recommendations are based.

¹ In the German system the Grade 1 describes the best marking while the Grade 6 is the worst marking as "insufficient".

Although in Germany recommendations seem to be based mainly on achievement, that is, the given grades (Arnold et al., 2007; Ditton & Krüsken, 2006; Ditton et al., 2005; Kristen, 2002; Lehmann & Peek, 1997), teachers also include other characteristics in this complex diagnostic decision. Arnold et al. (2007) found that the perceived parental involvement in school issues, educational valuation, and the cultural fit are also considered. Moreover, the influence of several characteristics of a social and cultural nature has been established. There is, however, no consensus on whether the sources of those effects are intentionally discriminating teachers or whether they should be interpreted as more or less subconsciously perceived general dispositions and aptitudes, which teachers, like any other person, are not immune against.

In Germany students with migration background are, on average, of an academically less successful group and often have highly adverse educational careers. But, despite common belief, migration background of students does not seem to affect track recommendations. After controlling for general cognitive competences and achievement scores, Arnold et al. (2007), Ditton et al. (2005), Kristen (2002, 2006), and Tiedemann and Billmann-Mahecha (2007) could not make out significant effects on the given grades or on track recommendations. Furthermore, Lehmann and Peek (1997) found smaller *critical values* for children with migration background for an academic track recommendation. They also found that females were more likely to obtain a recommendation to the academic track when their academic achievement levels were controlled. Additionally, Arnold et al. (2007) and Trautwein and Baeriswyl (2007) found moderate effects for academic self-concept, fear to fail, and willingness to make an effort after controlling for academic achievement.

Overall, it thus seems that track recommendations are mainly based on academic achievement levels. However, this impression has to be relativized in the sense that recommendations are mainly based on grades given to the students, which in turn are highly dependent on the average ability in the class. Already in some classical analyses (Marsh, 1987; Ingenkamp, 1969), and more often still in recent investigations (e.g. Ditton et al., 2005; Lehmann & Peek, 1997; Trautwein & Baeriswyl, 2007; Treutlein, Roos, & Schöler, 2008) a systematic negative relationship between the average academic achievement of a class and the given individual grade has been revealed. Given two students with comparable ability, the one in a relatively low achieving class receives a better grade than the one in a high achieving class. This is due to teachers evaluating students with reference to the group they teach rather than with regard to external criteria such as performance standards or competency levels. To the extent that teachers' track recommendations are based on grades, students with similar competencies may receive different track recommendations. Reference group effects are one of the main sources leading to the broad overlap of competencies observed in and between the secondary school tracks in Germany.

In another study, Kristen (2002) examined whether the composition of the class with regard to migration background shows similar effects. She found that chances to receive a recommendation for the academic or intermediate track decrease with increasing percentage of students with migration backgrounds in the class while controlling for achievement. Instead, Lehmann and Peek (1997) found that chances to receive an academic track recommendation decrease with decreasing percentage of students with migration background in the class and Tiedemann and Billmann-Mahecha (2007) found no such effects at all. Moreover, the latter found beneficial effects of an averagely disadvantaged class composition with regards to socioeconomic background for chances to receive an academic track recommendation.

3. Data

The data stem from the German longitudinal study ELEMENT ($N = 4925$; Lehmann & Lenkeit, 2008; Lehmann & Nikolova, 2005). In the years 2003–2005, ELEMENT gathered academic achievement, socioeconomic, and demographic information of students in Berlin at the beginning of Grade 4, end of Grade 5, and end of Grade 6. The study covered 3168 untracked students (64%) and 1757 (36%) students already assigned to the academic track (*Grundständiges Gymnasium*). Because this study is concerned with the mechanisms underlying track recommendations at the end of Grade 6, i.e., the regular point of transition in Berlin, students in the academic track are not part of the target population. Also, out of the original sample ($N = 3168$), students with less than three time point observations (27%) and those who have attended different classes over time (2%) are excluded in order to obtain relatively reliable information on growth and to allow for teachers to observe the progress of students over these grades prior to their recommendations, respectively. The sample available for analysis consists of 2242 students.

Although students excluded from the original sample come from less advantaged backgrounds and have lower academic achievement, differences between the original sample and the analytic sample are very small (see Table A.1 in Appendix A). For instance, the percentage of foreign students (25% versus 26%), the percentage of German students with immigrant background (10% versus 11%), the percentage of females (48% versus 49%), the average parental occupational status (46.3 points and 46.9 points), and the average SES (0 points and 0.03 points) are notably similar. Sample attrition is therefore unlikely to seriously bias model estimates.

Missing values were imputed for mother's education (26%), father's education (28%), mother's vocational training (25%), father's vocational training (27%), family's occupational status (25%), and track recommendations (11%). Multiple imputation methods were used to predict missing values of these variables from the available data (including dependent variables). Particularly, multiple imputation by chained equations (MICE) was carried out to generate 5 imputed versions of the raw data (Royston, 2004, 2005) and the Rubin's rule (1987) was applied to estimate standard errors that account for missing data uncertainty. Mean and standard errors of dependent and independent variables are reported in Table A.1 in Appendix A.

3.1. Dependent variables

Math achievement: It is derived from a battery of 49 selected items from the LAU study in the city of Hamburg (Lehmann, Gänßfuß, & Peek, 1998), the IGLU study (Bos et al., 2003), and the QuaSum study in the federal state of Brandenburg (Lehmann et al., 2000). Essentially, test items measure skills in arithmetic and geometry. They were scaled using Item Response Theory (IRT) with 15 over-lapping items vertically equated to create a longitudinally comparable scale suited to assess student growth. Math IRT scores ($M = 100$, $SD = 15$) are reliable ($\alpha = 0.92$).

Math school grade: They are math school grades given by teachers in Grade 6. Scores range from 1 to 6, where 1 indicates best performance.

Track recommendation: It is a dichotomous variable distinguishing academic track recommendations (value of 1) from recommendations to lower tracks (value of 0).

3.2. Independent variables

Basic cognitive abilities: Two sub-tests including 44 items of the *Basic Cognitive Skills Tests* (KFT for the abbreviation in German; Heller & Perleth, 2000) evaluate basic cognitive skills of students

by the end of Grade 4. Specifically, they assess verbal and figurative reasoning and provide an indication of fluid intelligence. The complete version of the test is reliable ($\alpha = 0.93$). The basic cognitive skills variable is the raw score (0–44).

Age: It is the age of the student in years.

Sex: It is a dichotomous variable distinguishing females (value of 1) from males (value of 0).

Parental schooling: Parents reported their highest level of schooling. Responses were classified into: (1) none/special education, (2) secondary school – lowest track, (3) secondary school – intermediate track, (4) admission level for technical college, and (5) admission level for university. The parental schooling variable is the higher schooling level of either parent.

Parental vocational training: Parents were asked if they had obtained any vocational training certificate: (0) no training certificate, (1) apprenticeship certificate, (2) college or commercial school certificate, (3) technical college, master craftsman, or technical school certificate, (4) technical degree or diploma, (5) university degree, and (6) doctoral degree. The mother's and father's vocational training variable corresponds to the highest vocational training certificate obtained, where *no training certificate* (value of 0) is the lowest certificate and *doctoral degree* the highest (value of 6). The parental vocational training variable is the higher level of vocational training of either parent.

Parental occupational status (HISEI): Occupational data for both the father and the mother were obtained with open-ended questions. Lehmann and Nikolova (2005) classified these responses in accordance with Erikson, Goldthorpe and Portocarero (1979) and then mapped them to the International Socioeconomic Index of Occupational Status (ISEI; Ganzeboom, de Graaf, & Treiman, 1992). HISEI corresponds to the higher ISEI score of either parent. Scores range from 16 to 85, where higher values indicate a higher occupational status.

Family SES: It is a composite measure of 5 variables: mother's education, father's education, mother's vocational training, father's vocational training, and parental occupational status (HISEI). Principal component analysis was applied to these data to obtain a single SES variable. The SES index was standardized to have a mean of 0 and a SD of 1 for the student population represented by the analytic sample.

Migration background: Two dichotomous variables were created to distinguish German students with migration background and foreign students from German students without migration background (reference group). Lehmann and Nikolova (2005) used data on German citizenship, the student's mother tongue, language spoken at home, and the country of birth of the student, and his/her parents to define these categories.

4. Two-stage analytical strategy

In a first stage, growth models of math measurements (level 1) nested within students (level 2) are estimated to characterize individual achievement growth trajectories and examine the effect of various socioeconomic and demographic variables on math achievement growth. Also, reliability-adjusted measures of initial achievement level and achievement growth are calculated for each student with the Empirical Bayes (EB) estimator. And potential sources of bias due to ceiling effects, achievement growth measuring ability rather than skills, and regression toward the mean are considered.

In the second stage, multilevel models of students (level 1) nested within classes (level 2) are estimated to evaluate the effect of achievement growth (EB), achievement levels (EB), family SES, migration background, gender, and group reference characteristics on the formation of teacher's track recommendations. Particularly, predictors of the math school grades given by teachers in Grade 6 and of the probability of obtaining a recommendation to the academic track are evaluated. The analyses place special attention to the influence of the EB estimate of achievement growth in math over the last three years of primary school.

5. Math growth curves

Table 1 reports estimates of math growth models. Independent variables were grand-mean centered and age was centered at 10 years to have a meaningful intercept and reduce the degree of multicollinearity arising from the correlation between age and its squared term. The initial status (intercept) can be interpreted as the expected value of math achievement at age 10 given a student population with average characteristics in independent variables.

Table 1

Math growth models: Socioeconomic background predictors of initial status and growth (unstandardized regression coefficients).

Fixed effects	(1)	(2)	(3)	(4)	(5)
Initial status (intercept)	92.01***	91.08***	91.08***	91.07***	91.26***
Family SES			5.95***		
Parental schooling				2.20***	1.94***
Parental vocational training				−0.01	0.10
Parental occupational status				0.25***	0.19***
German with migration background (ref: German)					−5.58***
Foreign (ref: German)					−5.95***
Sex (female = 1)					−5.29***
Growth rate (age)	9.59***	11.28***	11.25***	11.26***	11.01***
Family SES			0.79***		
Parental schooling				0.30*	0.32
Parental vocational training				0.07	0.07
Parental occupational status				0.02	0.02
German with migration background (ref: German)					−0.34
Foreign (ref: German)					0.02
Sex (female = 1)					0.74***
Acceleration rate (age ²)		−0.56***	−0.55***	−0.55***	−0.47***
Random effects			SD		
Intercept	13.14***	13.56***	12.20***	12.29***	11.70***
Age	2.95***	2.75***	2.62***	2.64***	2.66***

Note: Estimation method is hierarchical linear models (HLM). Sample consists of 2242 students. All variables were grand mean centered. Five imputed datasets account for missing data uncertainty and cases were weighted to represent the student population.

* $p < 0.10$.

*** $p < 0.01$.

The growth rate (age coefficient) is the rate of change in math achievement at age 10. The acceleration parameter (age-squared coefficient) captures the acceleration in the entire growth trajectory. Random effects were introduced for the initial status and growth rates to allow for these parameters to vary among students.

Students grow significantly in their math skills with age as they advance in school from Grade 4 to 6. Measured by the growth rate coefficient, they grow in 9.6 score points per year, that is, two-thirds of a SD of the math measure (see model 1 in Table 1). They do not grow at a constant rate of change over this period, though. The negative estimate of the acceleration parameter (see model 2 in Table 1) indicates a curvilinear growth trajectory, particularly, that students grow in their math skills at a decreasing rate of change. The rate of change decelerates on average 12% per year, from 11.3 score points at the age of 10 to 7.9 score points at the age of 13.

Fig. 1 depicts actual growth trajectories for 100 students randomly selected from the population and the fitted trajectory for the population of students. As expected, the fitted line shows a slightly curvilinear growth curve. Additionally, observed growth trajectories anticipate significant variation in initial achievement levels and growth rates among students. Note that initial achievement levels and growth slopes differ among students. Model estimates confirm that variation around the grand mean of the initial status and the growth rate is statistically significant (see random effects in Table 1).

6. Predictors of initial status and math growth

Because random effects for the initial status and growth rate are statistically significant, it is possible to model inter-individual variation in these parameters. Models 3–5 in Table 1 include stepwise socioeconomic and demographic independent variables to evaluate their effect on the initial status and achievement growth of students. Family SES is positively related to initial achievement levels. For an increment of 1 SD in SES, math achievement increases in 5.95 score points (see model 3 in Table 1), that is, in about 40% of a SD of the math measure. Furthermore, results of model 3 (see Table 1) indicate that family SES contributes to higher growth rates. Measured by the reduction of the SD of the growth rate random component, SES accounts for 5% of the differences in growth among students.

When the effect of family SES is broken down (see model 4 in Table 1), the relationship between SES and initial achievement

levels is found to be mostly driven by parental schooling and parental occupational status. Children whose parents have attained higher educational levels or are employed in higher-prestige occupations have higher achievement levels. For 1 SD increment in parental education and parental occupational status, math achievement increases in 2.7 and 3.9 score points. There is weak evidence that parental schooling contributes to explain the relationship between family SES and achievement growth. Note that the parental education coefficient is positive but significant at 10% only (see model 4 in Table 1). Otherwise, none of the SES characteristics have a statistically significant effect on the growth rate when evaluated separately.

Irrespective of family SES, children with migration background perform worse than the rest in math. Differences in math achievement attributed to migration background when SES is controlled amount to about one-third of a SD of the math measure and are thus considerable (see model 5 in Table 1). There are no apparent differences in growth rates related to migration background. Girls have lower achievement levels than boys but, interestingly, they grow more rapidly in their math skills (see model 5 in Table 1). Gender differences at the age of 10 amount to one-third of a SD of the math achievement measure and reduce in about 40% by the age of 13.

Growth differences remain statistically significant when family SES, migration background, and sex are controlled. Yet, limited intra-individual variation in math achievement due to the three data point design precludes comprehensively evaluating predictors of growth.

7. Empirical Bayes estimates of initial status and growth

The EB or *shrunk* estimator is applied to model 1 (see Table 1) to obtain individual measures of initial status and growth. Essentially, the EB estimator penalizes OLS estimates for reliability. Its calculation is such that highly reliable OLS estimates of the initial status/growth tend to their individual values and unreliable OLS estimates are *pulled* towards the grand mean estimate (Bryk & Raudenbush, 2002; Lindley & Smith, 1972). Random effects in model 1 enable Bayesian *shrinkage*.

The reliability of OLS estimates of the initial status and the growth rate is 0.69 and 0.35, respectively, indicating that the individual estimate of the initial status is fairly precise and that growth rates are estimated with less precision. In this regard, while the longitudinal design of three math occasions spaced one year apart provides a greater source of intra-individual variability than in most past studies, it is still limited for reliably measuring growth. As a result, reliability-adjusted measures of individual growth calculated with the EB estimator will be biased towards the grand mean. Only with more data points and greater spacing between waves the reliability of the growth measurement can be improved (Willet, 1988).

It should be noted that while the calculation of the initial status estimate depends only upon the OLS initial status estimate and its reliability, the individual math growth estimate depends upon the OLS individual growth estimate, the OLS initial status estimate, and their corresponding reliabilities. In cases where the initial status and achievement growth are highly correlated, EB estimates of math growth can be equally affected by the OLS math growth estimate and the OLS initial status estimate, thereby making their behavior and interpretation more complex (Bryk & Raudenbush, 2002). Note also that the correlation between the initial status and growth varies for different choices of the time at which initial status is measured. Thus, the value of age for the initial status should be chosen for substantive reasons and needs to be declared. Here, the initial status was set at age 10, when most students attended Grade 4. At this point, the correlation between the initial

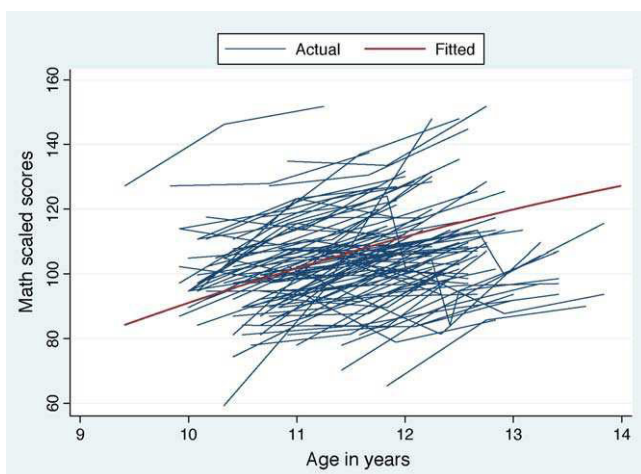


Fig. 1. Math observed and fitted achievement trajectories. Note: Observed trajectories are for 100 students randomly selected from the population. HLM estimates of the initial status, growth rate, and acceleration rate parameter of the math growth curve were used to construct the fitted line (see model 2 in Table 1)

status and math growth derived from model 1 (see Table 1) is 0.18. To the extent that the correlation is small, the calculation of the math growth EB estimate is fairly independent of the calculation for the initial status EB estimate.

8. Potential sources of bias

At least three sources of bias need to be considered when estimating individual measures of achievement growth and their effect on school track recommendations. The first is *ceiling effects* in math growth. Namely, that the math test was not sufficiently difficult to capture growth of best performing students. If this source of bias were present, one would expect a negative relationship between initial status and growth.

Fig. 2 depicts the relationship between the initial status and growth based on the EB estimates. Each dot represents a student and his/her math initial status and growth rate. The fitted line is the linear prediction of math growth using the initial status solely. Clearly, students starting with a higher initial status tend to grow more rapidly in their math skills. To the extent that initial status and growth are positively related, it can be somewhat ruled out that estimates of growth effects on school track recommendation models are an artifact of ceiling effects.

Theoretically, that the initial status and growth rate are correlated is not surprising because a student's status is a consequence of his/her growth history. Since each contains information about the other, an *ability bias* may arise if ability information embedded in the initial status is also reflected in the growth estimate. Growth effects on school track recommendations would confound both ability and skill effects. It should be noted, though, that while the initial status and growth rate are positively related (see Fig. 2), the relationship is not deterministic. A considerable number of students start with low achievement levels and exhibit relatively high growth rates and vice versa, suggesting that growth reflects other aspects besides achievement levels. Furthermore, school track recommendation models in the next section control for basic cognitive abilities and the initial status to counteract this source of bias.

The third source of bias is *regression toward the mean*. It occurs when scores far from the mean in a first observation tend to regress towards the mean in subsequent observations. This phenomenon is most apparent in two time point study designs, where most lucky and unlucky students on the first test will perform worse and better on the second test, respectively. If present, it gives the false impression that growth rates are higher and lower for worst and best performers on the first evaluation, respectively. Here, the

longitudinal design of three measurement points and the use of Bayesian shrinkage limit the effect of extreme unreliable scores on the growth rate and so counteract this source of bias.

9. Achievement growth and school track recommendations

OLS and logistic multilevel models of students (level 1) nested within classes (level 2) were estimated to evaluate the effect of achievement growth and other variables on the math school grades given by teachers by the end of primary school and on the probability of obtaining a recommendation to the academic track, respectively. For ease of interpretation of effect sizes, all except dummy variables were standardized to have a mean of 0 and SD of 1. Also, due to the German grading scale (1–6), where lower school scores represent better school performance and vice versa, the dependent variable is the negative value of math school grades. Estimates of math school grade models and school track recommendation models are reported in terms of unstandardized regression coefficients and odds ratios in Tables 2 and 3, respectively.

Irrespective of math initial achievement levels and basic cognitive skills, math growth contributes to better math school performance and, in turn, to obtain a recommendation to the academic track. For 1 SD increment in math growth, math school grades increase in 0.24 score points and the probability of obtaining a recommendation for the academic track increases by a factor of 1.61 (see model 1 in Tables 2 and 3). While the importance of math growth to primary school exit grades and the track recommendation is considerable, the contribution of initial achievement levels is even more pronounced. For 1 SD increment in the math initial status, math school grades increase almost by half point and the probability of being recommended to the academic track increases by a factor of 3.93 (see model 1 in Tables 2 and 3). The impact of the math initial status on the track recommendation is 2.4 times as large as the impact of math growth.

Students growing more rapidly in their math skills are more likely to obtain an academic track recommendation irrespective of their initial achievement levels. The effect of growth is not constant throughout the range of the initial status scale, though. Note that the interaction effect of the initial status and growth rate is negative and statistically significant (see model 2 in Tables 2 and 3). Apparently, growth effects on math school grades and the track recommendation decrease at higher levels of initial achievement. In other words, the growth of worst performing students at age 10 is valued more strongly by teachers for school track recommendations than the growth of best performing students.

Family SES mediates the relationship between math achievement and the track recommendation. The math initial status coefficient and the math growth coefficient reduce in 8% and 5% when SES is controlled (see models 2 and 3 in Table 3). Family SES is indirectly related to the track recommendation via the effect of math achievement (see model 3 in Table 1), but it also has direct effects. The family SES coefficient remains significant when academic achievement is controlled. For 1 SD increment in SES, the probability of getting a recommendation for the academic track increases by a factor of 1.8 (see model 3 in Table 3). Parental occupational status, parental schooling, and parental vocational training, in this order, drive the direct influence of family SES on the track recommendation.

Apparently, the SES effect on math school grades is mostly an indirect effect expressed via academic achievement. Evidence for an SES direct effect is weak. When academic achievement is controlled, the family SES coefficient amounts to less than 0.1 score points and is significant at the 90% confidence level only (see model 3 in Table 2). Also, when the SES effect is broken down, none of the SES components turns out to be statistically significant (see model 4 in Table 2). Furthermore, in unreported analyses the SES effect on

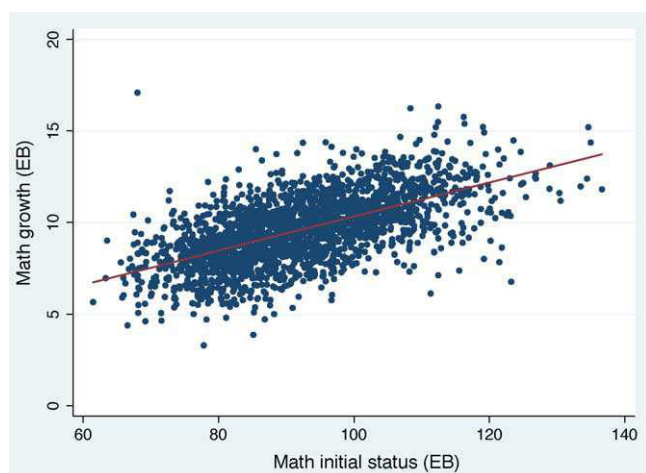


Fig. 2. Initial status and math growth over Grades 4–6.

Table 2

The relationship between socioeconomic background, academic achievement, and math performance in Grade 6.

Student level	Math school score (Grade 6)				
	Beta				
	(1)	(2)	(3)	(4)	(5)
Basic cognitive abilities	0.13***	0.13***	0.13***	0.12***	0.12***
Math initial status (BE)	0.45***	0.46***	0.45***	0.49***	0.52***
Math growth (BE)	0.24***	0.24***	0.24***	0.22***	0.23***
Initial status \times growth		–0.04***	–0.04***	–0.05***	–0.04***
Family SES			0.05*		
Parental schooling				0.07	0.07**
Parental vocational training				–0.01	0.00
Parental occupational status				–0.01	0.00
German with migration background (ref: German)				0.24***	0.18***
Foreign (ref: German)				0.23***	0.17***
Sex (female = 1)				0.18***	0.18***
<i>Class level</i>					
Mean math initial status					–0.18***
Foreign/German with migration background (%)					0.01

Note: Estimation method is random effects panel data regression (Baltagi, 2008). Sample consists of 2242 students. Effect sizes of all except dummy variables are for one SD change. Five imputed datasets account for missing data uncertainty and cases were weighted to represent the student population.

* $p < 0.10$.** $p < 0.05$.*** $p < 0.01$.**Table 3**

The relationship between socioeconomic background, academic achievement, and the track recommendation decision.

Student level	Track recommendation (gymnasium = 1)				
	Odds ratios				
	(1)	(2)	(3)	(4)	(5)
Basic cognitive abilities	1.67***	1.69***	1.64***	1.54***	1.55***
Math initial status (BE)	3.93***	4.22***	3.87***	5.39***	5.55***
Math growth (BE)	1.61***	1.70***	1.62***	1.62***	1.63***
Initial status \times growth		0.74***	0.75***	0.74***	0.74***
Family SES			1.80***		
Parental schooling				1.23***	1.24***
Parental vocational training				1.21**	1.21**
Parental occupational status				1.25**	1.26**
German with migration background (ref: German)				1.65***	1.45***
Foreign (ref: German)				1.24**	1.08
Sex (female = 1)				3.31***	3.34***
<i>Class level</i>					
Mean math initial status					0.85**
Foreign/German with migration background (%)					1.13**

Note. Estimation method is random effects logistic regression (Baltagi, 2008). Sample consists of 2242 students. Effect sizes of all except dummy variables are for one SD change. Five imputed datasets account for missing data uncertainty and cases were weighted to represent the student population.

** $p < 0.05$.*** $p < 0.01$.

math school grades was nonsignificant when both reading and math achievement were controlled, but it was still significant and considerable for the track recommendation.

Students with migration background perform lower in math and are less likely to be recommended to the academic track. On average, Germans without migration background obtain better grades in math (2.83 score points) than Germans with migration background (2.92 score points) and foreign students (3.05 score points). While 45% of German students without migration background obtain a recommendation for the academic track, only 27% of foreign students and 32% of German students with migration background obtain this recommendation. Interestingly, though, when family SES and achievement variables are controlled, the relationship reverses. Students with migration background are given higher grades in math (by 0.24 score points) than Germans without migration background (see model 4 in Table 2). Apparently, they are also more likely to obtain an academic track recommendation: Germans with migration background by a factor of 1.65 and foreign students by a factor of 1.24 (see model 4 in Table 3).

Evidence that girls had poorer math performance in the ELEMENT math test, but grew at faster rates than boys in their math skills was reported earlier (see model 5 in Table 1). Additionally, estimates of model 4 (see Table 2 and 3) indicate that, when academic and socioeconomic variables are controlled, girls are given higher grades in math by the end of primary education (by 0.18 score points) and are more likely to obtain an academic track recommendation by a factor of 3.31.

Other things being equal, students in classes with higher than average math achievement tend to obtain lower grades in math and are less likely to get an academic track recommendation (see model 5 in Tables 2 and 3). The class's mean initial status and proportion of students with migration background (i.e., German students with migration background or foreign students) account for about 25% of math grade differences attributed to migration background (see models 4 and 5 in Table 2). These class composition variables explain the relationship between migration background and the school track recommendation even to a greater extent (see models 4 and 5 in Table 3). Once they are

controlled, foreign students are no longer more likely to obtain an academic track recommendation. Note also that students have higher changes of getting an academic track recommendation in classes where the proportion of students with migration background is higher (see model 5 in Table 3).

10. Discussion

The presented analysis has important theoretical, policy, and methodological implications for the research on school tracking. Its limitations should be recognized and understood, though. One is that it is based on a local study in the city of Berlin restricted to a target group already *creamed off* from a very selective group of early transition into certain advanced placement programs and it is not certain that these findings hold if referring to data from another German federal state or educational system practicing between-school tracking. There is, however, no better data source for informing the issues addressed in this study. Available data sets contain fewer measurement points and/or neglect information on school track recommendations.

Another limitation is data loss due to missing values and attrition. The achievement test data are complete, but about 25% of socioeconomic data are missing. Yet, the multiple imputation by chained equations method (Royston, 2004, 2005) was applied to predict missing values and estimate standard errors that account for missing data uncertainty. Students with less than 3 math measurements (27%) are excluded from the analytic sample to safeguard the reliability of math growth rates. These students come from relatively less advantaged backgrounds. Nevertheless, differences between the original sample and analytic sample are small. Thus, data loss shall not seriously bias the model estimates.

Still another limitation is the low reliability of achievement growth ($\alpha = 0.35$). The EB estimator is used to counteract the lack of precision of individual achievement growth measures. And yet, while the EB estimator penalizes estimates for reliability, this deficiency could only be improved with additional points of measurement. Due to these limitations, the results are best considered suggestive, and certainly odds ratios and unstandardized coefficients reported should not be easily generalized. With these caveats understood, main findings emanating from this study are discussed next.

In general, the findings align well with the literature on school tracking. What is new from the results is that they offer evidence that students *growing* more rapidly in their math skills are more likely to obtain a recommendation for the academic track. This finding is not an artifact of ceiling effects, regression toward the mean, or growth measuring ability rather than skills. The relationship between achievement growth and school track recommendations remains significant even when achievement levels, ability, and a group of socio-demographic characteristics are controlled. Throughout the range of academic achievement, students growing at faster rates are more likely to obtain an academic track recommendation irrespective of their socioeconomic background or ability levels.

The relationship between achievement growth and school track recommendations is partly explained by the association between achievement growth and the math grades given by teachers at the end of primary education (Grade 6). Teachers seem to monitor and evaluate student progress individually in that, irrespective of math achievement levels, they reward growth in math with higher grades. Inasmuch as school track recommendations are largely determined by school grades, math grades directly mediate the relationship between math growth and school track recommendations.

The effect of growth is more pronounced for students starting with low levels of academic achievement than for students with

high initial levels of math achievement. Apparently, teachers reward more strongly the growth of originally academically disadvantaged students than the growth of students starting with higher achievement levels. And yet, while growth is valued for school track recommendations and may contribute to reduce inequalities attributed status characteristics, the level of achievement is a more critical factor. Particularly, the effect of the math achievement level on the track recommendation is 2.4 times as large as the effect of math growth.

Furthermore, from a set of socio-demographic and achievement variables, the level of achievement is the most important predictor of differences in school track recommendations. The influence of achievement growth and family SES, though considerable, is less significant when achievement levels are taken into account. Certainly, this finding raises questions on the extent to which school track recommendations ought to reflect achievement levels (status) and achievement growth (progress). Is the capacity of students to learn equally, more, or less important than their actual achievement levels for their future educational opportunities? This question is beyond the scope of this study but deserves the attention of further research and policy makers.

Other findings emanate from the analyses. Family SES is indirectly related to the track recommendation via its effect on math achievement levels and growth. Not only higher SES students perform better in math than lower SES students but they also grow more rapidly in their math skills (Alexander, Entwisle, & Olson, 2001; Becker, Stanat, Baumert, & Lehmann, 2008; Caro, McDonald, & Willms, 2009). Once achievement variables are controlled, family SES is unrelated to math school grades in Grade 6, but it is still positively associated to the track recommendation. To the extent that the recommendation is based on school grades and teacher individual assessment of students, direct effects of family SES appear to play a role in the individual assessment of students.

Strictly speaking, direct effects of family SES here are not equivalent to the so-called secondary effects of family SES because choices of parents have not been revealed yet (Boudon, 1974). But these effects do announce a source of inequality of opportunity in track enrollment due to teacher influences. Although they are less significant than the primary effects, their presence is particularly troubling because it suggests that high-achieving students of low SES families are in *double-jeopardy*. Compared to high SES students, they are less likely to be recommended to the academic track and, furthermore, even if they obtain the academic track recommendation, their parents are less likely to enroll them in the academic track (Bos et al., 2004; Ditton, 2007; Ditton et al., 2005; Maaz, Trautwein, et al., 2008; Maaz, Neumann, et al., 2008).

That parental occupational status and parental vocational training drive the direct effect of family SES may suggest that teachers perceive and reward the value students and their families attach to education (Arnold et al., 2007). In this regard, research shows that parental vocational training levels and parental occupational preferences are a source of class-based culture and values that influences the value students attach to education (Bourdieu, 1973; Karlsen, 2001; Koo, 2003). Also, scholars argue that socioeconomically advantaged parents instill in their offspring favorable attitudes towards education which, in turn, positively affect their educational plans (Carpenter & Fleishman, 1987; Crosnoe, Mistry, & Glen 2002; Eccles, Vida, & Barber, 2004; Hossler & Stage, 1992).

The results on the relationship between migration background and school track recommendations are quite interesting. When academic achievement and family SES are controlled, students with migration background are given better math school grades and are more likely to be recommended to the academic track (Lehmann & Peek, 1997; Limbos & Geva, 2001). In part, this is explained by the influence of class composition characteristics: where the proportion of immigrant students is higher, students

with migration background are more likely to obtain an academic track recommendation because referrals of teachers are not only based on individual performance but also on the class composition. But even in socioeconomically comparable classes, students with migration background are more likely to obtain an academic track recommendation.

One may speculate, for example, that teachers are less confident about their ability to distinguish academic performance of students with migration background and, thus, accept some level of underachievement in this group as part of their normal development (Limbos & Geva, 2001). Or that teachers in Berlin tend to share liberal beliefs and try to compensate immigrants for their disadvantaged position in the hierarchical social structure by providing them with better chances of pursuing academic careers. And still another argument could be that other class-level socioeconomic and achievement characteristics not considered here entirely explain this effect.

Finally, as with other studies, evidence of a negative relationship between the class mean achievement and the individual math grades and the probability of being recommended to the academic track was found (Maaz, Trautwein, et al., 2008; Maaz, Neumann,

et al., 2008; Tiedemann & Billmann-Mahecha, 2007; Trautwein & Baeriswyl, 2007). Given two students with comparable ability, the one in a class with higher mean math achievement receives a worse grade in math and is less likely to obtain an academic track recommendation. This is due to teachers evaluating students in relation to the group and not to an external criterion.

Acknowledgements

The authors wish to thank Petra Stanat and her research group for helpful comments and suggestions on this research. During the work on his dissertation, Daniel H. Caro was a pre-doctoral fellow of the International Max Planck Research School “The Life Course: Evolutionary and Ontogenetic Dynamics” (LIFE, www.imprs-life.mpg.de; participating institutions: MPI for Human Development, Humboldt-Universität zu Berlin, Freie Universität Berlin, University of Michigan, University of Virginia, University of Zurich).

Appendix A

See Table A.1.

Table A.1

Main statistics of dependent and independent variables.

Characteristic	Original sample (N = 3168)		Analytic sample (N = 2242)		Excluded sample (N = 926)	
	Mean	SE	Mean	SE	Mean	SE
<i>Dependent variables</i>						
Math achievement, Grade 4	95.66	(0.37)	96.58	(0.34)	93.41	(0.96)
Math achievement, Grade 5	105.60	(0.32)	106.78	(0.35)	102.71	(0.69)
Math achievement, Grade 6	113.91	(0.35)	115.03	(0.38)	111.13	(0.75)
Math School Grade, Grade 6	2.98	(0.02)	2.89	(0.02)	3.22	(0.04)
Track recommendation (Academic = 1)	0.36	(0.01)	0.40	(0.01)	0.27	(0.02)
<i>Independent variables</i>						
Basic cognitive abilities	25.24	(0.22)	25.83	(0.24)	23.80	(0.46)
Age in years, Grade 4	10.65	(0.01)	10.60	(0.01)	10.79	(0.03)
Age in years, Grade 5	11.57	(0.01)	11.52	(0.01)	11.70	(0.03)
Age in years, Grade 6	12.49	(0.01)	12.44	(0.01)	12.62	(0.03)
Sex (female = 1)	0.48	(0.01)	0.49	(0.01)	0.44	(0.02)
Parental schooling	3.44	(0.03)	3.48	(0.03)	3.36	(0.07)
Parental vocational training	2.46	(0.05)	2.47	(0.05)	2.43	(0.11)
Parental occupational status	46.27	(0.42)	46.86	(0.40)	44.83	(1.11)
Family SES	0.00	(0.02)	0.03	(0.03)	−0.07	(0.05)
German with migration background	0.10	(0.01)	0.11	(0.01)	0.09	(0.01)
Foreign	0.25	(0.01)	0.26	(0.01)	0.22	(0.02)

Note: Data were multiply imputed (5 data replicates) to account for missing data uncertainty and weighted by their sampling probability to represent the student population.

References

- Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, achievement and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis*, 23, 171–191.
- Arnold, K.-H., Bos, W., Richert, P., & Stubbe, T. C. (2007). Schullaufbahnpräferenzen am Ende der vierten Klassenstufe. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, K. Schwippert, & R. Valtin (Eds.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 271–297). Münster, Germany: Waxmann.
- Baltagi, B. (2008). *Econometric analysis of panel data* (4th edition). Chichester: John Wiley & Sons.
- Baumert, J., & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb [Family background, educational participation, and competence acquisition]. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann, & M. Weiß (Eds.), *PISA 2000 Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 323–407). Opladen, Germany: Leske, Budrich.
- Becker, M., Stanat, P., Baumert, J., & Lehmann, R. (2008). Lernen ohne Schule: Differenzielle Entwicklung der Leseleistungen von Kindern mit und ohne Migrationshintergrund während der Sommerferien. [Learning outside of school: Differential development of reading comprehension in children with and without immigrant background during summer vacation]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 48, 252–276.
- Bos, W., & Pietsch, M. (2005). Erste Ergebnisse aus KESS 4: Regionale, nationale und internationale Einordnung der Ergebnisse. In H.-G. Holtappels & K. Höhmann (Eds.), *Schulentwicklung und Schulwirksamkeit. Systemsteuerung, Bildungschancen und Entwicklung der Schule* (pp. 65–82). Weinheim, Germany: Juventa.
- Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Walther, G., & Valtin, R. (Eds.). (2003). *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich*. Münster/New York/München/Berlin: Waxmann.
- Bos, W., Voss, A., Lankes, E.-M., Schwippert, K., Thiel, O., & Valtin, R. (2004). Schullaufbahnpfehlungen von Lehrkräften für Kinder am Ende der vierten Jahrgangsstufe [Teachers' recommendations for student tracking at the end of fourth grade]. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin, & G. Walther (Eds.), *IGLU. Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* (pp. 191–220). Münster, Germany: Waxmann.
- Boudon, R. (1974). *Education, opportunity, and social inequality: Changing prospects in Western society*. New York: Wiley.
- Bourdieu, P. (1973). Cultural reproduction and social reproduction. In R. Brown (Ed.), *Knowledge, education and cultural change* (pp. 71–112). London: Tavistock.
- Bryk, A. S., & Raudenbush, S. W. (2002). *Hierarchical linear models: Applications and data analysis methods* (second edition). Thousand Oaks, CA: Sage.

- Caro, D. H., McDonald, T., & Willms, J. D. (2009). Socio-economic status and academic achievement trajectories from childhood to adolescence. *Canadian Journal of Education*, 32(3), 558–590.
- Carpenter, P. G., & Fleishman, J. A. (1987). Linking intentions and behavior: Australian students' college plans and college attendance. *American Educational Research Journal*, 24, 79–105.
- Condron, D. J. (2007). Stratification and educational sorting: Explaining ascriptive inequalities in early childhood reading group placement. *Social Problems*, 54(1), 139–160.
- Crosnoe, R., Mistry, R. S., & Elder, G. H. (2002). Economic disadvantage, family dynamics, and adolescent enrollment in higher education. *Journal of Marriage and the Family*, 64, 690–702.
- Dauber, S., Alexander, K., & Entwisle, D. (1996). Tracking and transitions through the middle grades: Channeling educational trajectories. *Sociology of Education*, 69, 290–307.
- Ditton, H. (2007). Der Beitrag von Schule und Lehrern zur Reproduktion von Bildungsungleichheit [The contribution of family and school to the reproduction of educational inequality]. In R. Becker & W. Lauterbach (Eds.), *Bildung als Privileg. Erklärungen und Befunde zu den Ursachen der Bildungsungleichheit* (pp. 243–272). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.
- Ditton, H., & Krüskens, J. (2006). Der Übergang von der Grundschule in die Sekundarstufe I [The transition from primary to secondary schools]. *Zeitschrift für Erziehungswissenschaft*, 9(3), 348–372.
- Ditton, H., Krüskens, J., & Schauenberg, M. (2005). Bildungsungleichheit—der Beitrag von Familie und Schule [Educational equality—the contribution of family and school]. *Zeitschrift für Erziehungswissenschaft*, 8, 285–303.
- Eccles, J. S., Vida, M. N., & Barber, B. (2004). The relation of early adolescents' college plans and both academic ability and task-value beliefs to subsequent college enrollment. *Journal of Early Adolescence*, 24, 63–77.
- Erikson, R., Goldthorpe, J. H., & Portocararo, L. (1979). Intergenerational class mobility in three Western European societies. *British Journal of Sociology*, 30, 415–441.
- Ganzeboom, H. B. G., de Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, 21, 1–56.
- Grundschulverordnung. (2005). *Verordnung über den Bildungsgang der Grundschule. Vom 19. Januar 2005, zuletzt geändert durch Verordnung vom 11. Dezember 2007*. Retrieved March 16, 2009 from <http://www.berlin.de/imperia/md/content/sen-bildung/rechtsvorschriften/grundschulverordnung.pdf>.
- Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision*. Göttingen, Germany: Beltz.
- Hossler, D., & Stage, F. K. (1992). Family and high school experience influences on the post secondary educational plans of ninth grade students. *American Educational Research Journal*, 29, 425–451.
- Ingenkamp, K. (1969). *Zur Problematik der Jahrgangsklasse*. Weinheim, Germany: Julius Beltz.
- Karlsen, U. D. (2001). Some things never change: Youth and occupational preferences. *Acta Sociologica*, 44, 243–255.
- Koo, A. (2003). *Social inequality and differential educational plans*. Paper presented at the EPUNet-2003 and BHPs-2003 Conferences.
- Kristen, C. (2002). Hauptschule, Realschule oder Gymnasium? Ethnische Unterschiede am ersten Bildungsübergang. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 54(3), 534–552.
- Kristen, C. (2006). Ethnische Diskriminierung in der Grundschule. Die Vergabe von Noten und Bildungsempfehlungen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 58(1), 79–97.
- Lehmann, R., & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis. Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin. Abschlussbericht über die Untersuchungen 2003, 2004 und 2005 an Berliner Grundschulen und grundständigen Gymnasien*. Berlin, Germany: Humboldt Universität.
- Lehmann, R., & Nikolova, R. (2005). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis-Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin. Bericht über die Untersuchung 2003 an Berliner Grundschulen und grundständigen Gymnasien*. Berlin, Germany: Humboldt Universität.
- Lehmann, R. H., & Peek, R. (1997). *Aspekte der Lernausgangslage und der Lernentwicklung von Schülerinnen und Schülern. die im Schuljahr 1996/97 eine fünfte Klasse an Hamburger Schulen besuchten -LAU 5- [Components of initial achievement and school learning among pupils who have attended Grade 5 classes in the city of Hamburg during the school year of 1996/97 -LAU 5-]*. Hamburg, Germany: Behörde für Schule, Jugend und Berufsbildung.
- Lehmann, R. H., Gänsfuß, R., & Peek, R. (1998). *Aspekte der Lernausgangslage und Lernentwicklung von Schülerinnen und Schülern an Hamburger Schulen. Klassenstufe 7 -LAU 7- [Components of initial achievement and school learning among pupils in the City of Hamburg. Grade 7 -LAU 7-]*. Hamburg, Germany: Behörde für Schule, Jugend und Berufsbildung, Amt für Schule.
- Lehmann, R., Peek, R., Gänsfuß, R., Lutkat, S., Mücke, S., & Barth, I. (2000). *Qualitätsuntersuchungen an Schulen zum Unterricht in Mathematik (QuaSUM)*. Potsdam, Germany: Ministerium für Bildung, Jugend und Sport des Landes Brandenburg.
- Limbos, M., & Geva, E. (2001). Accuracy of teacher assessments of second-language students at risk for reading disability. *Journal of Learning Disabilities*, 34(2), 136–151.
- Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34, 1–41.
- Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008a). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives*, 2(2), 99–106.
- Maaz, K., Neumann, M., Trautwein, U., Wendt, W., Lehmann, R., & Baumert, J. (2008b). Individuelle Lernkompetenzeinschätzungen von Lehrkräften beim Übergang von der Grundschule in die weiterführende Schule: Die Rolle von Schüler- und Klassenmerkmalen [Teacher assessments of individual student competence at the transition to secondary education: The role of student and class characteristics]. *Revue suisse des sciences de l'éducation*, 30(3), 519–548.
- Marsh, H. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79(3), 280–295.
- Merkens, H., & Wessels, A. (2002). *Zur Genese von Bildungsentscheidungen. Eine empirische Studie in Berlin und Brandenburg*. Baltmannsweiler: Schneider Verlag Hohengarten.
- Royston, P. (2004). Multiple imputation of missing values. *Stata Journal*, 4, 227–241.
- Royston, P. (2005). Multiple imputation of missing values: Update. *Stata Journal*, 5, 88–102.
- Rubin, D. B. (1987). *Multiple imputation for non-response in surveys*. New York: John Wiley & Sons.
- Schnabel, K., Alfeld, C., Eccles, J., Köller, O., & Baumert, J. (2002). Parental influence on students' educational choices in the United States and Germany: Different ramifications—same effect? *Journal of Vocational Behavior*, 60, 178–198.
- Sekretariat der Ständigen Konferenz der Kulturminder der Länder in der Bundesrepublik Deutschland. (2006). *Übergang von der Grundschule in Schulen des Sekundarbereichs I. Informationsgrundlage des Sekretariats der Kulturminderkonferenz*. Stand: März 2006. Retrieved March 16, 2009 from http://www.kmk.org/fileadmin/veroeffentlichungen_beschlusse/2006/2006_03_00_Uebergang_Grundschr_SekI_01.pdf.
- Thiel, O. (2008). *Modellierung der Bildungsgangempfehlung in Berlin: Wie Lehrkräfte zu ihrem Urteil kommen*. Saarbrücken, Germany: VDM-Verlag.
- Tiedemann, J., & Billmann-Mahecha, E. (2007). Zum Einfluss von Migration und Schulklassenzugehörigkeit auf die Übergangsempfehlung für die Sekundarstufe I. [The influence of ethnic criteria and frame of reference effects on teachers' recommendations in regard to transition from primary to secondary education]. *Zeitschrift für Erziehungswissenschaft*, 10(1), 108–120.
- Trautwein, U., & Baeriswyl, F. (2007). Wenn leistungsstarke Klassenkameraden ein Nachteil sind Referenzgruppeneffekte bei Übertrittsentscheidungen [When high-achieving classmates put students at a disadvantage: Reference group effects at the transition to secondary schooling]. *Zeitschrift für Pädagogische Psychologie*, 21(2), 119–133.
- Treutlein, A., Roos, J., & Schöler, H. (2008). Einfluss des Leistungsniveaus einer Schulklassen auf die Benotung am Ende des 3. Schuljahres. *Revue suisse des sciences de l'éducation*, 30(3), 579–593.
- Updegraff, K., Eccles, J., Barber, B., & O'Brien, K. (1996). Course enrollment as self-regulatory behavior: Who takes optional high school math courses? *Learning and Individual Differences*, 8(3), 239–259.
- Willet, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, 15, 345–422.

Daniel H. Caro is a doctoral candidate at the *Freie Universität Berlin* and a research analyst at the International Association for the Evaluation of Educational Achievement (IEA)-Data Processing and Research Center (DPC). He holds a BA in Economics and completed a Master's Degree in Interdisciplinary Studies at the University of New Brunswick (UNB) through the Canadian Research Institute for Social Policy (CRISP). His research examines the influence of family background on educational and labor force outcomes over the life course.

Jenny Lenkeit is a doctoral student and research assistant at the University of Hamburg, Faculty of Education, Psychology and Human Movement, Department of Education. She holds a Bachelor's degree in Education and a Master's degree in Empirical Educational Research and Educational Expertise from the Humboldt University of Berlin. Her research examines fairness of comparison and characteristics of effective schools.

Rainer Lehmann is professor of Educational Research Methods at Humboldt University of Berlin. He has contributed to several internationally comparative studies, concerned with written composition, reading, mathematics and science, adult literacy and civic education. He has also conducted a number of regional achievement studies, some of which had longitudinal designs. In recent years, the focus of his attention has moved beyond general education to include questions of vocational education and particular programs in special education.

Knut Schwippert holds the chair of "International Comparative Education and System Monitoring" at the University of Hamburg, Faculty of Education, Psychology and Human Movement. During PIRLS 2001 (Progress in International Reading Literacy Study), he was jointly responsible for coordinating Germany's involvement in PIRLS conducted by the International Association for the Development of Educational Achievement (IEA). He is a member of the PIRLS 2011 national steering committee for Germany. His main areas of research interest are methods in international large scale assessments, heterogeneity in schools, and organizational development.

CHAPTER 3 STUDY 2

Jenny Lenkeit (2012). Effectiveness measures for cross-sectional studies: A comparison of value-added models and contextualised attainment models. *School Effectiveness and School Improvement*. Available at: <http://www.tandfonline.com/doi/abs/10.1080/09243453.2012.680892>

Effectiveness measures for cross-sectional studies: a comparison of value-added models and contextualised attainment models

Jenny Lenkeit*

Department of Education, University of Hamburg, Hamburg, Germany

(Received 21 December 2010; final version received 19 January 2012)

Educational effectiveness research often appeals to *value-added models (VAM)* to gauge the impact of schooling on student learning *net* of the effect of student background variables. A huge amount of cross-sectional studies do not, however, meet VAM's requirement for longitudinal data. *Contextualised attainment models (CAM)* measure the influence of schools on student outcomes controlling for family background characteristics in cross-sectional studies. It is argued that the latter are adequate substitutes for student prior attainment. Drawing on data from a 3-point longitudinal study in the city of Berlin, Germany ($n = 3,074$), reading and mathematics achievement of primary students are investigated to assess effectiveness measures of schools. Estimates are compared for a 3-level growth curve analysis (VAM), a hierarchical linear model controlling for background characteristics (CAM), and one additionally controlling for prior achievement scores (prior attainment model). The article contributes to the enhancement of a feedback culture for cross-sectional study results.

Keywords: educational effectiveness; value-added models; contextualised attainment models; ranking

Introduction

Educational Effectiveness Research (EER) is primarily concerned with the effects classroom and school practices and policies have on student achievement (Creemers & Kyriakides, 2008; Postlethwaite & Ross, 1992; Reynolds, Teddlie, & Townsend, 2000; Scheerens, 1997).

EER is increasingly utilised to judge schools, principals, and teachers and sometimes confront them with high-stakes personnel decisions based on their results of effectiveness assessments. Many theoretical and empirical works, however, have repeatedly emphasised the weaknesses of the research field. Essentially, those weaknesses concern conceptual and methodological aspects such as the assumption that the complexity of what constitutes a teacher's or school's quality can be represented in a single score (Ballou, 2002; Kelly & Monczunski, 2007). Furthermore, claims that assessments of teacher and school effectiveness are usually norm referenced (Kupermintz, 2003; Sammons & Luyten, 2009) or that tests do not comprehensively represent the domain they claim to test (Koretz, 2000) seriously

*Email: jenny.lenkeit@uni-hamburg.de

question the use of effectiveness measures for accountability systems. This article, however, rather focuses on the scope of enhancing the methodological developments of the research field. The prerequisite to implement this kind of research is to find an appropriate way to operationalise the effectiveness of a class or school, that is, to define the measurement of the output.

Developments in EER have reached the common accord that for a “fair comparison” of a school’s performance, studies of effectiveness have to control for differences in student intake in order to hold educators accountable only for effects that are in their sphere of influence (Ballou, Sanders, & Wright, 2004; Martineau, 2006; Organisation for Economic Co-operation and Development [OECD], 2008; Thomas, 1998). Here, a measure of student achievement at an earlier measurement point is considered the best indicator to assess the effectiveness of a school, a teacher, or an instruction method. At least theoretically, this approach is the most sensible if we want to measure determinants of student achievement that are subject to educational policies. In practice, however, researchers are often confronted with cross-sectional studies that naturally lack measures of prior attainment and therefore have to fall back on other student characteristics to assess educational effectiveness of an institutional unit. In particular, this is the case for most international assessment studies of academic achievement. The aim of this article is to investigate whether effectiveness measures that are obtained by controlling for students’ background characteristics instead of prior attainment are suitable substitutes if cross-sectional data sets are the only ones available.

Conceptual framework

Theoretical background

What today comprehensively is called Educational Effectiveness Research (EER) captures a range of research areas from different waves and strands (Creemers & Kyriakides, 2008; Reynolds et al., 2000). It represents an integration of the fields of school effectiveness (school organisation and educational policy) and research aimed at the classroom level (teacher behaviour, instruction methods, and curriculum analyses). With a proceeding awareness of contextual impacts on learning processes, approaches were elaborated that viewed effectiveness as a multilevel phenomenon integrating cross-level relationships in the theoretical models. This development promoted the blending of the former approaches (Creemers & Kyriakides, 2008; Scheerens, 1997) to what has commonly been called *educational effectiveness*.

From a researcher’s point of view, the development of multilevel effectiveness models is clearly a conceptual and statistical improvement in effectiveness research. However, the interdependence of different levels poses further problems regarding accountability. Previously, teacher effectiveness studies were already criticized for their inability to disentangle teacher effects from student characteristics and class composition, thus making those studies a precarious endeavour, even more so in high-stakes accountability systems (Ballou et al., 2004; Kupermintz, 2003; McCaffrey, Lockwood, Mariano, & Setodji, 2005) as it became obvious that the quality of a teacher’s work does not solely depend on his or her individual professional skills. Consequently, the more we know about the interdependence of different levels in the learning and teaching process, the more difficult it seems for educational stakeholders to hold single teachers in, or principals of, a school accountable for their students’ attainment. Accountability could thus only be

claimed from the unit of the entire school, thereby in turn neglecting differences in teaching staff (and also departments) within schools as well as characteristics of the community the school is located in.

Associated with questions concerning accountability is the definition of the educational outcome. The distinction between academic and non-academic outcomes is primary for this definition. Most evaluations of effectiveness relate to cognitive domains of schooling (De Maeyer, Van den Bergh, Rymenans, Van Petegem, & Rijlaarsdam, 2010; Postlethwaite & Ross, 1992), arguing that they best represent the school's societal assignment and the areas in which schools can make a recognisable difference (Opdenakker & Van Damme, 2000). Others criticise to loosen this focus and include outcomes such as equity (Choi & Seltzer, 2003; Sammons, 2007), metacognitive skills and non-cognitive outcomes when examining the effects of teachers and schools on students (Campbell, Kyriakides, Muijs, & Robinson, 2003) as an indicator of the quality of a school's comprehensive work. Usually, however, the influence of classes and schools tends to be higher on cognitive domains than on non-cognitive domains (Opdenakker & Van Damme, 2000), a fact that mirrors that the school's responsibility in most countries across the world lies in the enhancement of knowledge rather than, for example, of well being.

The choices made regarding the educational outcome mirror "the conviction to the objective of education" (De Maeyer et al., 2010, p. 82), they determine whether or not a school is claimed to be effective (Coe & Fitz-Gibbon, 1998), and are of substantial theoretical and methodological importance. Unfortunately though, the judgement of teachers' and schools' work seems to be increasingly principled by political rather than pedagogical convictions.

Concepts and measures of effectiveness

In addition to the distinction of academic and non-academic domains, effectiveness studies differ with regard to the theoretical and empirical models they apply. As educational processes take place across levels of students, classes, and schools, multilevel modelling is indispensable although earlier models made use of single-level regression analysis and inferior ways to account for clustering (Aitkin & Longford, 1986; Bryk & Raudenbush, 1992; Hill & Rowe, 1996; Sammons & Luyten, 2009).

In essence, approaches to measure effectiveness can be differentiated by their conviction of the nature of student intake, on the one hand, and the respective modelling techniques, on the other hand. The models differ with regard to the number of measurement points, whether they additionally control for student characteristics, and their underlying theoretical assumptions.

Models for cross-sectional data designs control for student characteristics. Referring to the OECD publication (OECD, 2008) these type of models are entitled contextualised attainment models (CAM) in the following. Most often, family background measures such as social status indicators are included in these models. There are, however, a range of other variables that distinguish students from each other and that are related to achievement, for example, cognitive ability, student educational expectations, and motivation (Choi & Seltzer, 2003; Fortier, Vallerand, & Guay, 1995). The choice of control variables establishes a specific understanding of effectiveness, which needs to be theoretically questioned (Coe & Fitz-Gibbon, 1998).

Researchers that fall back on data designs with two and more measurement points consider student intake by means of controlling for prior scores of the output measure. The respective models are usually subsumed under the term value-added models (VAM).¹ A research design particularly employed in the assessment of school effectiveness includes cross-sectional studies that refer to prior attainment assessed multiple years previously. Concerned with the stability of school effects on levels of attainment, these studies examine different cohorts of students in a particular grade in different years (Luyten, 1994).² In those designs, there is a need to control for differences between the student cohorts with respect to individual characteristics (Kelly & Monczunski, 2007) and, as Willms and Raudenbush (1989) point out, changes at the level of community, such as an increase in local unemployment that may change student attitudes towards school and hence their academic attainment. Additionally, different cohorts of students are likely to have been taught by different teachers, making it difficult to attribute the effects to one teacher only. This particular research design will not be further discussed in detail since the database and the research question of the article requires a focus on students of the same cohort.

Value-added models consider the growth of student achievement as the most appropriate criterion to assess effectiveness (Teddle, Reynolds, & Sammons, 2000). It is argued that cross-sectional designs do not reflect the fact that learning itself is a process (Willet, 1988), that educational outcomes are a consequence of this process, that schools are changing, and that their respective effects are believed to be cumulative (Kennedy & Mandeville, 2000). Consequently, prior attainment is the most important and accurate factor that affects subsequent achievement, and rates of change are considered un-confounded with predictors of achievement such as socioeconomic and migration background of a student (Andrejko, 2004; Ballou et al., 2004; Zvoch & Stevens, 2008). A decisive advantage of growth curve modelling is, moreover, that two parameters of substance, achievement status and growth, are estimated for quantification of effectiveness (Stevens, 2005).

Nevertheless, there are discussions that question this unimpeachable standing of the analyses of attainment gains in effectiveness research. Researchers (Sammons, 1996; Teddle, Reynolds, et al., 2000) caution against the use of prior achievement data that have been surveyed proximal to the assessment within the same school, arguing that actual school effects are thus reduced. Consequently, as assessments in primary schools offer limited opportunities for prior achievement indicators that have been collected prior to the school entry, Teddle, Stringfield, and Reynolds, (2000) argue that use of background characteristics is more appropriate for effectiveness studies in primary school grades. Further disadvantages are that construct heterogeneity of the measured outcome, the choice of the metric, as well as intense mobility of students across schools (especially in urban areas) are threats to unbiased measures of effectiveness in longitudinal data designs (Doran & Cohen, 2005; Goldschmidt, Choi, Martinez, & Novak, 2010; Schmidt, Houang, & McKnight, 2005). Moreover, reliabilities of growth rates are normally weaker than those of status indicators (Stevens, 2005). Additionally, based on social inequality theories (Boudon, 1974; Bourdieu, 1983), a multitude of research findings has provided evidence of the systematic relationship between family background variables and academic achievement (e.g., Baumert, Stanat, & Watermann, 2006; Maaz, Baumert, & Trautwein, 2009; Mullis, Martin, Kennedy, & Foy, 2007; OECD, 2007), according to which students from disadvantaged social backgrounds are more

likely to attain lower academic achievement scores. Because of this strong and systematic association of student's socioeconomic background variables and their academic achievement, it is reasonable to assume that the former function as adequate controls for the prediction of academic achievement.

Value-added models additionally differ with regard to the measurement points they rely on to estimate attainment gains. Although many studies are based on measures obtained from two measurement points only (e.g., Sammons, Nuttall, & Cuttance, 1993; Strand, 2010; Tekwe et al., 2004; Thomas & Mortimore, 1996), the notion is that individual growth can only be obtained from three or more measurement points (Bryk & Raudenbush, 1992; Rogosa, 1995; Singer & Willet, 2003; Willet, 1988). Independent of the number of measurement points, value-added models further disagree on the inclusion of student background variables. Opponents argue that these variables are of no additional value once prior attainment has been taken into account (Sanders, 2000; Thomas & Mortimore, 1996), that they would excuse schools from the responsibility of their effects (Tekwe et al., 2004), or that they might "overcorrect" the statistical models (McCaffrey et al., 2005). Guo (1998) argued, however, that long-term exposure to disadvantaged social and economic situations may have cumulative effects on students' cognitive outcomes as they advance in their educational careers, and empirical findings on the influence of individual background characteristics on attainment gains maintain this argument (Alexander, Entwisle, & Olsen, 2001; Caro & Lehmann, 2009; Cortina, Carlisle, & Zeng, 2008).

Again, both arguments are subject to convictions about what educators shall be held accountable for and similarly apply to the discussion on the inclusion of compositional variables in evaluations of school effectiveness. As Raudenbush (1995) points out, individual characteristics and those of a shared social context "are correlated and they interact to shape development" (p. 169). Schools and classes with an advantaged social composition of students are also more likely to rely on higher average student abilities and motivational attitudes, higher parental support, and fewer disciplinary problems (Willms, 2000). These favourable contexts hence promote the development of academic achievement. Moreover, research has repeatedly given evidence that compositional variables affect academic achievement over and above individual background variables (Dreeben & Barr, 1988; Hattie, 2002; Nachtigall, Kröhne, Enders, & Steyer, 2008). On the one hand, it can be argued that these effects originate from the composition of the student body and can therefore not be influenced by teachers and school principals. They would thus have to be "controlled for" when assessing educational effectiveness. On the other hand, it might be argued that having controlled for selective school enrolment on the student level, compositions of the same variables play indeed a conceptually different role in the models, as they are likely to direct school policy and practices (Raudenbush, 2004; Raudenbush & Willms, 1995) and have also been found to interact with effects of school and classroom processes on performance (Opdenakker & Damme, 2007; Stevens, 2005). The above discussion highlights that the differences in the applied models represent differences in the theoretical assumptions of adequate controls in obtaining measures of effectiveness. Therewith connected are convictions regarding responsibilities associated with educators.

More recent methodological developments in EER are moreover concerned with reliable measures for class and school process variables that account for cluster effects of the individual perceptions of these variables within the respective classes

and school (D’Haenens, Van Damme, & Onghena, 2010), the adjustment of measurement error in covariates of value-added models (Battauz, Bellio, & Gori, 2011), the properties of status and growth metrics to examine the performance of schools for accountability purposes (Goldschmidt et al., 2010), and the assessment of school effects with an added year of schooling (regression discontinuity design) (Heck & Moriyama, 2010; Luyten, Tymms, & Jones, 2009).

Status models and prior attainment models have previously been compared regarding their amount of explained variance (e.g., Sammons et al., 1993). One of the few studies comparing status and growth models has been conducted by Zvoch and Stevens (2008). Their evaluation of appropriate measures for use in an accountability system did not, however, surpass a comparison of results from inferential analyses with descriptive statistics of student achievement status and growth. Thus, studies investigating whether effectiveness measures obtained from status and growth models differ with regard to their conceptual idea of the quality of a school’s work are, to my knowledge, lacking in the research literature on educational effectiveness.

Research questions

Most international studies such as the Programme for International Student Assessment (PISA) (OECD, 2007) and the Progress in International Reading Literacy Study (PIRLS) (Mullis et al., 2007) are implemented with cross-sectional data designs. Following the conviction that effectiveness can only be adequately measured by assessing attainment gains, these cross-sectional studies could not be used for the evaluation of effectiveness. The potential of these designs to obtain adjusted measures of school and respectively country performance by controlling for student intake factors is hardly ever utilised. Instead, results are presented in the format of league tables of unadjusted raw scores and regularly provoke heavy reactions in the public sphere (see Goldstein & Spiegelhalter, 1996, for further discussion).

It is argued that student background characteristics function as good substitutes for prior attainment, when only cross-sectional data are available to predict academic achievement levels. On that account, effectiveness measures obtained from the analyses of growth models, prior attainment models, and contextualised attainment models are comparatively analysed. Consequently, the article at hand poses the following questions: Do the different assumptions of effectiveness underlying the models also result in different effectiveness estimates? Are differences in effectiveness estimates of practical relevance, so that, for example, a school shows to be effective with regard to growth but ineffective with regard to status? It thus aims to contribute to the methodological developments in EER.

Analytical strategy

Data

Data stem from the longitudinal student achievement study ELEMENT (*Erhebung zum Lese- und Mathematikverständnis – Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin* [Assessment of reading and mathematics literacy – Developments in the Grades 4 to 6 in Berlin]) that was carried out from 2003 to 2005 (Lehmann & Lenkeit, 2008). Primary education in Berlin is split into groups of students who are educated in primary schools until the end of Grade 6 and those who prematurely

enter an academic track (Grundständiges Gymnasium) after Grade 4. The majority of students in Berlin remain in primary schools until the end of Grade 6, and only less than approximately 10% of the age cohort enters an early academic track. This group is highly selective with regard to academic achievement and family background characteristics.

As documented by the project director (Lehmann & Nikolova, 2005), the study is based on a representative sample of primary schools in Berlin in the school year 2002/2003 (Probability Proportional to Size [PPS] sample on class level), as well as a census of schools that offer an academic track for students after Grade 4. The study tested an overall sample of 4,926 students at the end of Grades 4, 5, and 6 in the domains of reading comprehension and mathematics. Students who migrated between schools within those 3 years and those in the early academic track³ were excluded from the analysis. Note that the analytic sample has thus been creamed off of a highly selective group of students as the latter represent approximately 95% of the excluded sample (37.6% of the overall sample or 1,852 students). The analytic sample includes 3,074 students in 71 schools.

However, in Berlin the demand for places in an early academic track exceeds its actual supply so that only few individual students of a primary school class prematurely change the school to enter an academic track. Thus, classes and schools remain intact with sufficient number of students. Nevertheless, heterogeneity of achievement levels in classes may be affected by the loss of even a small number of high-achieving students. This puts constraints on the representativeness of fourth-grade students in Berlin. Moreover, measures of social background on school level (as described in the Measures section) may be systematically underestimated. The underestimation will, however, be even across the schools. These measures are, moreover, no object to explicit model comparison of schools in the analysis.

Missing data on dependent and independent variables was imputed by using multiple imputation (Schafer & Olsen, 1998) so that relevant data were available for all three measurement points. The model was estimated in a one-step procedure.⁴ Table A1 (Appendix 1) shows the descriptive measures for the analytical and the excluded sample.

Measures

Reading and mathematics literacy: Test items were scaled using item response theory (IRT) with overlapping items vertically equated to create a longitudinally comparable scale suited to assess student growth. The reading scale ($\alpha = .84$) consists of 26 items in Grade 4, 32 items in Grade 5, and 24 items in Grade 6 with 17 and 13 overlapping items, respectively. The mathematics scale ($\alpha = .92$) consists of 24 items in Grade 4, 40 items in Grade 5, and 54 items in Grade 6 with 19 and 39 overlapping items, respectively. Items were scaled to have an average score of 100 and a standard deviation of 15 for the first measurement point (Lehmann & Lenkeit, 2008). The included measures are estimates of student achievement status in Grade 6 as well as student achievement growth which represent the dependent variables for the estimated models.

Sex is a dichotomous variable that distinguishes boys (0) from girls (1).

Number of books at home (books) was reported by parents and categorized in four values (1 = up to 25 books, 2 = up to 100 books, 3 = up to 200 books, 4 = more than 200 books).

Highest parent education (hiedu): Both mother and father reported their highest level of schooling; the variable presents the highest of both statements and is classified into (1) no qualification/special education, (2) lowest track of secondary education, (3) intermediate track of secondary education, (4) admission level for technical college, and (5) admission level for university.

Highest socioeconomic index (hisei) represents occupational data for both parents. It was obtained with open-ended questions, and responses were classified in accordance with Erikson, Goldthorpe, and Portocarero (1979) and then mapped to the International Socioeconomic Index of Occupational Status (ISEI; Ganzeboom, De Graaf, & Treiman, 1992) by Lehmann and Nikolova (2005). HISEI corresponds to the higher ISEI score of either parent. Scores range from 16 to 85, where higher values indicate a higher occupational status.

The independent variable at the aggregate level is the *school mean of highest socioeconomic index (school hisei)*.

No information of school practices is considered. Rather, the respective effects are conceived to be an unobservable part of the school effects that remains after controlling for effects of student background and contexts.

Analytical strategy

Models are estimated by means of hierarchical linear modelling (HLM) accounting for the multilevel structure of the data (Bryk & Raudenbush, 1992). To avoid multicollinearity, only one indicator of student intake is accounted for at the school level. Preceding analyses have shown that *hisei* has the strongest impact in explaining differences between schools. Covariates at student and school level are grand mean centred to control for them as hypotheses relate to contextual effects (Bryk & Raudenbush, 1992; Enders & Tofighi, 2007). No weights were used as statements referring to the population of the sample are not the cause of the investigation and value-added analyses can always only refer to the sample studied.

The contextualised attainment model (CAM) is represented as a two-level hierarchical linear model that controls for influences of background variables at student level (sex, number of books at home, highest parent education, and highest socioeconomic status) and includes the school mean of highest socioeconomic status at the school level.

The prior attainment model (PAM) is specified accordingly while simultaneously integrating prior achievement scores in Grade 4 (*prior4*) on student as well as its aggregate on school level.

The value-added model (VAM) is specified as a three-level linear growth model that models achievement status in Grade 6 as well as growth from Grades 4 to 6.⁵ Interindividual differences in achievement levels and achievement growth rates are estimated controlling for sex, number of books at home, highest parent education, and highest socioeconomic status. At the school level, the school mean of highest socioeconomic status is included to estimate differences in average achievement status and growth between schools.

Effectiveness measures are delineated by the school residuals of the conditional models CAM, PAM, and VAM that capture the deviation of a school from its expected outcome when the conditional model is applied, that is, differences between the expected and adjusted scores in achievement status or growth. Additionally,

differences between unadjusted raw scores and adjusted scores are brought in order to compare practical relevance of accorded performance scores.

Results

Table 1 gives an overview of the estimates for the two status models (CAM and PAM) and the growth model (VAM). Students in Grade 6 achieve an average score of 110 in reading achievement and have an average growth rate of 6.2 scores per grade. The status model yields an intra-class correlation (ICC) of 22.5%. The baseline statistics of the growth model indicate that the percentage of variation between schools is 27.8% for the status and 7.7% for the growth parameter. Correlation of status and growth in reading achievement is negative ($r = -.201$) indicating that, on average, students with high status scores in Grade 6 had lower growth rates. Ceiling effects represent a possible source of bias if negative correlations of status and growth are found, suggesting that the test is not sufficiently difficult to capture growth of the best performing students. If indeed ceiling effects existed, they would leave little room for improvement for high-achieving students and thus defeat the possibility to measure an added value (Koedel & Betts, 2008). The proportion of students at each grade level that achieve the maximum test score is negligibly small with 0.1%, 0.03%, and 0.1% for reading achievement in Grades 4 to 6 (0.3%, 0.1%, and 0.03% for mathematics achievement, respectively). Distributions for both domains and the three measurement points fit a normal distribution nicely. Hence, the phenomenon does not give clues as a potential source for the negative correlation found. Also, regression towards the mean as a

Table 1. Status (Grade 6) and growth model estimates for reading and mathematics achievement.

	reading achievement	mathematics achievement
fixed effects:	coefficient	coefficient
<i>status model</i>		
mean status in Grade 6, β_{00}	109.99***	114.98***
<i>growth model</i>		
mean status in Grade 6, γ_{000}	110.38***	115.15***
mean growth rate, γ_{100}	6.19***	9.21***
random effects (as percentage of variation):	variance	variance
<i>status model</i>		
status in Grade 6, u_0	22.5%***	21.5%***
<i>growth model</i>		
status in Grade 6, u_{00}	27.8%***	23.9%***
growth rate, u_{10}	7.7%***	22.8%***
reliability		
<i>status model</i>		
mean status in Grade 6, β_{0j}	0.92	0.92
<i>growth model</i>		
mean status in Grade 6, β_{0ij}	0.72	0.81
mean growth rate, β_{1ij}	0.29	0.26

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

potential source of bias has been considered. It occurs when scores far from the mean in an initial observation tend to regress towards the mean in subsequent observations. It is considered that the extreme initial scores are exceptional and will most likely not occur in a second or third observation. If regression towards the mean exists, it gives the impression that the initially lowest achieving students have higher growth rates than the highest achieving students (and vice versa) and consequently biases the results because of unreliable growth estimates. This has relevant consequences for the interpretation of, for example, school or programme effects. Indeed, there is evidence that scores of the highest achieving students as well as the lowest achieving students in particular (upper and lower 10% for the first measurement point) do regress towards the mean in the following measurement points (for both reading and mathematics achievement). In the three measurement points growth model of the study, the applied Bayesian shrinkage, however, limits the effect of extreme unreliable growth estimates. It does so by pulling highly unlikely values closer towards the mean (Bryk & Raudenbush, 1992).

For mathematics, students on average achieve 115 scores in Grade 6 and have gained 9.2 scores per grade. The ICC for the status model indicates that differences in status are attributable to differences between schools by 21.5%. Percentage of variation that lies between schools according to the growth model is 23.9% for achievement status in Grade 6 and 22.8% for the growth parameter. The latter variation attributable to differences between schools is considerably higher than for reading achievement. In this regard, reading skills have to be considered much more home taught than mathematics skills. Consequently, status and growth in mathematics are also positively associated by $r = .474$.

Reliability of status estimates are 0.92 for both reading and mathematics. Estimates for individual growth parameter are small, that is, unreliable with 0.29 for reading and 0.26 for mathematics. There are two possible sources for this low reliability: Either the precision with which individual slopes are measured is low or the variability in growth rates between the tested students is small. Both effects, within-person precision and between-person heterogeneity, are confounded in the reliability measure of the growth parameter (Bryk & Raudenbush, 1992; Singer & Willet, 2003; Willet, 1988). Either source sets constraints on reliability of the results of the VAM analyses and their interpretation. Moreover, unreliable growth estimates additionally bias the relationship of achievement status in Grade 6 and estimates of achievement growth.

Detailed results of model estimates are presented in Tables A2 (reading achievement) and A3 (mathematics achievement) in Appendix 1. In both domains, the models explain different amounts of achievement variance between students and schools. The PAM explains the highest amounts of variance within and between schools. With regard to Grade 6 reading achievement status, covariates of the conditional PAM can account for 43.9% of variance within and 93.1% between schools. The same holds true for mathematics with 46.7% (within) and 85.6% (between) variance explained, respectively. Variables considered for growth rates in reading and mathematics explain considerably less variance in both domains; 2.6% within and 21.4% between schools in reading and 4.4% and 18.5% in mathematics, respectively.

Figures 1 to 3 illustrate the mean residuals for schools; that is, the deviations from their expected outcome in reading achievement accounting for the influence of the variables controlled in the conditional models.

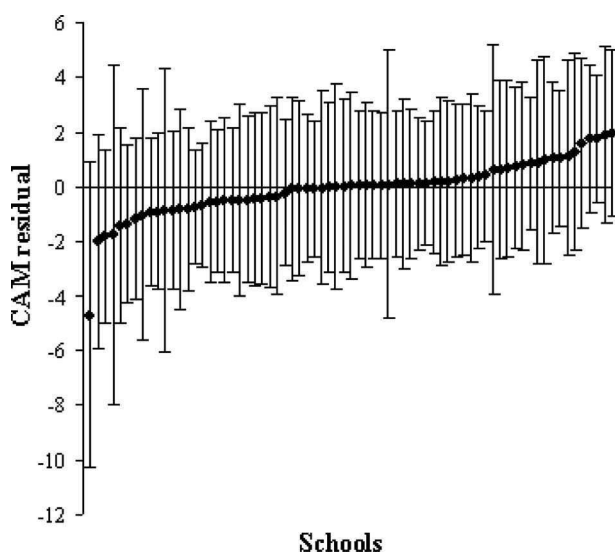


Figure 1. Mean residuals for reading achievement status in Grade 6 and confidence intervals by schools.

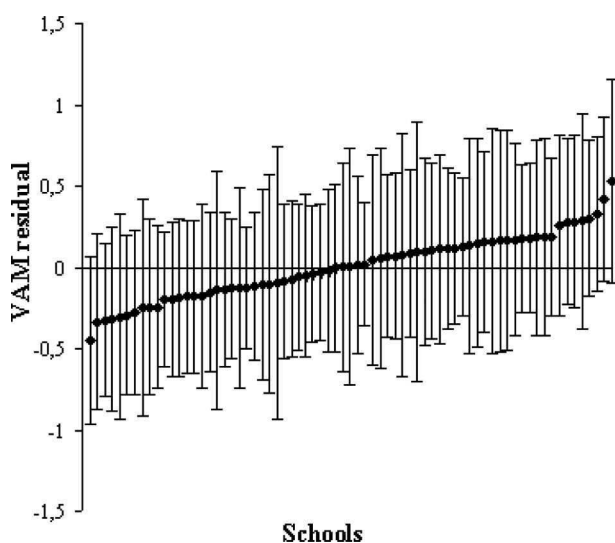


Figure 2. Mean residuals for reading achievement growth and confidence intervals by schools.

Regardless of whether the models control only for background variables or additionally for prior achievement scores or whether they consider growth rates, confidence intervals are strikingly wide, allowing no precise distinction between schools with regard to their effectiveness.⁶ The respective results for mathematics achievement (not depicted) show essentially the same picture.

The wide confidence intervals moreover indicate that within schools there are students who are outperforming their expected results and students who stay below

it. In this regard, analysis of schools' differential effectiveness for distinctive groups of students is sensible as accomplished by Sammons et al. (1993) and Strand (2010).

To investigate whether residuals obtained from the different model approaches represent the same idea of effectiveness, residuals are correlated (Table 2). Residuals were obtained from the school level controlling for the respective intake variables. Residuals from the PAM and VAM are almost in complete agreement ($r = .974$ for reading and $r = .933$ for mathematics), indicating that effectiveness measured by growth rates and by inclusion of prior achievement scores to predict status is nearly identical for both reading and mathematics achievement.

The association of residuals from PAM and CAM is very high for both subjects ($r = .700$ for reading and $r = .710$ for mathematics achievement). Effectiveness measures from the latter are however less strongly correlated with those from VAM ($r = .627$ for mathematics), considerably so for reading achievement ($r = .532$), indicating that effectiveness measures are captured quite differently. The comparison of CAM residuals with both, PAM and VAM residuals, shows that there are schools

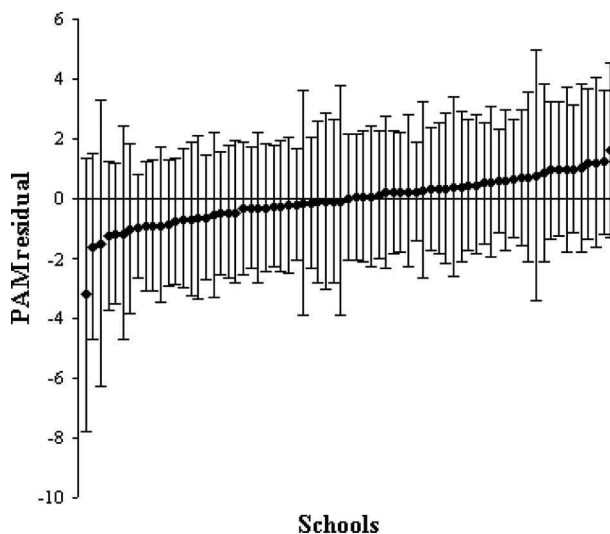


Figure 3. Mean residuals for reading achievement status and confidence intervals by schools.

Table 2. Correlations of residuals obtained from different models for reading and mathematics achievement.

	reading achievement			mathematics achievement		
	CAM	PAM	VAM	CAM	PAM	VAM
CAM		.700	.532		.710	.627
PAM	.700		.974	.710		.933
VAM	.532	.974		.627	.933	

Note: All correlations are significant $p < 0.001$.

that are effective with regard to achievement status but ineffective when measures of prior achievement scores are taken into account and vice versa (not depicted).

To illustrate the practical relevance of these adjustment procedures, schools have been ranked by their mean of student raw scores and subsequently by their mean adjusted scores. Comparisons of school rank position of unadjusted and adjusted estimates were made. Although ranking is rightfully criticised in assessing schools' work, value-added procedures are applied exactly with that aim – to bring schools in an order of their relative effectiveness. With practical relevance, it is thus also referred to the presentation of results in publications of international studies, which illustrate performance of countries in the form of league tables and regularly provoke heavy reactions in the public sphere.

In Figures 4 and 5, rank differences of unadjusted raw scores and adjusted scores obtained from the different models for reading achievement are depicted. Generally, differences are larger when prior achievement scores have been taken into account as in PAM or in VAM. Comparison of CAM and VAM reveal that in 31 (reading achievement) and 16 (mathematics achievement; not depicted) out of 71 cases schools are ranked into opposite direction.

In this regard, it is initially suggestive to assume that effectiveness is differentially captured in the two approaches. However, the weak reliability of the growth rate estimates has to be considered as a potential source of bias that yields unreliable effects on and of growth as well. It is thus that estimates obtained from VAM can only be observed with great uncertainty (Stevens, 2005). Therefore, rank differences

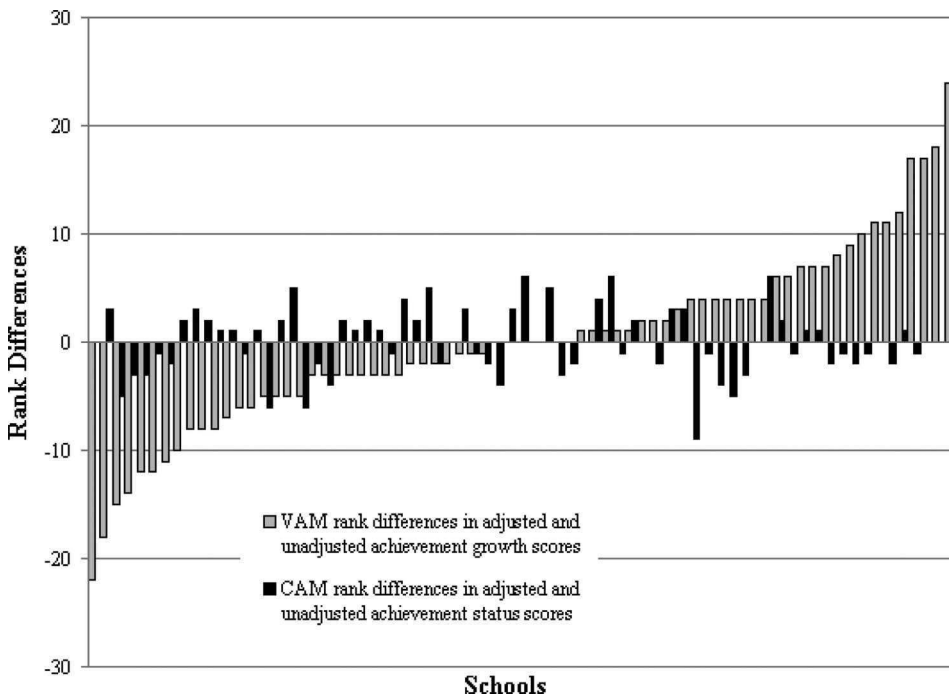


Figure 4. Differences in ranks of unadjusted raw scores and adjusted scores in reading achievement by models (VAM/CAM) and schools.

Note: 31 schools are ranked in the opposite direction.

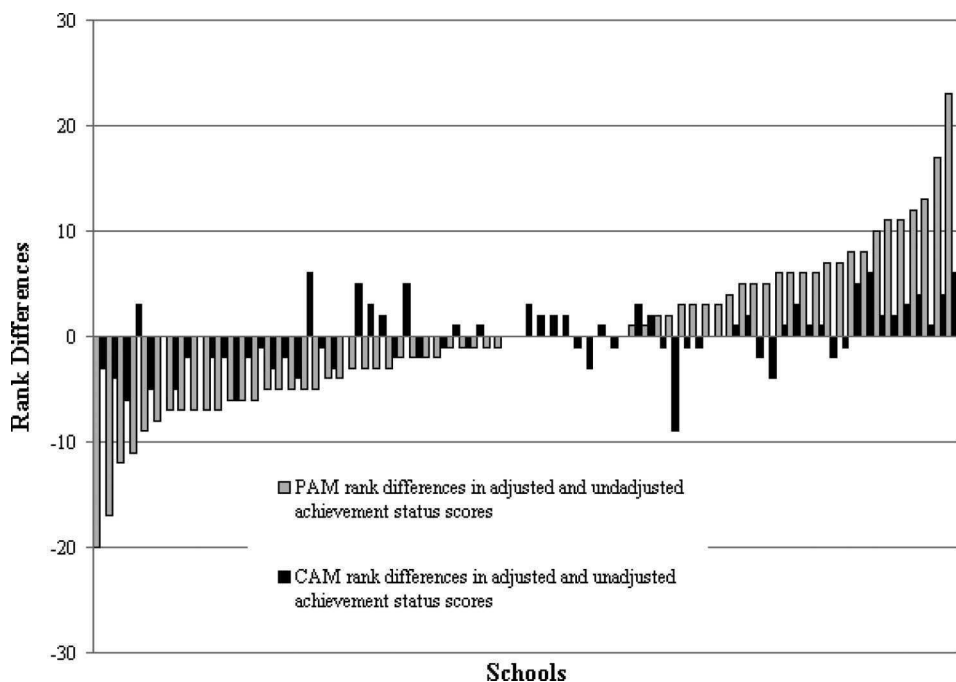


Figure 5. Differences in ranks of unadjusted raw scores and adjusted scores in reading achievement by models (PAM/CAM) and schools.

Note: 16 schools are ranked in the opposite direction.

between the more reliable CAM and PAM are considered. In the case of reading (Figure 5), considerably fewer schools are ranked into opposite direction (16) when comparing rank differences of those models and 17 for mathematics achievement (not depicted). Consequently, an order of schools by their effectiveness would be considerably different to ranking schools by raw scores. Nevertheless, schools would have different rank positions according to different adjustment models applied and to the relative change of positions of other schools in the sample.

The results overall indicate that the exclusion of prior achievement scores comprised lower amounts of explained variance and less strong shifts in ranks than for models that do include such measures. Although obtaining effectiveness measures from growth rates is conceptually sensible, statistically these measures are acquired with high uncertainty, at least with three measurement points only. A tangible alternative for the VAM is represented by PAM. Their residuals are highly correlated, but the latter yields more reliable results. Consequently and with regard to the research question, a comparison of the CAM and PAM is appropriate. Here, stronger associations of residuals have been found ($r = .700$ for reading and $r = .710$ for mathematics), and ranks have been shifted in opposite directions less often.

Correlation of effectiveness measures for the domain of reading and mathematics was conducted to investigate whether schools are equally effective or ineffective for both subjects. In fact, this can be negated for effectiveness measures obtained from PAM ($r = .496$) and VAM ($r = .449$). These results indicate that schools which are effective for one domain, do not necessarily have to be effective in another domain as

well; that is, they can be differentially effective. Estimates for the CAM model are higher with $r = .727$.

Discussion and conclusion

The analysis at hand has important practical implications for the research on educational effectiveness. It illustrated a comparison of different effectiveness models with regard to the actual consequences for the ranking of schools if those models were applied. It thus exemplified how cross-sectional data sets could be employed in educational effectiveness research. Moreover, the article contributes to the existing literature by applying a three-level growth model to assess the impact of schools on students' learning processes, an area which has been neglected in school effectiveness research.

Findings align with research on school effectiveness. Including prior achievement scores in the prediction of achievement status considerably contributes to explaining differences between students and schools in comparison to a model with family background characteristics only. Prior attainment thus has to be considered the most relevant of factors influencing achievement status, although the additional contribution of family background characteristics should not be disregarded. Nevertheless, effectiveness measures yielded by the growth model are accompanied with high uncertainty. The presumption that growth models of educational achievement yield reliable measures of effectiveness has, although theoretically and conceptually maintainable, been set methodological limits when only few measurement points are available. In research designs that yield reliable measures for both status and growth estimates, these models would hold more information for the evaluation of effectiveness, yielding both status and growth estimates. For the study presented here, however, I argue that although reliability of achievement status within the growth model is sufficiently high, uncertain information about a school's effectiveness based on growth estimates obtained from this model may be misleading and should not be the foundation for decisions, especially high-stakes ones. This considered, the comparison of the two status models (CAM and PAM) led to more trustworthy results. Correlation estimates showed high associations of residuals from both models, indicating a strong relationship of the obtained effectiveness measures (Hill & Rowe, 1996). Although shifts in ranks for unadjusted and adjusted school mean scores for both models were not identical regarding their extent, for the most part classification of schools to "underperforming" and "overperforming" institutions was coherent. Conclusively, the application of CAM for illustrating effectiveness of organisational units (classes, schools, tracks, or systems) is advocated when only cross-sectional data are available. Although comparison of the different models applied overall does not yield definite results in favour of the outlined research question, findings legitimate the adjustment of achievement scores in cross-sectional studies to obtain measures of effectiveness. Accounting for differences of individual students and the student body of a school will yield a more reliable evaluation of a school's work than applying no adjustments at all.

Furthermore, the article responds to the imperative that a school's work cannot rightfully be judged by assessing its effectiveness in one subject only (Hill & Rowe, 1996; Sammons et al., 1993). Hence, effectiveness was investigated by two of the most basic and societal relevant skills, reading achievement and mathematics. Nevertheless, a comprehensive evaluation of a school's work should by no means be

focused solely on cognitive outcomes but rather be additionally considerate of non-cognitive outcomes.

Analyses showed that 22.8% of growth parameter variation in mathematics and only 7.7% in reading is attributable to the school at all. This is a sobering result. However, only approximately a fifth of this potential radius of operation for schools is occupied by students' socioeconomic and sociodemographic characteristics. The remaining four fifths therefore lie within the range of action and responsibility of schools.

However, there are limitations to this investigation. One is that the analytic sample has been creamed off of 37.6% of the original sample mostly for a selective group of students who prematurely attend an academic track after Grade 4. Results of the model comparisons are thus afflicted by the limited representativeness of students under investigation within schools. Another limitation is that school and class level are not differentiated as no representative samples of classes within all schools have been at disposition in the analytic sample. Hence, it cannot be precluded that school effects can also be class effects. Furthermore, although student intake variables in the models have been chosen in accordance with established theoretical and empirical findings, the possibility cannot be ruled out that other important covariates of educational achievement have been missed out and would thus bias the results.

Reliability of the individual growth estimates is poor and reduces the accuracy of the results. Despite the limit that Bayesian shrinkage sets on unreliable growth parameters, regression towards the mean has to be considered an important impact on the reliability of the growth estimates and of the respective results. Although the three measurement points of the study provided greater intra-individual variability than, for example, two measurement points, Raudenbush (1995) states the gain of an additional data point usually has substantial positive effects on the precision of the within-person estimates and can thus benefit the reliability of the growth parameter. Thus, reliance of growth estimates in longitudinal designs for obtaining effectiveness measures depends essentially on their reliability.

Additionally, overlapping confidence intervals demonstrate a high uncertainty of the investigation of school effectiveness. Therefore, the superiority of adjustments is questionable towards the context of the analyses (Goldstein, 1991) as obviously none of the observed schools achieves salient results once student intake has been taken into account. As mentioned before, in-depth analyses of differential school effectiveness would be necessary to make more funded claims on a school's impact regarding different groups of students. Moreover, the estimated effects have to be understood as associations and by no means interpreted as causal effects as is often implied. Consequently, interpretation of effectiveness measures has to be oriented towards this constraint. For a more detailed discussion on causality in value-added assessments, see Rubin, Stuart, and Zanutto (2004) and Schatz, VonSecker, and Alban (2005). Furthermore, it has to be kept in mind that a school's effectiveness is always based on its relative position in comparison with the schools included in the analysed data of the study (OECD, 2008).

To apply the outlined adjustment procedures in the presentation of country league tables of international large-scale assessments needs careful consideration of the challenges international comparisons bring about. Socioeconomic background variables, for example, constitute as highly efficient predictors of educational achievement in Germany; however, associations are often less strong in other

countries (Mullis et al., 2007; OECD, 2007). Comparisons thus might be most sensitive within groups of similar economic, societal, and cultural characteristics. Nevertheless, adjusting achievement scores in international studies of student performance to account for different student compositions within and between the various countries would certainly contribute to lessen negative reactions towards the established illustration of respective results.

Acknowledgements

The author thanks the reviewers for their helpful comments as well as the Senatsverwaltung für Bildung, Wissenschaft und Forschung, Berlin, and the Research Data Centre (Forschungsdatenzentrum, FDZ) at the Institute for Educational Progress (Institut zur Qualitätsentwicklung im Bildungswesen, IQB) in Berlin for providing the ELEMENT data.

Notes

1. Following the idea that school contributes to student achievement, the term value-added is in some studies also used for analysis of achievement status when controlling for student background characteristics only (Stevens, 2005).
2. This research design is also found in international studies of student assessments such as PIRLS, PISA, or TIMSS, which examine the achievement of students in a particular grade in regular cycles.
3. Although data are available for three measurement points for students who prematurely entered an academic track, they were excluded from the analytic sample as they were first tested at the beginning of Grade 5 (rather than the end of Grade 4) to gather their initial achievement levels. A comparison of initial achievement levels and succeeding growth estimates with students of the primary school sample would be biased considering the research findings on summer learning gaps (e.g., Alexander, Entwisle, & Olsen, 2007).
4. All analyses are conducted with imputed data provided by the Research Data Centre (Forschungsdatenzentrum, FDZ) at the Institute for Educational Progress (Institut zur Qualitätsentwicklung im Bildungswesen, IQB).
5. The grade indicator has been recoded so that -2 represents achievement score in Grade 4, -1 in Grade 5, and a value of 0 represents achievement score in Grade 6.
6. For comparison of Figures 1 to 3 note, however, that deviations are displayed in accordance to actual scores and that figures of status and growth have different scales.

Notes on contributor

Jenny Lenkeit is a doctoral student and research assistant at the University of Hamburg, Department of Education. She holds a Bachelor's degree in Education and a Master's degree in International Educational Research and Expertise from the Humboldt-University of Berlin. Her research examines status and growth as predictors of educational outcomes and educational effectiveness.

References

- Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General)*, 149, 1–43.
- Alexander, K.L., Entwisle, D.R., & Olsen, L.S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis*, 23, 171–191.
- Alexander, K.L., Entwisle, D.R., & Olsen, L.S. (2007). Lasting consequences of the summer learning gap. *American Sociological Review*, 72, 167–180.
- Andrejko, L. (2004). Value-added assessment: A view from a practitioner. *Journal of Educational and Behavioral Statistics*, 29, 7–9.
- Ballou, D. (2002). Seizing up test scores. *Education Next*, 2(2), 10–15.
- Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal for Educational and Behavioral Statistics*, 29, 37–65.

- Battaui, M., Bellio, R., & Gori, E. (2011). Covariate measurement error adjustment for multilevel models with application to educational data. *Journal of Educational and Behavioral Statistics*, 36, 283–306.
- Baumert, J., Stanat, P., & Watermann, R. (Eds.). (2006). *Herkunftsbedingte Disparitäten im Bildungswesen. Vertiefende Analysen im Rahmen von PISA 2000* [Class-related disparities in the educational system. In-depth analyses in the scope of PISA 2000]. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.
- Boudon, R. (1974). Educational growth and economic equality. *Quality and Quantity*, 8, 1–10.
- Bourdieu, P. (1983). Ökonomisches Kapital, kulturelles Kapital, soziales Kapital [Economic capital, cultural capital, social capital]. In R. Kreckel (Ed.), *Soziale Ungleichheiten* (pp. 183–198). Göttingen, Germany: Schwartz.
- Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models*. London, UK: Sage.
- 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, 347–362.
- 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, 407–431.
- Choi, K., & Seltzer, M. (2003). *Addressing questions concerning equity in longitudinal studies of school effectiveness and accountability: Modelling heterogeneity in relationships between initial status and rates of change*. Los Angeles, CA: Center for the Study of Evaluation, University of California.
- Coe, R., & Fitz-Gibbon, C.T. (1998). School effectiveness research: Criticisms and recommendations. *Oxford Review of Education*, 24, 421–438.
- 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, 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, UK: Routledge.
- De Maeyer, S., Van den Bergh, H., Rymenans, R., Van Petegem, P., & Rijlaarsdam, G. (2010). Effectiveness criteria in school effectiveness studies: Further research on the choice for a multivariate model. *Educational Research Review*, 5, 81–96.
- D'Haenens, E., Van Damme, J., & Onghena, P. (2010). Multilevel exploratory factor analysis: Illustrating its surplus value in educational effectiveness research. *School Effectiveness and School Improvement*, 21, 209–235.
- Doran, H.C., & Cohen, J. (2005). The confounding effects of linking bias on gains estimated from value-added models. In R. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 80–110). Maple Grove, MN: JAM Press.
- Dreeben, R., & Barr, R. (1988). Classroom composition and the design of instruction. *Sociology of Education*, 61, 129–142.
- Enders, C.K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12, 121–138.
- Erikson, R., Goldthorpe, J.H., & Portocarero, L. (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. *The British Journal of Sociology*, 30, 415–441.
- Fortier, M.S., Vallerand, R.J., & Guay, F. (1995). Academic motivation and school performance: Toward a structural model. *Contemporary Educational Psychology*, 20, 257–274.
- Ganzeboom, H.B.G., De Graaf, P.M., & Treiman, D.J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, 21, 1–56.
- 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, 337–357.
- Goldstein, H. (1991). Better ways to compare schools? *Journal of Educational Statistics*, 16, 89–91.
- Goldstein, H., & Spiegelhalter, D.J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159, 385–443.

- Guo, G. (1998). The timing of the influences of cumulative poverty on children's cognitive ability and achievement. *Social Forces*, 77, 257–287.
- Hattie, J.A.C. (2002). Classroom composition and peer effects. *International Journal of Educational Research*, 37, 449–481.
- Heck, R.H., & Moriyama, K. (2010). Examining relationships Among elementary schools' context, leadership, instructional practices, and added-year outcomes: A regression discontinuity approach. *School Effectiveness and School Improvement*, 21, 377–408.
- Hill, P.W., & Rowe, K.J. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement*, 7, 1–34.
- 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 Researcher*, 36, 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, UK: Routledge.
- Koedel, C., & Betts, J.R. (2008). Test-score ceiling effects and value-added measures of school quality. In *JSM Proceedings, Social Statistics Section* (pp. 2370–2377). Alexandria, VA: American Statistical Association.
- Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources*, 37, 752–777.
- Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis*, 25, 287–298.
- Lehmann, R., & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis. Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin. Abschlussbericht über die Untersuchungen 2003, 2004 und 2005 an Berliner Grundschulen und grundständigen Gymnasien* [Assessment of reading and mathematics literacy – Developments in the Years 4 to 6 in Berlin. Final report on the studies in 2003, 2004, and 2005 in Berlin primary schools and pre-academic tracks]. Berlin, Germany: Humboldt Universität.
- Lehmann, R., & Nikolova, R. (2005). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis – Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin. Bericht über die Untersuchung 2003 an Berliner Grundschulen und grundständigen Gymnasien* [Assessment of reading and mathematics literacy – Developments in the Years 4 to 6 in Berlin. Report on the study in 2003 in Berlin primary schools and pre-academic tracks]. Berlin, Germany: Humboldt Universität.
- 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, 197–216.
- Luyten, H., Tymms, P., & Jones, P. (2009). Assessing school effects without controlling for prior achievement? *School Effectiveness and School Improvement*, 20, 145–165.
- Maaz, K., Baumert, J., & Trautwein, U. (2009). Genese sozialer Ungleichheit im institutionellen Kontext der Schule: Wo entsteht und vergrößert sich soziale Ungleichheit? [Emergence of social inequality in the institutional context of school: Where does social inequality emerge and grow?]. *Zeitschrift für Erziehungswissenschaft Special Issue* (12), 11–46.
- 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, 35–62.
- McCaffrey, D.F., Lockwood, J.R., Mariano, L.T., & Setodji, C. (2005). Challenges for value-added assessment of teacher effects. In R. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 111–144). Maple Grove, MN: JAM Press.
- 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 on 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. Harting, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 315–335). Göttingen, Germany: Hogrefe & Huber.

- Organisation for Economic Co-operation and Development. (Ed.). (2007). *PISA 2006. Science competencies for tomorrow's world*. Paris, France: Author.
- Organisation for Economic Co-operation and Development. (2008). *Measuring improvements in learning outcomes. Best practices to assess the value-added of schools*. Paris, France: Author.
- Opdenakker, M.-C., & Van Damme, J. (2000). Effects of schools, teaching staff and classes on achievement and well-being in secondary education: Similarities and differences between school outcomes. *School Effectiveness and School Improvement*, 11, 165–196.
- 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, 179–206.
- Postlethwaite, T.N., & Ross, K.N. (1992). *Effective schools in reading. Implications for educational planners*. Hamburg, Germany: The International Association for Evaluation of Educational Achievement.
- Raudenbush, S.W. (1995). Hierarchical linear models to study the effects of social context on development. In J.M. Gottman (Ed.), *The analysis of change* (pp. 165–201). Mahwah, NJ: Lawrence Erlbaum Associates.
- 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, 121–129.
- Raudenbush, S.W., & Willms, D.J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20, 307–335.
- Reynolds, D., Teddlie, C., & Townsend, T. (2000). An introduction to school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp. 3–25). London, UK: Routledge.
- Rogosa, D. (1995). Myths and methods: “Myths about longitudinal research” plus supplemental questions. In J.M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah, NJ: Lawrence Erlbaum Associates.
- Rubin, D.B., Stuart, E.A., & Zanutto, E. (2004). A potential outcome view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, 29, 103–116.
- Sammons, P. (1996). Complexities in the judgement of school effectiveness. *Educational Research and Evaluation*, 2, 113–149.
- Sammons, P. (2007). *School effectiveness and equity: Making connections. A review of school effectiveness and improvement research and its implications for practitioners and policy makers*. London, UK: CfBT.
- Sammons, P., & Luyten, H. (2009). Editorial article for special issue on alternative methods of assessing school effects and schooling effects. *School Effectiveness and School Improvement*, 20, 133–143.
- Sammons, P., Nuttall, D., & Cuttance, P. (1993). Differential school effectiveness: Results from a reanalysis of the Inner London Education Authority's Junior School Project data. *British Educational Research Journal*, 19, 381–405.
- Sanders, W.L. (2000). Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education*, 14, 329–339.
- Schafer, J.L., & Olsen, M.K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33, 545–571.
- Schatz, C.J., VonSecker, C.E., & Alban, T.R. (2005). Balancing accountability and improvement: Introducing value-added models to a large school system. In R. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 1–18). Maple Grove, MN: JAM Press.
- Scheerens, J. (1997). Conceptual models and theory-embedded principles on effective schooling. *School Effectiveness and School Improvement*, 8, 269–310.
- Schmidt, W.H., Houang, R.T., & McKnight, C.C. (2005). Value-added research: Right idea but wrong solution? In R. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 145–165). Maple Grove, MN: JAM Press.
- Singer, J.D., & Willet, J.B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

- 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, MN: JAM Press.
- Strand, S. (2010). Do some schools narrow the gap? Differential school effectiveness by ethnicity, gender, poverty, and prior achievement. *School Effectiveness and School Improvement*, 21, 289–314.
- 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, UK: Routledge.
- Teddlie, C., Stringfield, S., & Reynolds, D. (2000). Context issues within school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp. 160–186). London, UK: Routledge.
- Tekwe, C.D., Carter, R.L., Ma, C.-X., Algina, J., Lucas, M.E., Roth, J., ... Resnick, M.B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, 29, 11–36.
- Thomas, S. (1998). Value-added measures of school effectiveness in the United Kingdom. *Prospects*, 28, 91–108.
- Thomas, S., & Mortimore, P. (1996). Comparison of value-added models for secondary-school effectiveness. *Research Papers in Education*, 11, 5–33.
- Willet, J.B. (1988). Chapter 9: Questions and answers in the measurement of change. *Review of Research in Education*, 15, 345–422.
- Willms, J.D. (2000). Monitoring school performance for “standards-based reform”. *Evaluation & Research in Education*, 14, 237–253.
- Willms, J.D., & Raudenbush, S.W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, 26, 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, 569–595.

Appendix 1

Table A1. Descriptives of the analytical and excluded sample with mean (*SD*), % for sex.

	analytical sample (<i>n</i> = 3,074)	excluded sample (<i>n</i> = 1,852)
Reading achievement (Grade 4)	97.52 (15.1)	112.79 (12.3)
Reading achievement (Grade 5)	104.75 (12.8)	118.42 (10.2)
Reading achievement (Grade 6)	109.85 (12.6)	122.41 (11.5)
Math achievement (Grade 4)	96.36 (13.7)	112.58 (13.2)
Math achievement (Grade 5)	106.25 (14.1)	123.29 (13.6)
Math achievement (Grade 6)	114.84 (15.2)	132.88 (14.3)
Sex (female)	48.7%	51.7%
Books	2.41 (1.1)	3.19 (0.9)
Hiedu	3.47 (1.2)	4.40 (1.0)
Hisei	46.55 (15.7)	59.64 (14.9)

Note: Differences in estimates of Table A1 and Table 1 are due to different estimation algorithms.

Table A2. Model estimates for reading achievement.

<i>fixed effects</i>	Contextualised Attainment Models			Value-Added Models			Prior Attainment Models		
	Model 0	Model 1	Model 2	Model 0	Model 1	Model 2	Model 0	Model 1	Model 2
<i>student level b0</i>									
prior4								0.51***	0.50***
sex		1.84***	1.82***		1.70***	1.69***		0.33	0.33
books		1.92***	1.81***		1.81***	1.72***		0.76***	0.67***
hisei		0.13***	0.12***		0.13***	0.12***		0.05**	0.05*
hiedu		1.08***	0.99***		1.05***	0.96**		0.54***	0.49***
<i>student level b1</i>									
sex					−0.59**	−0.59**			
books					−0.24**	−0.22**			
hisei					−0.02	0.00			
hiedu					−0.02	−0.01			
<i>school level g0</i>									
intercept	109.99***	109.90***	109.97***	110.38***	110.29***	110.34	109.99***	109.87***	109.91***
school prior4									0.19**
school hisei			0.44***			0.42***			0.04
<i>school level g1</i>									
grade				6.19***	6.20***	6.20***			
school hisei						−0.02			
<i>random effects (in %)</i>									
<i>status</i>									
ICC	22.5			27.8			22.5		
explained student-level variance		10.3	10.4		14.2	14.3		43.9	43.9
explained school-level variance		58.4	83.6		59.1	85.0		87.8	93.1

(continued)

Table A2. (Continued).

	Contextualised Attainment Models			Value-Added Models			Prior Attainment Models		
	Model 0	Model 1	Model 2	Model 0	Model 1	Model 2	Model 0	Model 1	Model 2
<i>fixed effects</i>									
<i>growth rate</i>									
growth parameter				7.7					
variance									
explained student-					2.6	2.6			
level variance									
explained school-					19.2	21.4			
level variance									

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A3. Model estimates for mathematics achievement.

	Contextualised Attainment Models			Value-Added Models			Prior Attainment Models		
<i>fixed effects</i>	Model 0	Model 1	Model 2	Model 0	Model 1	Model 2	Model 0	Model 1	Model 2
<i>student level b0</i>									
prior4								0.70***	0.69***
sex		−3.17***	−3.19***		−3.27***	−3.28***		0.07	0.02
books		1.70***	1.59***		1.64***	1.53***		0.58***	0.51***
hisei		0.14***	0.13**		0.15***	0.13***		0.07**	0.06**
hiedu		1.51***	1.41***		1.56***	1.46***		0.79***	0.74***
<i>student level b1</i>									
sex					0.73***	0.73***			
books					0.00	0.00			
hisei					0.02	0.02			
hiedu					0.22*	0.22*			
<i>school level g0</i>									
intercept	114.98***	114.88***	114.95***	115.15***	115.06***	115.12***	114.98***	114.79***	114.83***
school prior4									0.15
school hisei			0.50***			0.49***			0.13
<i>school level g1</i>									
grade				9.21***	9.20***	9.20***			
school hisei						0.05			
<i>random effects (in %)</i>									
<i>status</i>									
ICC	21.5			23.9			21.5		
explained student-level variance		9.0	9.0		11.5	11.5		46.7	46.7
explained school-level variance		50.8	76.4		52.2	77.9		80.6	85.6

(continued)

Table A3. (Continued).

	Contextualised Attainment Models			Value-Added Models			Prior Attainment Models		
	Model 0	Model 1	Model 2	Model 0	Model 1	Model 2	Model 0	Model 1	Model 2
<i>fixed effects</i>									
<i>growth rate</i>									
growth parameter				22.8					
variance									
explained student-level					4.4	4.4			
variance									
explained school-level					11.1	18.5			
variance									

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

CHAPTER 4 STUDY 3

Jenny Lenkeit (in press). How effective are educational systems? A value-added approach to study trends in PIRLS. *JERO Journal of Educational Research Online*.

How Effective Are Educational Systems? A Value-Added Approach to Study Trends in PIRLS.

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

Jenny Lenkeit

Department of Education, University of Hamburg, Hamburg, Germany

Correspondence details: e-mail: jenny.lenkeit@uni-hamburg.de, Binderstr. 34, 20146 Hamburg, Germany; Tel: +49-(40)-42838-7559, Fax: +49-(40)-42838-2709;

How Effective Are Educational Systems? A Value-Added Approach to Study 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 as 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

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.

Keywords: 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. 63 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 (Organization for Economic Cooperation 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; Organization for Economic Cooperation and Development [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, 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 LSAs 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) (Teddle & 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. The 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; Mislavy, 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). Erikan (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 LSAs 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 LSAs 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 examined 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 levels, 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 decisions 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).

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 $\alpha=0.674$ and indicated that constituent items explained 32.2 % of the latent SES construct. Both imputed and non-imputed data showed a very similar reliability ($\alpha=0.657$ 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 (see Table 2).

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.

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	Parents' highest employment status	Parents' highest occupational 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
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

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		
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

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

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 (Human Development Report Office [HDRO], o.J.).

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

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_{0jk} is the mean achievement of school j in system k . At the school level π_{0jk} is a function of average achievement in system k (β_{00k}) and the error term that represents the schools deviation from the expected average achievement (u_{0jk}) (2). γ_{000} represents the average achievement across educational systems and u_{00k} 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_{0jk} 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_{00k} 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:

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

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).

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

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:

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

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

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

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

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. 24 % of the overall achievement variance is attributed to differences between schools and 16 % 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 roughly 10 % of the student level variance, 48 % of school level variance and 35 % of system level variance. 65 % of the overall system level variance thus remain unexplained and may be subject to other (potentially malleable) system level factors.

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, SESEM, AGE and HDI from the predicted score based on the model specifications (u_{00k} 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.

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

<i>fixed effects</i>		Unconditional Model		Conditional Model	
		Coefficient	SE	Coefficient	SE
intercept		521.3 *	6.7	525.3 *	5.6
<i>student level</i>					
	SES			27.6 *	1.6
<i>school level</i>					
	SESSM			28.1 *	6.9
<i>system level</i>					
	AGE			16.8	13.8
	HDI			-4.7	6.6
<i>random effects (in %)</i>					
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<0.05

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 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 positions lower, respectively.

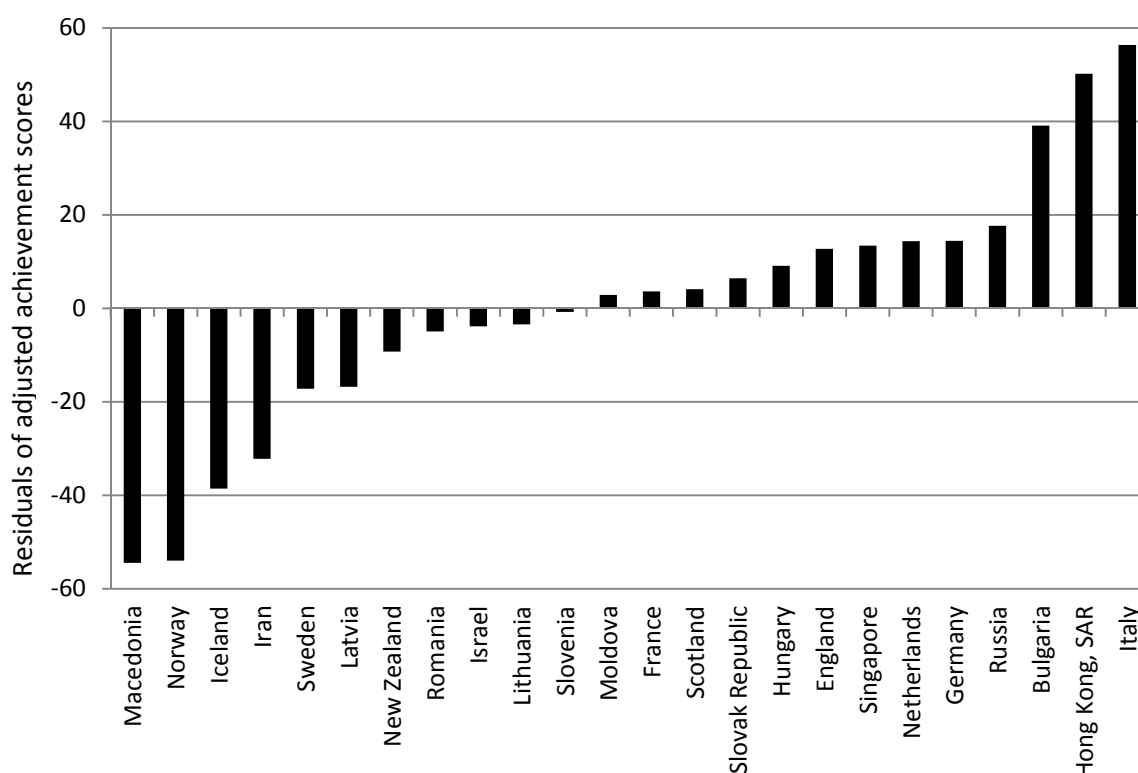


Figure 1. Residuals of adjusted achievement scores (i.e. effectiveness measures) in 2006 by educational system

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).

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 % of the overall variance is attributed to differences between schools and 15 % 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.

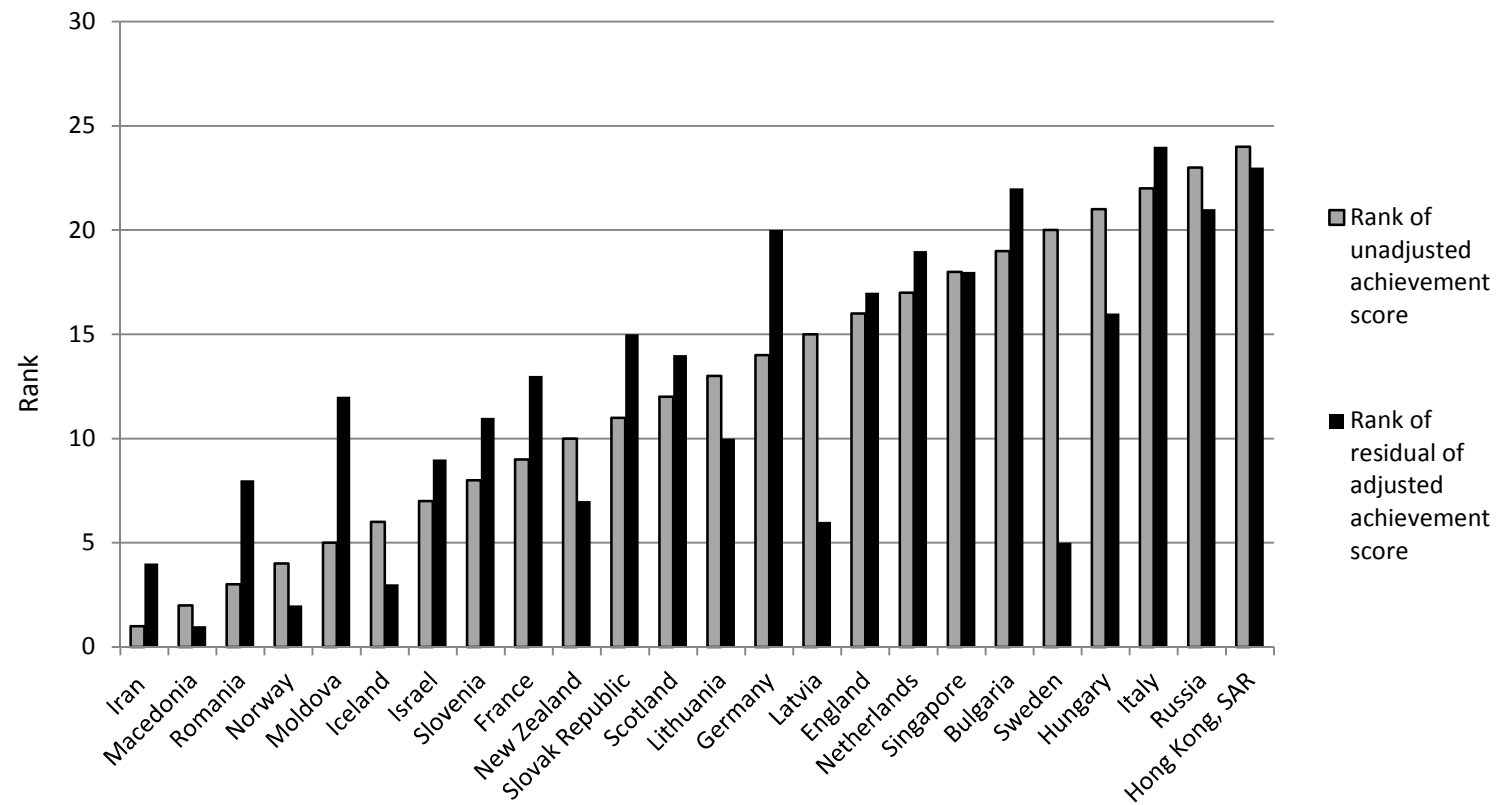


Figure 2. Differences in ranks of unadjusted achievement scores and residuals of adjusted achievement scores (i.e. effectiveness measures) by educational system

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.

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

<i>fixed effects</i>	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
<i>student level</i>						
SES					26.8 *	1.5
<i>school level</i>						
SESSM					28.8 *	5.7
COHORT			2.2	2.9	-11.6 *	3.6
<i>system level</i>						
AGE					35.3	37.7
HDI					1.3	14.6
cohort (b01)						
AGED					19.3 *	8.1
HDI					-2.6	2.5
<i>random effects (in %)</i>						
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<0.05

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 % of the student level variance, 46 % of school level variance and 41 % of system level variance. The included predictor variables moreover explain 27 % of the cohort parameter variance. 73 % 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.

Figure 3 illustrates the distribution of change score residuals. It can be seen that, taking SES, SESEM, 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.

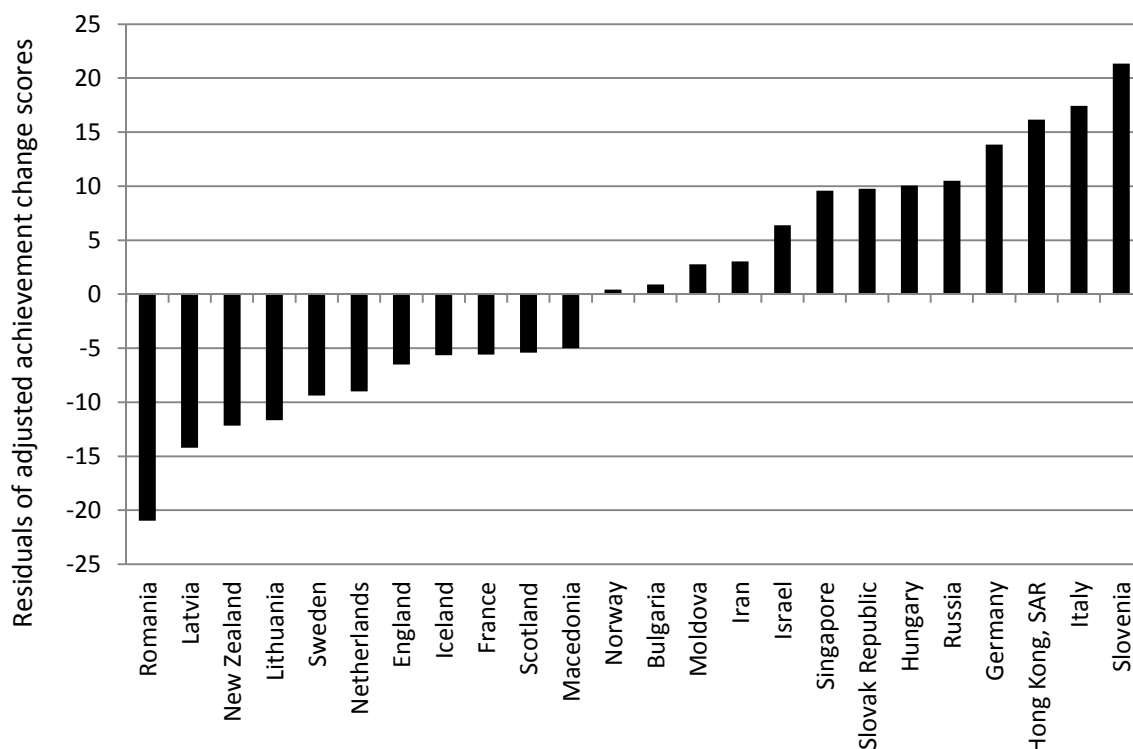


Figure 3. Residuals of adjusted achievement change scores (i.e. effectiveness measures) from 2001 to 2006 by educational system

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.

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 ($r = 0.451$). 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. 7 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.

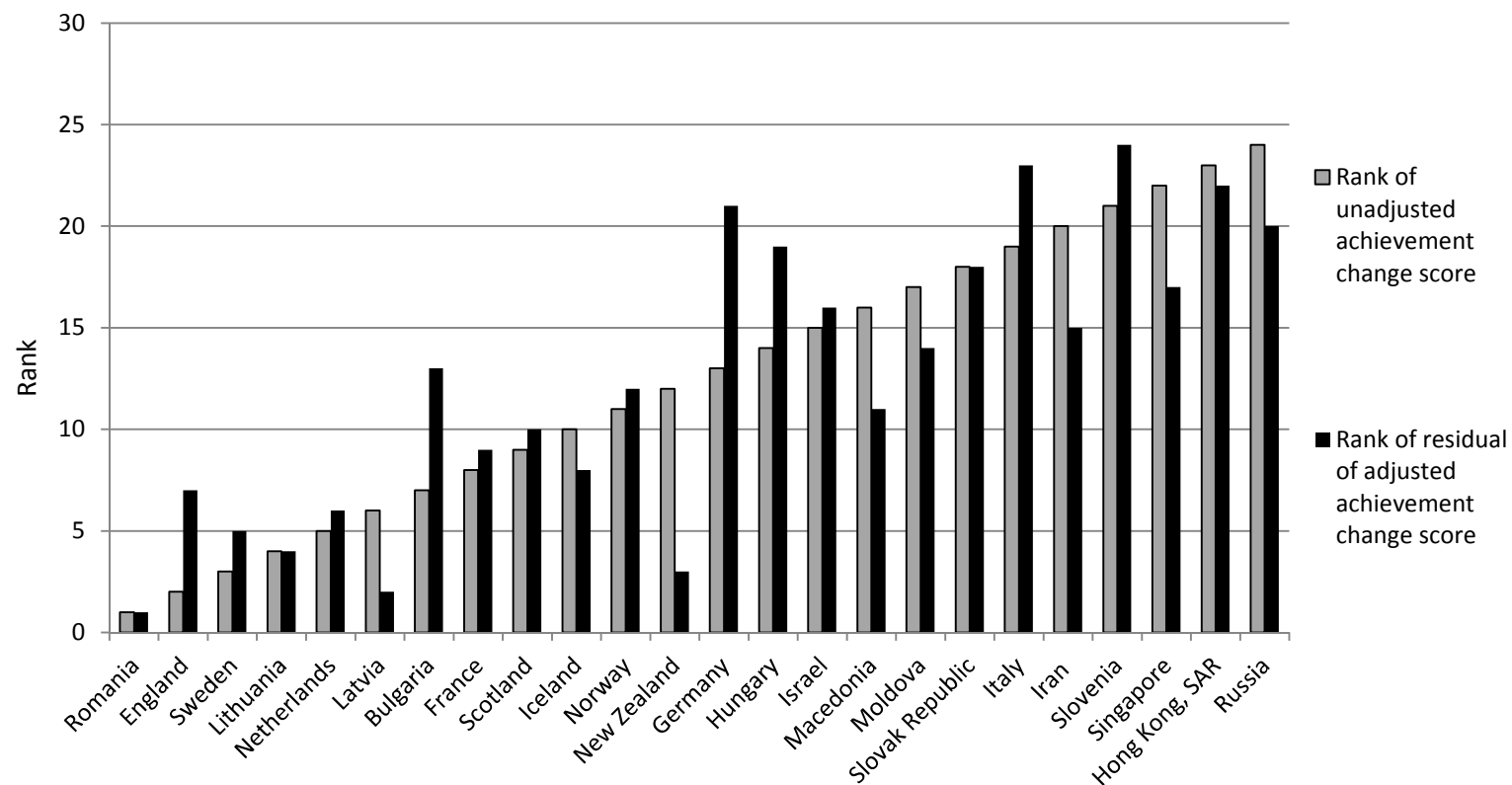


Figure 4. Differences in ranks of unadjusted achievement change scores and residuals of adjusted achievement change scores (i.e. effectiveness measure) by educational system

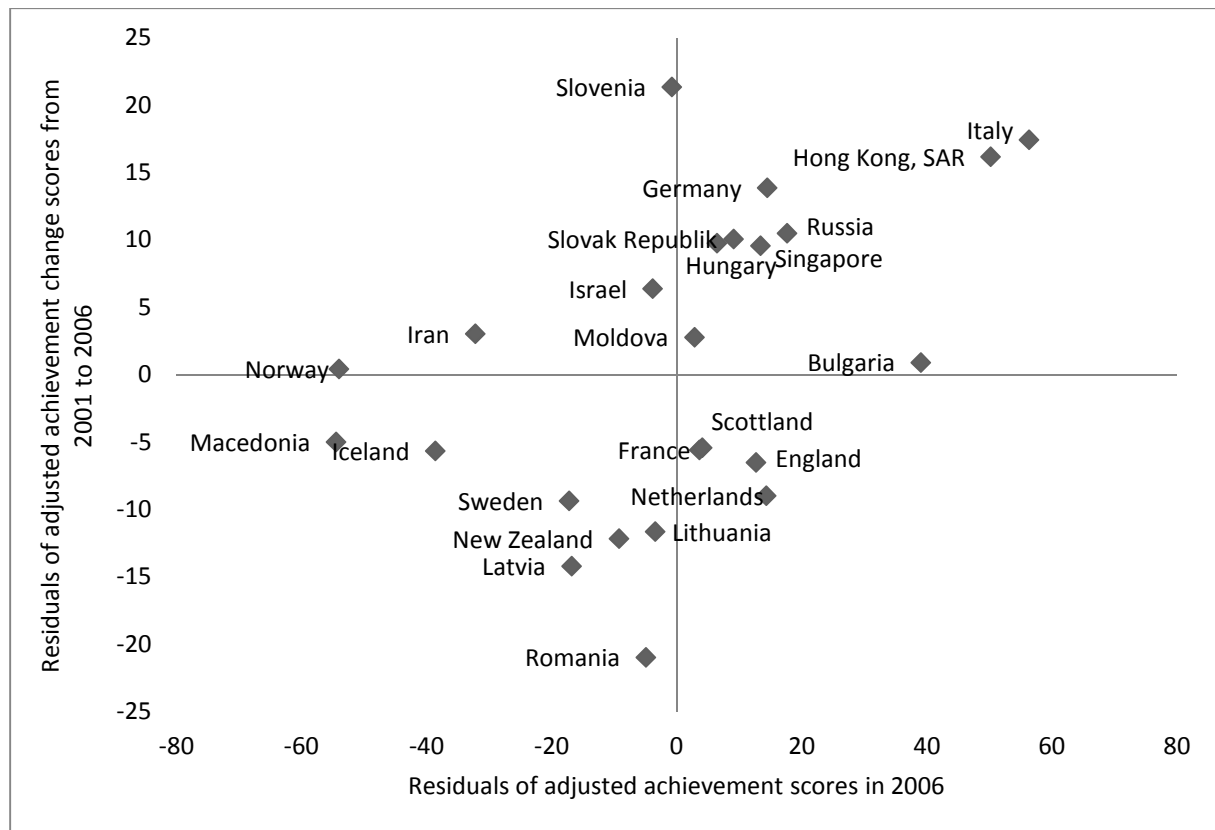


Figure 5. Correlation of residuals of adjusted achievement score in 2006 and residuals of adjusted achievement change scores from 2001 to 2006

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 investigating 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.

Notes

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?.

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).

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., Wittenboer, G. v. d., Hox, J. J., & De Glopper, C. (1999). The validity of comparative educational studies. *Educational Measurement: Issues and Practice*, Fall, 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 D.C.: 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 (Methodological)*, 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.
- 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.
- Human Development Report Office (o.J.). *Indices & Data. Human Development Index. Human Development Reports (HDR). United Nations Development Programme (UNDP)*. Retrieved from: <http://hdr.undp.org/en/statistics/hdi/> [24.11.2011].
- 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 Researcher*, 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: Routledge.

- 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*, 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*. Dissertations available from ProQuest. Paper AAI3073031.
- 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*. 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 on 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 (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools*. Paris: OECD.
- OECD (2009). *PISA 2009: Assessment framework: Key competencies in reading, mathematics and science*. Paris: OECD.
- OECD (2010a). *PISA 2009 results: Learning trends: Changes in student performance since 2000 (Volume V)*. Paris: OECD.
- OECD (2010b). *TALIS 2008: Technical Report*. Paris: OECD.
- OECD (2010c). *PISA 2009 Results: Overcoming social background: Equity in learning opportunities and outcomes (Volume 2)*. 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: The 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 surveys*. New York: 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 [25.06.2005].
- 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 Finland's role as projection surface 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.

CHAPTER 5 CONCLUSION AND DISCUSSION

5.1 Aims and objectives recapitulated

The thesis contrasts conceptual and methodological arguments for the use of effectiveness measures based on status and growth/change models and argues that the discrepancies between those arguments can be overcome. It was argued that achievement levels reflect ability, a more stable trait that comprises cognitive abilities and is strongly associated with environmental factors such as socioeconomic background (e.g. Guo, 1998). In contrast achievement growth is related to accomplishment which reflects students' capacity to acquire skills through effort, determination, and motivation (ibid.). In line with this argumentation Educational Effectiveness Research (EER) holds that achievement growth is more appropriate to measure effectiveness because it is independent of achievement levels and hence unconfounded with the socioeconomic status of students (Ballou et al., 2004; Teddlie, et al., 2000).

The thesis investigates if this distinction can be empirically observed, despite the fact that controlling for SES in status models theoretically purges the ability component in effectiveness measures obtained from achievement levels. It compares the practical consequences and statistical properties of effectiveness measures obtained for achievement levels and for achievement growth. The research agenda of the thesis comprises the following questions: Do teachers differentiate concepts associated with achievement levels (ability) and growth (accomplishment) and do they acknowledge growth in track recommendations? Are effectiveness measures obtained from status and growth models comparable (practical relevance, reliability of estimates)? Can effectiveness be established for educational systems and are effectiveness measures for performance levels and change in performance comparable in the sense that they capture the same notion of educational quality?

By asking these questions, the thesis wants to challenge the impeccable standing of growth measures for the evaluation of educational quality in EER. In particular, it investigates if and how this distinction is reflected in the specific notion of educational quality EER refers to: a "contextualized quality" that takes non-malleable contextual conditions into account. The relevance of this objective is reflected in the fact that in practice many study designs are cross-sectional, including designs of international comparative assessments. Based on arguments favoring growth measures, cross-sectional studies are often excluded for investigations on educational effectiveness. It further aims at broadening the way results of international achievement studies are reported and at enhancing the methodological

developments of educational effectiveness in cross-national comparative research. It moreover contributes to the discussion about the conceptual use and statistical appropriateness of effectiveness measures in educational research and policy. Rather than expanding research specifically aimed at factors enhancing educational effectiveness, the thesis accentuates the significance of establishing a conceptual and methodological sound basis on which studies aiming to identify effectiveness enhancing factors are at all sensible.

5.2 Summary and discussion of results

This section first summarizes the findings from the three individual articles. Secondly, it conflates and discusses these findings in the general context of the thesis and addresses limitations. Thirdly, some implications for researchers and policymakers are offered.

Investigations made in the first article (Caro et al., 2009) have shown that teachers take students' achievement growth into consideration for track recommendations. The independent influence of students' achievement growth on track recommendation suggests that teachers reward achievement levels and achievement growth as different concepts presumably related to the distinction of ability and accomplishment. However, unlike researchers concerned with educational effectiveness, for teachers as evaluators growth seems to be far less important than levels of achievement for track recommendations.

Further the article confirmed previous findings that SES, migration background, and gender (being a girl) negatively affect achievement levels in mathematics. Students with and without migration background grow equally in their mathematical skills, but girls and higher SES students have higher growth rates, although the impact of SES is smaller on achievement growth than on achievement levels. These findings are in line with the conceptual distinction, that achievement levels are much more confounded with environmental factors such as SES and that achievement levels are, therewith, a stronger representation of ability. In comparison, achievement growth is less influenced by environmental factors and more associated with accomplishments. Interestingly, the growth of students with lowest initial achievement levels is rewarded more strongly by teachers for school track recommendations than growth of the best performing students. Teacher case studies would have to confirm if this reward is related to a stronger estimation of those students' efforts and determination (accomplishments) to promote their learning rates despite disadvantaged initial positions.

The second article (Lenkeit, 2012) investigated the comparability of effectiveness measures for achievement levels and achievement growth. The overall results show that

measures for achievement levels and achievement growth are comparable with restrictions only. Correlations of effectiveness measures for achievement levels and achievement growth are medium with $r = .627$ for mathematics and $r = .532$ for reading achievement, indicating that empirical differences between the respective measures are indeed notable. The results further suggest that schools would be categorized differentially effective if ratings are based on effectiveness measures for achievement levels or achievement growth. These findings suggest that the conceptual distinction between achievement levels (representing ability) and achievement growth (representing accomplishment) are observable even when the confounding factors of ability (SES) are controlled for as is done in EER's notion of "contextualized quality".

Nevertheless, further investigations in the article relativize this finding. It was also shown that the growth measure itself can only be estimated with poor precision, i.e. low reliability of the achievement growth estimate. This finding is also supported with evidence from the first article (Caro et al., 2009) and restricts the comparisons made in the investigation. On that account effectiveness measures of change models were included in the investigations of the second article (Lenkeit, 2012). These measures conceptually comprise a longitudinal perspective by including prior achievement in a status model but have more reliable measures on which to base the interpretation of results. Effectiveness measures for change were highly correlated with effectiveness measures for achievement growth (above $r = .900$) and thus seem to capture a very similar notion of effectiveness. Effectiveness measures for achievement levels and for achievement change were less strong, but still also highly correlated ($r = .700$ for reading and $r = .710$ for mathematics). Thus, schools that outperformed their expected achievement levels also tended to outperform their expected changes in achievement. This finding indicates that once the confounding effects of SES are controlled for, achievement levels and change in achievement yield comparable indicators of "contextualized quality".

These findings initiated the investigations made in the third article (Lenkeit, in press) because they suggest that once confounding factors of ability (as represented in achievement levels) are controlled, achievement levels similarly capture what has been entitled accomplishment. Obtaining effectiveness measures from status models thus seems a reasonable endeavor. Hence, the third article investigated how effectiveness measures from cross-sectional data designs can be obtained for educational systems. The investigation was further motivated by shortcomings of international LSA to present adjusted performance scores for cross-country comparisons. The results show that the obtained effectiveness

measures present indeed a different picture of high and low performing educational systems compared with unadjusted performance scores.

Moreover, and in line with the thesis' research agenda, the comparability of performance levels and change in performance at the system level was also examined. The repeated cross-sectional design of PIRLS allowed for this longitudinal perspective in effectiveness at the level of educational systems. By controlling for relevant non-malleable intake factors, a notion of quality as "contextualized quality" was established in line with the convictions of the EER field. The findings suggest that educational systems which are effective for their performance levels also tend to be effective regarding their change in performance. However, correlation of the respective estimates suggest a lower relation between those measures than found in the second article (Lenkeit, 2012) with $r = .451$. Consequently, at the systemic level the conceptual distinction between levels of performance (ability at individual levels) and the capacity to improve less effective systemic structures and processes (accomplishment at the individual level) is more pronounced.

Regarding the comprehensive scope of this thesis and with the data at hand results are summarized as follows. Investigations made in this thesis showed that SES-adjusted achievement scores yield very different results than unadjusted achievement scores in the context of quality evaluation. Adjusted achievement scores apply a different standard on which the quality of educational institutions can be evaluated. It has been argued that an evaluation on these standards provides a notion of "contextualized quality". In the context of this notion of quality educational units can exhibit quality, i.e. be effective, although their descriptive results suggest otherwise. This is possible because the influence of non-malleable student characteristics is accounted for. Therewith the approach acknowledges that educators can be effective in the sense that their students may exceed their expected outcomes. While this finding is merely a confirmation of previous evidence in EER, it is new to research conducted at the level of educational systems. By taking the effect of economic and developmental differences between educational systems into account the demonstrated approach helps distinguishing high and low achieving systems from effective and ineffective systems.

Further, it was argued that controlling for environmental factors as represented by SES would level the distinction made between measures of achievement status and achievement growth and consequently yield similar estimates of effectiveness. The conceptual distinction between ability and accomplishment as reflected in achievement levels and achievement

growth would then also level out through the statistical adjustments applied. This could, however, not be conclusively observed through the investigations made. Results of analyses with the data at hand yielded strong empirical differences between effectiveness measures for achievement status and achievement growth and therewith initially suggest conceptual differences of the two.

It is argued that medium correlations of effectiveness measures from growth models and status models occur due to low reliabilities of growth measures. With only three measurement points growth measures are clearly obtained with too high uncertainty to allow for valid comparisons. This finding supports earlier evidence on the low reliability of growth measures (Raudenbush, 1995; Stevens, 2005). As an initial consequence, approaches obtaining effectiveness measures from cross-sectional data designs would be advocated for the following reason: Especially in high stakes systems as well as in research concerned with school improvement and development processes, results obtained from longitudinal data analysis form a highly valued evidence base to implement relevant policy decisions. However, if growth measures are acquired with high uncertainty as is the present case here, they build a misleading evidence base. Consequently, results that stem from more reliable estimates of achievement status are advocated over those stemming from unreliable growth estimates.

This conclusion is, however, far from suggesting the abandonment of longitudinal research designs. Their scientific value is unquestionable. Rather, it calls upon an increased amount of study designs with more than three measurement points. As Raudenbush (1995) points out one additional data point usually has substantial positive effects on the precision of estimates and can thus benefit the reliability of the growth parameter. Consequently, respective longitudinal data designs would be able to support the investigations strived for in this thesis. In particular they would allow valid investigations into whether the conceptual distinction of achievement levels and achievement growth is reflected in empirical data.

This also holds true for participation in international LSA. The third article mainly demonstrated a methodological approach to obtain effectiveness measures for educational systems including information from two assessment cycles from PIRLS. Clearly, other international LSA with more measurement points exist. Nevertheless, it is emphasized that more measurement points would allow to model growth, rather than just achievement change, on the level of educational systems. Consequently, governmental actors would do well in supporting the further participation in international LSA, much more so because

implementation of reforms on systemic level are lengthy and intended changes and effects will show only after several years (Goy, et al., 2010).

Nevertheless, prior achievement is one of the strongest predictors of subsequent achievement levels and should not be easily disregarded in models seeking to identify educational effectiveness (Ballou et al., 2004; Thomas & Mortimore, 1996). Therefore, as an alternative to achievement growth, it is suggested to include prior achievement indicators in status models. The obtained effectiveness measures are based on more reliable estimates and consequently yield more valid results. Thus, for studies with few measurement points only the use of change rates is advocated over growth measures.

Further, findings from these comparisons of achievement levels and achievement growth should not disregard potential flaws inherited in effectiveness measures from achievement levels. One major flaw is that indicators of SES are often measured with poor precision or are not equally valid to indicate SES across different groups of students (Hansson & Gustafsson, 2011). Considering that status models control for SES in order to eliminate its influence and consequently make inferences about educational quality, error prone measures of SES are a threat to the validity of respective inferences. This threat is of particular relevance in cross-national research as has been conducted in Lenkeit (in press). In fact, the comparability of the SES construct is limited by the different structures of social stratification across educational systems (Buchmann, 2000) and a different choice of constitutive variables would alter the association of the SES construct with achievement (Chudgar, Luschei, Fagioli, and Lee, 2012). Moreover, Caro and Cortés (2012) have provided evidence that the constitutive items are not equally indicative of SES across educational systems and further Caro, Sandoval-Hernandez, and Lüdtke (2012) have shown that even support of weak invariance for combinations of two educational systems was scarce.

The sometimes insufficient precision and comparative validity of SES indicators provides a further argument for a strong appeal for longitudinal data designs with sufficient measurement points in order to evaluate effectiveness of educational institutions. Reliable growth measures would be to a great extent unbiased by (potentially imprecise) SES measures and would yield a purer estimate that is conceptually related to what has been referred to as accomplishment.

Two more points shall not pass unnoticed as they, too, have implications for further research on educational effectiveness and the construction of study designs.

First, in the first article (Caro et al., 2009) and elsewhere (Guo, 1998; Alexander et al., 2001; Caro & Lehmann, 2009; Cortina et al., 2008) evidence was provided that SES and other background variables are also associated with achievement growth, not only achievement status. In line with researchers criticizing the distinction of the ability and accomplishment concept (Humphreys, 1974; Lohman, 2006), it has to be noted, that cognitive ability and SES also influence achievement growth (and therewith accomplishment) and that the concepts are therefore, at least empirically, not clearly distinguishable with the achievement tests administered in studies used within this thesis. Not only is family SES positively related to initial achievement levels, it also contributes to higher growth rates. This finding may be used to call upon a combined application of ability and achievement tests in respective assessments. While achievement levels are to a great extent a reflection of SES, they are far more associated with cognitive ability (e.g. Lehmann & Lenkeit, 2008). Thus, for a more accurate evaluation of the institutional effect, measures of cognitive ability and SES provide necessary information to model achievement growth (accomplishment) independent of the effects of ability.

Second, findings also confirm that schools are differentially effective for different groups of students (Sammons et al., 1993; Strand, 2010) as is indicated by overlapping confidence intervals of effectiveness measures. Differential effectiveness occurs mostly along stratifying student background characteristics such as ethnicity, gender, and SES (ibid.). However, another source of overlapping confidence intervals may be discussed. In the study presented here, school and class levels are not differentiated as no representative samples of classes within all schools have been at disposition in the analytic sample. Therefore, effectiveness measures for the school are partly obtained from more than two classes within the school. Students of those classes may have different background variables and may have been taught by different teachers. Therewith the effectiveness measure is somewhat a mixture of the effectiveness of two (or more) different teachers who may have worked with students of very different background characteristics. On this account an appeal is made to ensure that sufficient amounts of all units in the hierarchical educational system are administered (students, classes, tracks, schools, systems).

Conclusively, these results provide implications for different groups of actors within educational contexts.

Policymakers (decision-making in national assessment contexts): Although the evaluation of teachers' and schools' quality relies increasingly on effectiveness measures obtained from

growth models, results have shown that this procedure is vague when based on unreliable growth measures. It becomes intangibly unfair when high stakes decisions are involved. Consequently, the results not only caution against the use of growth measures for school-related decision making processes. They moreover provide arguments to invest in study designs with higher numbers of measurement points which would then be capable of uniting the conceptual arguments for the use of growth measures with the statistical quality requirements of these estimates. Respective investments should also be addressed to comprehensive samplings of classes within schools and within educational tracks.

Policymakers (decision-making in international assessment contexts): Adjusting performance scores in cross-national assessments and thereby differentiating less from more effective educational systems is especially relevant when researchers and policymakers compare their own educational systems with the seemingly most successful ones, hoping that the adaption of certain structures will remedy their own deficits. Thus a more careful handling of information drawn from league tables of unadjusted achievement scores is advocated. Further, Germany's more recent, but continuous participation in international large scale assessments should be maintained. Data from several more measurement points (assessment cycles) will allow researchers to model changes at the system level with sufficient precision and consequently derive stronger evidence about the impact of recent reform measures and structural adjustments.

Researchers (within national contexts): For researchers concerned with school development processes the results show that the identification of effectiveness enhancing factors (e.g. in "best-practice schools") is dependent on the effectiveness measure chosen. Identification and subsequent research based on unreliable growth measures can seriously bias further investigations and initiate misdirected reform measures. Consequently, further research is needed on the reliability of growth estimates and the conceptual and methodological comparability of effectiveness measures obtained from growth and status estimates.

Researchers (within international contexts): The third article has shown that it is possible (and reasonable) to transfer educational effectiveness research to research concerned with cross-national assessments of achievement. Including measures of "contextualized quality" in international reports provides additional information, especially relevant for policymakers. The applied methodological approach moreover includes a longitudinal perspective to measure change rates of the educational systems. Rather than using merely difference scores to measure the average progress made, hierarchical models adjusting for non-malleable

factors at different levels are recommended. Moreover, as PIRLS data will be available for the 2011 cycle in due time, the application of a model with three following cohorts would provide information about the potential impact of reforms and programs initiated after the first cycle, simply due to the fact that a sufficient amount of time has passed.

CHAPTER 6 SUMMARIES

6.1 English summary

Educational researchers increasingly use measures of achievement growth to evaluate the quality of classes and schools. However, neither is there consensus about the nature of quality that is captured by growth measures, nor about the advantages of the methodological approaches applied to obtain them. The doctoral thesis discusses conceptual arguments for the use of achievement level and achievement growth measures to evaluate educational quality. Three independent investigations compare the conceptual meaning of those quality indicators and discuss their statistical properties and the therewith associated suitability.

Recent decades have seen a mentionable increase in national and international assessments of educational outcomes in various school stages and domains around the world. This increase can only partly be ascribed to a scientific interest in the functioning and structures of educational institutions and systems, but is rather linked with a strong shift of steering mechanisms towards performance based evaluations (Küssau & Brüsemeister, 2007). This development has initiated a lively discussion among researchers and practitioners alike about the concept of quality in schools and educational systems (Helmke, Hornstein & Terhart, 2000). The respective quality is supposedly measured through standardized tests that aim to capture domain specific achievement. Next to aims and tasks of institutionalized education (development of cognitive, non-cognitive, and meta-cognitive skills) (Campbell, Kyriakides, Muijs, & Robinson, 2003) the discussion is mainly concerned with the applied quality criteria and fair comparisons (Nachtigall, Kröhne, Enders, & Steyer, 2008). Therewith connected are questions concerning the accountability of educators and their responsibility for the academic outcomes.

In this context quality as an accomplishment of a defined set of standards can be distinguished from “contextualized quality” which takes characteristics of the object itself for its evaluation into account. Within an educational context this notion of quality takes contextual characteristics that influence achievement into account when evaluating the quality of e.g. educators or schools. These characteristics are mainly associated with the social, economic, and ethnic background of students (also referred to as the student intake) which are non-malleable by educators but are decisively related to students’ academic achievements. Within this notion of quality, the object consequently functions as a quality criterion itself, i.e. has an individual reference. In this context, educational research refers to *educational*

effectiveness (Creemers & Kyriakides, 2008). Effectiveness is thereby defined as the relation of the observed outcome and the expected outcome (Scheerens & Creemers, 1989; Teddlie & Reynolds, 2000). What can be expected is a function of the student intake that influences academic achievement.

Educational Effectiveness Research (EER) further differentiates “contextualized quality” and its indicators into measures of adjusted achievement levels and measures of achievement growth. Achievement levels reflect to a substantial degree ability levels which are confounded with students’ socioeconomic background variables (Guo, 1998). Thus achievement levels have to be adjusted for these characteristics in order to obtain a measure of effectiveness. Measures of achievement growth better reflect the capacity of students to acquire knowledge and skills over their school careers and therewith capture the cumulative process of learning itself (Willet, 1988). Growth measures are commonly regarded as unconfounded with socioeconomic background characteristics (Andrejko, 2004) and therewith function as indicators of “contextualized quality”. As a consequence of this conceptual distinction EER promotes measures of achievement growth while adjusted measures of achievement levels are often regarded as inappropriate. Thus, data from cross-sectional study designs are often neglected in EER.

In fact the research literature lacks investigations which compare the notions of quality captured with measures of adjusted achievement levels and measures of achievement growth. This comparison is necessary in order to examine whether the conceptual distinction between those measures is also of practical relevance for the evaluation of classes, schools, or educational systems. In this sense it would be possible that, e.g. schools are effective with regard to enhancing students’ achievement growth, but ineffective with regard to levels of achievement.

Which measure is applied for the evaluation of educational institutions is, however, not only a question of theoretical considerations. Rather, statistical properties of effectiveness measures are of essential relevance. Particularly in high stakes evaluation systems such as in the USA and England educators and schools are judged on the basis of those evaluation results. These judgments often lead to severe personnel decisions sometimes extending up to the closure of entire schools (Kohn, 2000). There is consequently no question that effectiveness measures have to demonstrate statistical properties that safeguard respective decisions at least empirically.

Against this background, the thesis addresses the following questions:

1. Do teachers differentiate abilities and the capacity of students to acquire knowledge and skills when giving track recommendations in the sense that they consider students achievement growth across three academic years in their decisions?
2. Are effectiveness measures obtained from achievement levels and achievement growth comparable in the sense that they capture similar notions of quality? Or, are conceptual distinctions reflected in the empirical data in the form of different effectiveness estimates?
3. Can the concept of effectiveness be transferred to the level of educational systems and can effectiveness measures for educational systems be obtained? In the light of league tables of unadjusted raw scores typically presented in reports of international large scale assessments, it is further asked how a comparison of raw scores and effectiveness measures changes the picture of high and low performing educational systems.
4. Are differences between effectiveness measures obtained from achievement levels and achievement growth of practical relevance, in the sense that schools or educational systems are effective with regard to achievement growth, but ineffective with regard to levels of achievement?
5. Which statistical properties do estimates of achievement levels and achievement growth have and how reliable are they for the evaluation of educational effectiveness and quality, respectively?

These research questions are addressed in three articles. The first article (Caro, Lenkeit, Lehmann, & Schwippert, 2009) investigates if teachers consider students' achievement growth in school track recommendations over and above achievement levels at the end of primary school. Results show that, while controlling for student background characteristics and cognitive ability, achievement growth is indeed a relevant predictor of track recommendation. Nevertheless, it is also shown that the predictive strength of achievement levels is three times higher.

The second article (Lenkeit, 2012) investigates the comparability and appropriateness of effectiveness measures of achievement levels and achievement growth for the evaluation of schools' quality. The study shows that correlations of measures of achievement levels and growth are medium and that schools are categorized differently effective depending on the measure applied. It is moreover shown, that the reliability of growth measures is dissatisfactory low and that inferences based on these measures are highly charged with

uncertainty. Alternatively, measures of achievement levels are compared with measures of achievement change. Compared with measures of achievement growth those of achievement change are obtained from status models and include only two measurement points. Findings show that measures of achievement levels and achievement change are highly correlated and capture comparable concepts of effectiveness.

The third article (Lenkeit, in press) applies the methodological approach to obtain effectiveness measures to data from international large scale assessments which results are typically presented in the form of league tables of unadjusted raw scores. Controlling for differences in socioeconomic and developmental characteristics induces changes in ranks between educational systems. Further, measures of performance levels are compared with measures of change in performance. Measures demonstrate medium correlation and therewith suggest conceptual differences between effectiveness with regard to performance levels and effectiveness with regard to change in performance.

In summary, conceptual and empirical investigations made in the thesis suggest that achievement levels and achievement growth capture different notions of quality. Due to the low reliability of the growth measure it can, however, not be empirically investigated with certainty, if this distinction is also reflected in the obtained effectiveness measures. The low reliability of the growth measure is, however, not only a statistical restriction. Rather, it points out the therewith related tentativeness of the results and cautions against an evaluation of educational institutions based on these measures. Where only few measurement points are available, it is recommended to base quality judgments on effectiveness measures for achievement levels and/or achievement change, since they have been proved to be empirically comparable and show high reliability estimates.

These findings are, however, far from suggesting the abandonment of longitudinal research designs. Their scientific value is unquestionable. Nevertheless, findings point out that three measurement points are insufficient to obtain reliable growth measures. Consequently, they also stress the necessary caution to use those measures for evaluations of educational quality. Moreover, results of the thesis show that the impeccable standing of effectiveness measures for achievement growth can also be conceptually questioned.

6.2 German summary – Deutsche Zusammenfassung

Die Verwendung von Leistungszuwachsmäßen als Qualitätsmerkmal von Schule und Unterricht hat in der Bildungsforschung erheblich an Aufmerksamkeit gewonnen. Dennoch besteht derzeit weder Konsens über die mit diesen Mäßen erfasste Definition von Qualität noch über die statistische Angemessenheit der Güte von Leistungszuwachsmäßen. Die vorliegende kumulative Dissertationsschrift bespricht konzeptionelle Argumente für die Verwendung von Schätzern des Leistungsstatus und des Leistungszuwachses zur Beurteilung von Qualität im Bildungsbereich. Drei unabhängige empirische Untersuchungen vergleichen deren konzeptionelle Bedeutung als Qualitätsindikatoren und diskutieren die jeweiligen statistischen Eigenschaften sowie die damit verbundene Eignung.

Die Bedeutung von Leistungsstudien hat in den letzten Jahrzehnten in nationalen und internationalen Kontexten stark zugenommen. Dies ist nicht zuletzt eine Folge der Umorientierung hin zu output-orientierter Evaluation im Bildungsbereich, deren Ergebnisse Entscheidungsträgern Informationen zur Steuerung des Bildungssystems liefern sollen (Küssau & Brüsemeister, 2007). Diese Entwicklung hat unter Wissenschaftlern wie Praktikern eine angeregte Debatte über das Verständnis von schulischer und systemischer Qualität hervorgerufen (Helmke, Hornstein, & Terhart, 2000), die mit standardisierten, auf fachbezogene Leistung abzielenden Tests gemessen werden soll. Neben der generellen Diskussion um Ziele und Aufgaben institutionalisierter Bildung (Entwicklung kognitiver, non-kognitiver und meta-kognitiver Fähigkeiten) (Campbell, Kyriakides, Muijs, & Robinson, 2003), stehen die angelegten Qualitätskriterien und „faire Vergleiche“ im Fokus (Nachtigall, Kröhne, Enders, & Steyer, 2008). Damit verbunden sind Fragen der Rechenschaftslegung und Verantwortlichkeit für die erbrachten Leistungen, die sich in den festgelegten Qualitätskriterien widerspiegeln.

In diesem Kontext kann zwischen Qualität als Erfüllung von fest definierten Standards und „kontextualisierter Qualität“, welche spezifische Eigenschaften des Objektes zu dessen Bewertung berücksichtigt, differenziert werden. Bezogen auf den Bildungsbereich bedeutet „kontextualisierte Qualität“ demnach, dass unterrichts- und schulexterne Einflussfaktoren auf die Leistung bei der Qualitätsbeurteilung von z.B. Lehrkräften oder Schulen berücksichtigt werden. Dies sind vorrangig Merkmale des sozialen, ökonomischen und ethnischen Hintergrundes von Schülerinnen und Schülern, welche von den Lehrkräften nicht beeinflusst werden können, jedoch die Leistungen der Lernenden entscheidend mitbestimmen. Im Falle „kontextualisierter Qualität“ fungiert das Objekt somit selbst als Qualitätskriterium, hat also

eine individuelle Referenz. In der Bildungsforschung wird in diesem Zusammenhang auch von *educational effectiveness* gesprochen (Creemers & Kyriakides, 2008). Effektivität wird hier definiert als das Verhältnis von beobachteter Leistung und der zu erwartenden Leistung (Scheerens & Creemers, 1989; Teddlie & Reynolds, 2000). Die erwartete Leistung ist dabei eine Funktion der die Leistung beeinflussenden Hintergrundmerkmale.

Die Effektivitätsforschung unterscheidet „kontextualisierte Qualität“ und deren Indikatoren konzeptionell weiterhin in adjustierte Maße des Leistungsstatus und Leistungszuwachsmäße. Maße des Leistungsstatus erfassen Wissen und Fertigkeiten zu einem bestimmten Zeitpunkt, welche mit den Hintergrundmerkmalen der Schüler und Schülerinnen konfundiert sind und um diese adjustiert werden müssen. Leistungszuwachsmäße beziehen sich auf das Potenzial und die Fähigkeit, sich Wissen und Fertigkeiten anzueignen, und reflektieren damit, dass Lernen an sich ein kumulativer Prozess ist (Willet, 1988). Leistungszuwachsmäße werden allgemein hin als unabhängig von Hintergrundmerkmalen erachtet (Andrejko, 2004) und fungieren damit als Indikatoren „kontextualisierter Qualität“. Als Folge dieser konzeptionellen Unterscheidung werden in der Effektivitätsforschung Leistungszuwachsmäße befürwortet und adjustierte Maße des Leistungsstatus oft als unangemessen bezeichnet. Mithin werden Querschnittstudien für Fragestellungen der Effektivitätsforschung oft vernachlässigt.

Tatsächlich fehlen in der Forschungsliteratur jedoch Untersuchungen, welche die Verständnisse von Effektivität, i.e. Qualität, die über adjustierte Leistungsstatus- und Leistungszuwachsmäße gemessen werden, vergleichen, um damit Aussagen darüber treffen zu können, ob Leistungszuwachsmäße und Lernstatusmaße neben der konzeptionellen Unterscheidung auch praktisch zu differenzieren sind, wenn es um Qualitätsbeurteilungen von Unterricht, Schulen oder auch Systemen geht. In diesem Sinne gäbe es z.B. Schulen, die effektiv hinsichtlich ihres Leistungsstatus, jedoch ineffektiv hinsichtlich des Leistungszuwachses sind.

Welches Maß für die Beurteilung von Qualität verwendet wird, ist jedoch nicht nur eine Frage theoretischer Überlegungen. Vielmehr sind die statistischen Eigenschaften von Effektivitätsmaßen von ausschlaggebender Bedeutung. Vor allem in „high stakes“ Evaluationssystemen wie den USA oder England führen Qualitätsurteile auf der Basis von Effektivitätsmaßen zu teils schwerwiegenden Entscheidungen über Personal und Ressourcenverteilungen, bis hin zur Schließung ganzer Schulen (Kohn, 2000). Es steht somit

außer Frage, dass Effektivitätsmaße statistische Eigenschaften aufweisen müssen, welche die entsprechenden Entscheidungen zumindest empirisch absichern.

Vor diesem Hintergrund stellen sich für die vorliegende Arbeit folgende Fragen/Fragenkomplexe:

1. Differenzieren Lehrkräfte Fähigkeit und das Potenzial sich Wissen und Fertigkeiten anzueignen bei Übergangsempfehlungen in der Form, dass sie den Leistungszuwachs von Schülerinnen und Schülern über drei Schuljahre hinweg in den Empfehlungen berücksichtigen?
2. Sind Effektivitätsmaße des Leistungszuwachses und des Leistungsstatus vergleichbar in dem Sinne, dass sie gleiche Konzepte von Qualität erfassen? Oder lassen sich die konzeptionellen Unterscheidungen auch in den empirischen Daten als unterschiedliche Schätzer abbilden?
3. Kann das Effektivitätskonzept auf eine systemische Ebene übertragen und können somit Effektivitätsmaße von Bildungssystemen entwickelt werden? Mit Blick auf die bisher anhand von Rohwerten erstellten Ranglisten in der Berichtslegung international vergleichender Studien, stellt sich die Frage, wie sich dadurch das Bild erfolgreicher und weniger erfolgreicher Bildungssysteme verändert.
4. Sind Unterschiede zwischen Effektivitätsmaßen des Leistungsstatus und des Leistungszuwachses von praktischer Relevanz, in dem Sinne, dass Schulen bzw. Bildungssysteme als effektiv hinsichtlich des Leistungsstatus, jedoch ineffektiv des Leistungszuwachses kategorisiert werden können?
5. Welche statistischen Kennwerte weisen Leistungsstatus- und Leistungszuwachsmäße auf und wie verlässlich sind diese für die Beurteilung von Effektivität bzw. Qualität?

Die Forschungsfragen werden in drei Artikeln aufgegriffen. Der erste Artikel (Caro, Lenkeit, Lehmann, & Schwippert, 2009) untersucht, ob Lehrkräfte für die Übergangsempfehlungen der Schülerinnen und Schüler über den Leistungsstand am Ende der Grundschulzeit hinaus auch deren Leistungszuwachs berücksichtigen. Es zeigt sich, dass unter Kontrolle von Hintergrundmerkmalen und der kognitiven Fähigkeiten der Leistungszuwachs in der Tat ein signifikanter und relevanter Prädiktor der Übergangsempfehlung ist, obschon sich der Einfluss des Leistungsstatus als mehr als drei Mal so stark erweist.

Der zweite Artikel (Lenkeit, 2012) vergleicht die Bedeutung und Eignung von Effektivitätsmaßen des Leistungsstatus und des Leistungszuwachses für die Qualitätsbeurteilung von Schulen. Die Studie zeigt, dass Zuwachsmäße und Statusmäße nur mittelhoch miteinander korrelieren und Schulen hinsichtlich ihrer Effektivität je nach Maß teils unterschiedlich kategorisiert werden. Darüber hinaus zeigt sich, dass die Reliabilität des Zuwachsmäßes unbefriedigend niedrig ist und darauf basierende Schlussfolgerungen in höchstem Maße mit Unsicherheit belastet sind. Alternativ zum Leistungszuwachsmäß werden deshalb Statusmäße mit Veränderungsmaßen, die im Vergleich zum Zuwachsmäß nur zwei Erhebungszeitpunkte einschließen, verglichen. Es zeigt sich, dass Statusmäße und Veränderungsmaße hoch miteinander korrelieren und vergleichbare Konzepte von Effektivität erfassen.

Der dritte Artikel (Lenkeit, im Druck) überträgt das methodologische Vorgehen für die Generierung von Effektivitätsmaßen auf Daten international vergleichender Schulleistungsstudien, deren Ergebnisse in der Regel in Form von unadjustierten Leistungswerten dargestellt werden. Die Kontrolle von Unterschieden in der ökonomischen Entwicklung zwischen den Bildungssystemen führt zu Verschiebungen in deren Rangfolgen und Leistungspunkten. Darüber hinaus wurden hier ebenfalls Statusmäße und Veränderungsmaße auf Systemebene verglichen. Beide Effektivitätsmäße korrelieren mittelstark miteinander und suggerieren damit, dass auf Systemebene unterschiedliche Konzepte von Effektivität erfasst werden.

Zusammenfassend zeigen die konzeptionellen und empirischen Untersuchungen der Arbeit, dass Leistungsstatus und Leistungszuwachs unterschiedliche Konzepte von Effektivität, i.e. Qualität erfassen. Aufgrund der niedrigen Reliabilität des Zuwachsschätzers kann jedoch nicht mit Sicherheit empirisch überprüft werden, ob diese Unterscheidung sich auch in den Ergebnissen der statistischen Modelle widerspiegelt. Die niedrige Reliabilität des Leistungszuwachsschätzers stellt dabei nicht nur ein statistisches Problem dar. Vielmehr verweist die damit verbundene Unsicherheit der Ergebnisse auf erhöhte Vorsicht, wenn hierauf basierend Bildungsinstitutionen und ihre Qualität beurteilt werden sollen. Für die vorliegenden Daten wird daher empfohlen, Qualitätsurteile auf der Basis von Veränderungsmaßen bzw. Statusmaßen vorzunehmen, da diese sich als konzeptionell vergleichbar erwiesen haben und höhere statistische Reliabilitäten aufzeigen.

Dennoch ist die Intention der Arbeit nicht in einer Diskreditierung von Längsschnittstudien und deren Maßen zu sehen. Deren wissenschaftlicher Wert ist

unbestreitbar. Die Ergebnisse verweisen jedoch darauf, dass drei Messzeitpunkte ungenügend sind, um verlässliche Schätzer des Leistungszuwachses zu generieren, und betonen die notwendige Sensibilität, diese für Qualitätsbeurteilungen zu nutzen. Darüber hinaus zeigen die Ergebnisse, dass die überhöhte Stellung von Effektivitätsmaßen basierend auf dem Leistungszuwachs auch konzeptionell hinterfragt werden kann.

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.
- Andrejko, L. (2004). Value-added assessment: A view from a practitioner. *Journal of Educational and Behavioral Statistics*, 29 (1), 7–9.
- 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. (2002). *Sizing up test scores*. Retrieved from: www.educationnext.org [Summer 2002].
- 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.
- Baumert, J. & Schümer, G. (2001). Familiäre Lebensverhältnisse. Bildungsbeteiligung und Kompetenzerwerb [Family background, educational participation, and competence acquisition]. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann, & M. Weib (Eds.), *PISA 2000 Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 323–407). Opladen, Germany: Leske, Budrich.
- Baumert, J., Stanat, P., & Watermann, R. (Eds.). (2006). *Herkunftsbedingte Disparitäten im Bildungswesen. Vertiefende Analysen im Rahmen von PISA 2000* [Class-related disparities in the educational system. In-depth analyses in the scope of PISA 2000]. Wiesbaden: Verlag für Sozialwissenschaften.
- Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Report*. Chestnut Hill, MA: Boston College
- Bernstein, B. (1975). *Class, codes and control*. London: Routledge & Kegan Paul. (Vol. 3).
- 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.
- Bos, W., Bonsen, M., Gröhlich, C., Jelden, D., & Rau, A. (2006). *Erster Bericht zu den Ergebnissen der Studie „Kompetenzen und Einstellungen von Schülerinnen und Schülern – Jahrgangsstufe 7“ (KESS 7)* [First report of results from the „Kompetenzen und Einstellungen von Schülerinnen und Schülern – Jahrgangsstufe 7“ study (KESS 7)]. Forschungsbericht für die Behörde für Bildung und Sport der Stadt Hamburg.

- Bos, W., Hornberg, S., Arnold, K.-H., Faust, G., Fried, L., Lankes, E.-M., Schwippert, K., & Valtin, R. (Eds.). (2007). IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich [IGLU 2006. Reading competencies of primary students in Germany in an international comparison]. Münster: Waxmann.
- Bos, W., Hornberg, S., Arnold, K.-H., Faust, G., Fried, L., Lankes, E.-M., Schwippert, K., & Valtin, R. (Eds.). (2008). IGLU-E 2006. Die Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich. [IGLU-E 2006: National and international comparison of the German Länder]. Münster: Waxmann.
- Bos, W., Stubbe, T. C., & Buddeberg, M. (2010). Einkommensarmut und schulische Kompetenzen [Poverty and educational competences]. In J. Fischer & R. Merten (Eds.), *Armut und soziale Ausgrenzung von Kindern und Jugendlichen. Problembestimmungen und Interventionsansätze* (pp. 58–72). Baltmannsweiler: Schneider Verlag Hohengehren.
- Bourdieu, P. (1983). Forms of Capital. In J. E. Richardson (Ed.), *Handbook of Theory and Research for the Sociology of Education* (pp. 241–258). New York: Greenwood.
- Bourdieu, P. & Passeron, J.-C. (1977). *Reproduction in Education, Society and Culture*. London: Sage Publications.
- 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 D.C.: National Academy Press.
- Bundesministerium für Bildung und Forschung (BMBF) (2007) (Eds.). Zur Entwicklung nationaler Bildungsstandards. Eine Expertise [On the development of national educational standards. An expertise]. Bonn: Bundesministerium für Bildung und Forschung (BMBF), Referat Öffentlichkeitsarbeit.
- Burbules, N. C. (2004). Ways of thinking about educational quality. *Educational Researcher*, 33 (6), 4–10.
- 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, Volume 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., Lenkeit, J., Lehmann, R., & Schwippert, K. (2009). The role of academic achievement growth in school track recommendations. *Studies in Educational Evaluation*, 35 (4), 183–192.

- Caro, D. H., McDonald, J. T., & Willms, D. J. (2009). Socio-economic status and academic achievement trajectories from childhood to adolescence. *Canadian Journal of Education* 32 (3), 558–590.
- Caro, D. H., Sandoval-Hernandez, A., & Lüdtke, O. (2012). *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.
- Choi, K. & Seltzer, M. (2003). *Addressing questions concerning equity in longitudinal studies of school effectiveness and accountability: Modeling heterogeneity in relationships between initial status and rates of change*. Los Angeles: University of California.
- 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.
- Coleman, J. S. (1988). Social capital and the creation of human capital. *The American Journal of Sociology*, 94, 95–120.
- Condrón, D. J. (2007). Stratification and educational sorting: Explaining ascriptive inequalities in early childhood reading group placement. *Social Problems*, 54 (1), 139–160.
- 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. (Contexts of learning).
- Decker, D. M. & Bolt, S. E. (2008). Challenges and opportunities for promoting student achievement through large-scale assessment results: Research, reflections and future directions. *Assessment for Effective Intervention*, 34 (1), 43–51.
- De Maeyer, S., van den Bergh, H., Rymeanans, R., Van Petegem, P., & Rijlaarsdam, G. (2010). Effectiveness criteria in school effectiveness studies: Further research on the choice for a multivariate model. *Educational Research Review*, 5 (1), 81–96.
- Doran, H. C. & Cohen, J. (2005). The confounding effects of linking bias on gains estimated from value-added models. In R. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 80–110). Maple Grove, Minnesota: JAM Press.
- Fend, H. (2000). Qualität und Qualitätssicherung im Bildungswesen [Quality and quality assurance in the educational system]. In A. Helmke, W. Hornstein & E. Terhart (Eds.), *Qualität und Qualitätssicherung im Bildungsbereich: Schule, Sozialpädagogik, Hochschule* (S. 55–72). Weinheim: Beltz.

- Fend, H. (2008). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität* [Designing schools. System control, school development and quality of instruction]. Wiesbaden: Verlag für Sozialwissenschaften.
- 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.
- 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.
- Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education: Principles, Policy and Practice*, 11 (3), 319–330.
- Goy, M., Gröhllich, C., Strietholt, R., Stubbe, T. C., Bos, W., & Kanders, M. (2010). Panelstudien als Antwort auf Forschungsdesiderate in der Sekundarstufe I [Panel studies as an answer to research desiderata in lower secondary schools]. In N. Berkemeyer, W. Bos, H. G. Holtappels, N. McElvany, & R. Schulz-Zander (Eds.), *Jahrbuch der Schulentwicklung* (pp. 37–70). Weinheim: Juventa.
- Guo, G. (1998). The timing of the influences of cumulative poverty on children's cognitive ability and achievement. *Social Forces*, 77 (1), 257–288.
- Hansson, A. & Gustafsson, J.-E. (2011). Measurement invariance of socioeconomic status across migrational background. *Scandinavian Journal of Educational Research*. Available at: <http://dx.doi.org/10.1080/00313831.2011.625570>.
- Harvey, L. & Green, D. (2000). Qualität definieren. Fünf unterschiedliche Ansätze [Defining quality. Five different approaches]. *Zeitschrift für Pädagogik*, 41 (Beiheft), 17–39.
- Haveman, R. & Wolfe, B. (1995). The determinants of children's attainments: A review of methods and findings. *Journal of Economic Literature*, 33 (4), 1829–1878.
- Heckhausen, H. (1981). Chancenausgleich [Compensation of opportunities]. In H. Schiefele & A. Krapp (Eds.), *Handlexikon zur Pädagogischen Psychologie* (pp. 54–61). München: Ehrenwirt.
- Heid, H. (2000). Qualität: Überlegungen zur Begründung einer pädagogischen Beurteilungskategorie [Quality: Considerations on the legitimation of a pedagogical evaluation category]. *Zeitschrift für Pädagogik, Beiheft*, 41, 41–51.
- Helmke, A. (2007). *Unterrichtsqualität erfassen, bewerten, verbessern* [Quality of instruction - coverage, evaluation, improvement]. Seelze: Kallmeyer.
- Helmke, A., Hornstein, W., & Terhart, E. (Eds.). (2000). *Qualität und Qualitätssicherung im Bildungsbereich: Schule, Sozialpädagogik, Hochschule* [Quality and quality assurance within the educational system: School, social work and education, higher education]. Weinheim: Beltz.
- Hill, P. W. & Rowe, K. J. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement*, 7 (1), 1–34.
- Hofer, M., Kuhnle, C., Kilian, B., & Fries, S. (2012). Cognitive ability and personality variables as predictors of school grades and test scores in adolescents. *Learning and Instruction*, 22, 368–357.

- Holtappels, H. G. (2003). *Schulqualität durch Schulentwicklung und Evaluation* [School quality through school development and evaluation]. München: Luchterhand.
- Humphreys, L. G. (1974). The misleading distinction between aptitude and achievement tests. In Green, D. R. (Ed.), *The Aptitude-Achievement Distinction* (pp. 262–267). McGraw-Hill: CTB.
- 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 Researcher*, 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: Routledge.
- Klieme, E. & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education*, 16, 383–400.
- Kohn, A. (2000). Burnt at high stakes. *Journal of Teacher Education*, 51 (4), 315–327.
- Köller, O. (2009). Evaluation pädagogisch-psychologischer Maßnahmen [Evaluation of pedagogical-psychological programs]. In E. Wild & J. Möller (Eds.), *Pädagogische Psychologie* (pp. 333–352). Heidelberg: Springer.
- Kuper, H. (2002). Stichwort: Qualität im Bildungssystem [Keyword: Quality in the education system]. *Zeitschrift für Erziehungswissenschaft* 5 (4), 533–551.
- Küssau, J. & Brüsemeister, T. (2007). Educational Governance: Zur Analyse der Handlungskoordination im Mehrebenensystem der Schule [Educational governance: The analysis of coordinative action in the multilevel system of the school]. In H. Altrichter, T. Brüsemeister, & J. Wissinger (Eds.), *Educational Governance: Handlungskoordination und Steuerung im Bildungswesen* (pp. 15–54). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Kyriakides, L. (2004). Differential school effectiveness in relation to sex and social class: Some implications for policy evaluation. *Educational Research and Evaluation: An International Journal on Theory and Practice*, 10 (2), 141–161.
- Lehmann, R. H. (2006). Zur Bedeutung der kognitiven Heterogenität von Schulklassen für den Lernstand am Ende der Klassenstufe 4 [The relevance of cognitive heterogeneity in classes for achievement levels at the end of grade 4]. In A. Schröder-Lenzen (Ed.), *Risikofaktoren kindlicher Entwicklung. Migration, Leistungsangst und Schulübergang* (pp. 109–121). Wiesbaden: Verlag für Sozialwissenschaften.
- Lehmann, R. & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis. Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin. Abschlussbericht über die Untersuchungen 2003, 2004 und 2005 an Berliner Grundschulen und grundständigen Gymnasien* [Assessment of Reading and Mathematic Literacy - Developments in the Years 4 to 6 in Berlin. Final Report on the Studies in 2003, 2004, and 2005 in Berlin Primary Schools and Pre-Academic Tracks]. Berlin, Germany: Humboldt Universität.
- Lehmann, R. H. & Nikolova, R. (2005). *ELEMENT. Erhebungen zum Lese- und Mathematikverständnis. Entwicklungen in den Klassenstufen 4 bis 6 in Berlin. Bericht über die Untersuchung 2003 an Berliner Grundschulen und grundständigen Gymnasien* [ELEMENT. Survey for reading and mathematic literacy. Developments

- in grades 4 to 6 in Berlin. Research report on the survey in 2003 in primary schools and undergraduate academic tracks in Berlin] Berlin: Senatsverwaltung für Bildung, Jugend und Sport.
- Lehmann, R. H. & Peek, R. (1997). *Aspekte der Lernausgangslage von Schülerinnen und Schülern der fünften Jahrgangsstufe an Hamburger Schulen. Bericht über die Untersuchung im September 1996* [Aspects of initial achievement status of 5th grade students in schools in Hamburg. Research report on the September 2006 survey]. Hamburg: Behörde für Schule, Jugend und Berufsausbildung, Amt für Schule.
- Lehmann, R. H., Peek, R., Gänsfuss, R., Lutkat, S., Mücke, S., & Barth, I (2001). *QuaSUM. Qualitätsuntersuchung an Schulen zum Unterricht in Mathematik. Ergebnisse einer repräsentativen Untersuchung im Land Brandenburg* [QuaSUM. Quality survey of mathematic instruction in schools. Results of a representative survey in the federal state Brandenburg]. Potsdam: Ministerium für Bildung, Jugend und Sport des Landes Brandenburg.
- Lenkeit, J. (2012). Effectiveness measures for cross-sectional studies: A comparison of value-added models and contextualised attainment models. *School Effectiveness and School Improvement*. Available at: <http://www.tandfonline.com/doi/abs/10.1080/09243453.2012.680892>
- Lenkeit, J. (im Druck). How effective are educational systems? A value-added approach to study trends in PIRLS. *JERO Journal of Educational Research Online*.
- Lohman, D. F. (2006). Beliefs about differences between ability and accomplishment: From folk theories to cognitive science. *Roeper Review*, 29 (1), 32–40.
- Maaz, K., Baumert, J., & Trautwein, U. (2009). Genese sozialer Ungleichheit im institutionellen Kontext der Schule: Wo entsteht und vergrößert sich soziale Ungleichheit? [Emergence of social inequality in the institutional context of school: Where does social inequality emerge and grow?]. *Zeitschrift für Erziehungswissenschaft, Special Issue* (12), 11–46.
- 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 scales student achievement data for growth-based, value-added accountability. *Journal for Educational and Behavioral Statistics*, 31 (1), 35–62.
- 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.
- Mueller, C. W. & Parcel, T. L. (1981). Measures of socioeconomic status: Alternatives and recommendations. *Child Development*, 52 (1), 13–30.
- Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report*. 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 on 40 countries*. Chestnut Hill, MA: Boston College.
- OECD. (2008). *Measuring improvements in learning outcomes. Best practices to assess the value-added of schools*. Paris: OECD Publishing.

- OECD. (2009). *PISA 2009. Assessment framework. Key competencies in reading, mathematics and science*. Paris: OECD Publishing.
- OECD (2010). *PISA 2009 results: Learning trends. Changes in student performance since 2000 (Volume V)*. Paris: OECD.
- Olsen, R. V. (2005). *Achievement test from an item perspective*. Oslo: University of Oslo.
- Opdenakker, M.-C. & Van Damme, J. (2000). Effects of schools, teaching staff and classes on achievement and well-being in secondary education: Similarities and differences between school outcomes. *School Effectiveness and School Improvement*, 11 (2), 165–196.
- 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. & Ross, K. N. (1992). *Effective schools in reading. Implications for educational planners*. Hamburg: The International Association for Evaluation of Educational Achievement.
- Raudenbush, S. W. (1995). Hierarchical linear models to study the effects of social context on development. In J. M. Gottman (Ed.), *The analysis of change* (pp. 165–201). New Jersey: Lawrence Erlbaum Associates.
- 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.
- Raudenbush, S. W. & Willms, D. J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20 (4), 307–335.
- Reynolds, D., Teddlie, C., & Townsend, T. (2000). An introduction to school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp. 3–25). London: Routledge.
- Rice, J. K. (2010). Principal effectiveness and leadership in an era of accountability: What research says. Brief 8. *National Center for Analysis of Longitudinal Data in Education Research*. Washington: The Urban Institute.
- Sammons, P. (1996). Complexities in the judgement of school effectiveness. *Educational Research and Evaluation*, 2 (2), 113–149.
- Sammons, P. (2007). *School effectiveness and equity: Making connections. A review of school effectiveness and improvement research and its implications for practitioners and policy makers*. London: CfBT.
- Sammons, P., Nuttall, D., & Cuttance, P. (1993). Differential school effectiveness: Results from a reanalysis of the inner London education authority's junior school project data. *British Educational Research Journal*, 19 (4), 381–405.
- Sanders, W. L. (2000). Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education*, 14, 329–339.

- Scheerens, J. (1997). Conceptual models and theory-embedded principles on effective schooling. *School Effectiveness and School Improvement*, 8 (3), 269–310.
- Scheerens, J. (2004) Perspectives on Education Quality, Education Indicators and Benchmarking. *European Educational Research Journal*, 3 (1), 115–138.
- Scheerens, J. & Creemers, B. P. M. (1989). Conceptualizing school effectiveness. *International Journal of Educational Research*, 13 (7), 691–706.
- Schnabel, K. U., Alfeld, C., Eccles, J. S., Köller, O., & Baumert, J. (2002). Parental influence on students' educational choices in the United States and Germany: Different ramifications-same effect? *Journal of Vocational Behavior*, 60, 178–198.
- Schwippert, K. & Goy, M. (2008). Leistungsvergleichs- und Schulqualitätsforschung [Research on achievement assessment and school improvement]. In W. Helsper & J. Böhme (Eds.), *Handbuch der Schulforschung* (2., durchgesehene und erweiterte Aufl., pp. 387–421). Wiesbaden: Verlag für Sozialwissenschaften.
- Schwippert, K. & Lenkeit, J. (2012). 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. (Studies in International Comparative and Multicultural Education).
- Schwippert, K. & Walker, M. (2003). Homogenous and high performing classes: The case of optimal classes. *Studies in Educational Evaluation*, 29 (2), 109–128.
- Singer, J. D. & Willet, J.B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.
- Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 75 (3), 417–453.
- 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, Minnesota: JAM Press.
- Strand, S. (2010). Do some schools narrow the gap? Differential school effectiveness by ethnicity, gender, poverty, and prior achievement. *School Effectiveness and School Improvement*, 21 (3), 289–314.
- 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.
- 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.

- Tiedemann, J. & Billmann-Mahecha, E. (2007). Zum Einfluss von Migration und Schulklassenzugehörigkeit auf die Übergangsempfehlung für die Sekundarstufe I. [The influence of ethnic criteria and frame of reference effects on teachers' recommendations in regard to transition from primary to secondary education]. *Zeitschrift für Erziehungswissenschaft*, 10 (1), 108–120.
- Tramonte, L. & Willms, D. J. (2010). Cultural capital and its effects on education outcomes. *Economics of Education Review*, 29 (2), 200–213.
- Updegraff, K., Eccles, J., Barber, B., & O'Brien, K. (1996). Course enrollment as selfregulatory behavior: Who takes optional high school math courses? *Learning and Individual Differences*, 8 (3), 239–259.
- Wang, M., Haertel, G., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research*, 63 (3), 249–294.
- Willet, J. B. (1988). Chapter 9: Questions and answers in the measurement of change. *Review of Research in Education*, 15, 345–422.
- Willms, J. D. (2000). Monitoring school performance for “Standards-based Reform”. *Evaluation and Research in Education*, 14 (3), 237–253.
- Willms, D. J. (2003). *Ten hypotheses about socioeconomic gradients and community differences in children's developmental outcomes*. Quebec: Applied Research Branch.
- 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.

TABLE OF FIGURES

		Page
CHAPTER 1	INTRODUCTION	
Figure 1	Theoretical framework of the relationship between student achievement and its determinants	10
Figure 2	The dynamic model of educational effectiveness	17
CHAPTER 2	STUDY 1: THE ROLE OF ACADEMIC ACHIEVEMENT GROWTH IN SCHOOL TRACK RECOMMENDATIONS	
Figure 1	Math observed and fitted achievement trajectories	33
Figure 2	Initial status and math growth over Grades 4–6	34
CHAPTER 3	STUDY 2: EFFECTIVENESS MEASURES FOR CROSS-SECTIONAL STUDIES: A COMPARISON OF VALUE-ADDED MODELS AND CONTEXTUALISED ATTAINMENT MODELS	
Figure 1	Mean residuals for reading achievement status in Grade 6 and confidence intervals by schools	50
Figure 2	Mean residuals for reading achievement growth and confidence intervals by schools	50
Figure 3	Mean residuals for reading achievement status and confidence intervals by schools	51
Figure 4	Differences in ranks of unadjusted raw scores and adjusted scores in reading achievement by models (VAM/CAM) and schools	52
Figure 5	Differences in ranks of unadjusted raw scores and adjusted scores in reading achievement by models (PAM/CAM) and schools	53
CHAPTER 4	HOW EFFECTIVE ARE EDUCATIONAL SYSTEMS? A VALUE-ADDED APPROACH TO STUDY TRENDS IN PIRLS	
Figure 1	Residuals of adjusted achievement scores (i.e. effectiveness measures) in 2006 by educational system	85
Figure 2	Differences in ranks of unadjusted achievement scores and residuals of adjusted achievement scores (i.e. effectiveness measures) by educational system	86
Figure 3	Residuals of adjusted achievement change scores (i.e. effectiveness measures) from 2001 to 2006 by educational system	88
		129

Figure 4	Differences in ranks of unadjusted achievement change scores and residuals of adjusted achievement change scores (i.e. effectiveness measure) by educational system	90
Figure 5	Correlation of residuals of adjusted achievement score in 2006 and residuals of adjusted achievement change scores from 2001 to 2006	91

Table of Tables	Page
CHAPTER 2	STUDY 1: THE ROLE OF ACADEMIC ACHIEVEMENT GROWTH IN SCHOOL TRACK RECOMMENDATIONS
Table 1	Math growth models: Socioeconomic background predictors of initial status and growth (unstandardized regression coefficients)..... 32
Table 2	The relationship between socioeconomic background, academic achievement, and math performance in Grade 6 35
Table 3	The relationship between socioeconomic background, academic achievement, and the track recommendation decision 35
Table A.1	Main statistics of dependent and independent variables 37
CHAPTER 3	STUDY 2: EFFECTIVENESS MEASURES FOR CROSS-SECTIONAL STUDIES: A COMPARISON OF VALUE-ADDED MODELS AND CONTEXTUALISED ATTAINMENT MODELS
Table 1	Status (Grade 6) and growth model estimates for reading and mathematics achievement 48
Table 2	Correlations of residuals obtained from different models for reading and mathematics achievement 51
Table A1	Descriptives of the analytical and excluded sample with mean (SD), % for sex 60
Table A2	Model estimates for reading achievement 61
Table A3	Model estimates for mathematics achievement 63
CHAPTER 4	HOW EFFECTIVE ARE EDUCATIONAL SYSTEMS? A VALUE-ADDED APPROACH TO STUDY TRENDS IN PIRLS
Table 1	Missing rates of constitutive SES index variables per educational system and cohort, in percent 78
Table 2	Descriptives of average reading achievement scores and average SES measures by educational system and measurement point 79
Table 3	Estimates for reading achievement across educational systems in 2006 84
Table 4	Estimates for change in reading achievement across educational systems from 2001 to 2006 87

ANNEX

A.1 Liste der Einzelarbeiten

- Caro, D. H., Lenkeit, J., Lehmann, R., & Schwippert, K. (2009). The role of academic achievement growth in school track recommendations. *Studies in Educational Evaluation*, 35 (4), 183–192.
- Lenkeit, J. (2012). Effectiveness measures for cross-sectional studies: A comparison of value-added models and contextualised attainment models. *School Effectiveness and School Improvement*. Available at: <http://www.tandfonline.com/doi/abs/10.1080/09243453.2012.680892>
- Lenkeit, J. (im Druck). How effective are educational systems? A value-added approach to study trends in PIRLS. *JERO Journal of Educational Research Online*.

A.2 Curriculum Vitae

Entfällt aus datenschutzrechtlichen Gründen.

A.3 Liste der Publikationen und Präsentationen

Publikationen

In Bearbeitung

Lenkeit, J., Rau, A. & Jordan, A.-K. (in Bearbeitung). Reading achievement and teacher's instructional profiles in PIRLS 2006.

Angenommen

Lenkeit, J. (im Druck). How effective are educational systems? A value-added approach to study trends in PIRLS. *JERO Journal of Educational Research Online*.

2012

Lenkeit, J. (2012). Effectiveness measures for cross-sectional studies: A comparison of value-added models and contextualised attainment models. *School Effectiveness and School Improvement*. Online verfügbar unter: <http://www.tandfonline.com/doi/abs/10.1080/09243453.2012.680892>

Caro, D. H. & Lenkeit, J. (2012). An analytical approach to study educational inequalities: 10 hypothesis tests in PIRLS 2006. *International Journal of Research and Methods in Education*, 35 (1), 3–33.

Lenkeit, J., Goy, M. & Schwippert, K. (2012). The Impact of PIRLS in Germany. In K. Schwippert & J. Lenkeit (Eds.), *Progress in reading literacy in national and international context. The impact of PIRLS 2006 in 12 countries* (S. 85–105). Münster: Waxmann.

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

Schwippert, K. & Lenkeit, J. (2012). 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, Germany: Waxmann.

2011

Kuhl, P., Lenkeit, J., Pant, H. A. & Wendt, W. (2011). Die Kontextuierung von Leistungswerten bei Vergleichs- und Prüfungsarbeiten. In S. Müller, M. Pietsch & W. Bos (Hrsg.), *Schulinspektion in Deutschland. Eine Zwischenbilanz aus empirischer Sicht* (S. 237–259). Münster: Waxmann.

Liegmann, A. B., Lenkeit, J., van Ackeren, I. & Schwippert, K. (2011). Strategien im Umgang mit Befunden aus PIRLS. Eine explorativ-vergleichende Studie zur Rezeption von Befunden zu Chancengleichheit in 12 Ländern. In F. Dietrich, M. Heinrich & N. Thieme (Hrsg.), *Neue Steuerung – alte Ungleichheiten? Steuerung und Entwicklung im Bildungssystem* (S. 165–176). Münster: Waxmann.

2009

Caro, D. H., Lenkeit, J., Lehmann, R. & Schwippert, K. (2009). The role of academic achievement growth in school track recommendations. *Studies in Educational Evaluation*, 35, 183–192.

2008

- Lehmann, R. & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis - Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin. Abschlussbericht über die Untersuchungen 2003, 2004 und 2005 an Berliner Grundschulen und grundständigen Gymnasien*. Berlin: Humboldt-Universität zu Berlin.
- Schwippert, K. & Lenkeit, J. (2008). Von der Schule in den Beruf: Youth in Transition Survey - Eine kanadische Längsschnittuntersuchung. *Tertium Comparationis*, 14 (2), 154–167.

Präsentationen

2012

- Lenkeit, J.: How effective are educational systems? A value-added approach to study trends in PIRLS. Sixth Biennial Meeting der EARLI Special Interest Group 1 “Assessment and Evaluation” im August 2012, Brüssel, Belgien.
- Lenkeit, J., Rau, A., & Jordan, A.-K.: Reading achievement and teachers’ instructional profiles in PIRLS 2006. European Conference on Educational Research (ECER) der European Educational Research Association (EERA) im September 2012, Cadix, Spanien.

2011

- Lenkeit, J. (Autor & Chair): Symposium mit dem Thema “Strategies and implications of changing educational systems: Consequences of large-scale assessments”. European Conference on Educational Research (ECER) der European Educational Research Association (EERA), Berlin, Deutschland.
- Lenkeit, J., Rau, A., Jordan, A. & Schwippert, K.: Zusammenhang von Leseleistung und Merkmalen des Unterrichts aus Lehrerperspektive anhand der PIRLS 2006-Daten. Vortrag, 76. Tagung der Arbeitsgruppe für Empirisch Pädagogische Forschung (AEPF), Klagenfurt, Österreich.

2010

- Caro, D. H. & Lenkeit, J.: An analytical approach to study socioeconomic gradients: Ten hypothesis tests in PIRLS 2006. Vortrag, 4th IEA International Research Conference, Göteborg, Schweden.
- Lenkeit, J.: Effectiveness measures for cross-sectional studies: A comparison of value-added models and contextualised attainment models. Vortrag, 2nd Biennial Meeting of the EARLI Special Interest Group 18 “Educational Effectiveness: Models, Methods and Applications”, Leuven, Belgien.
- Lenkeit, J.: Effectiveness measures for cross-sectional studies: A comparison of value-added models and contextualised attainment models. Poster, IV European Congress of Methodology der European Association of Methodology (EAM), Potsdam, Deutschland.
- Liegmann, A. B., Lenkeit, J., van Ackeren, I. & Schwippert, K.: Impact of PIRLS 2006: Rezeption von Befunden aus Large Scale Assessments im internationalen Vergleich.

Vortrag, KBBB-Herbsttagung 2010 der Sektion Empirische Bildungsforschung der DGfE, Dortmund, Deutschland.

2009

Caro, D. H., Lenkeit, J., Lehmann, R. & Schwippert, K.: The role of academic achievement growth in school track recommendations. Poster, European Conference on Educational Research (ECER) der European Educational Research Association (EERA), Wien, Österreich.

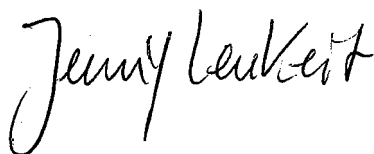
Lenkeit, J.: Fragen zum Vergleich von Gruppen - Herausforderung in nicht-experimentellen Forschungsdesigns am Beispiel einer Länderstudie. Vortrag, DGfE Summer School, Lingen, Deutschland.

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorgelegte Arbeit „Achievement status and growth as predictors of educational outcomes and effectiveness“ selbständig verfasst habe. Andere als die angegebenen Hilfsmittel habe ich nicht verwendet. Ich habe keine kommerzielle Promotionsberatung in Anspruch genommen. Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden und wird zur Veröffentlichung eingereicht.

Die angeführten Koautoren des ersten Artikels (Caro, Lenkeit, Schwippert & Lehmann, 2009) werden bestätigen, dass ich maßgeblich und entscheidend zur Konzeption und Darlegung des Artikels beigetragen habe. Mein Beitrag bezieht sich insbesondere auf die folgenden Punkte: Aufarbeitung und Darlegung des Forschungsstandes zu Übergangsentscheidungen insbesondere im deutschen Bildungssystem; Aufbereitung der ELEMENT-Daten für die durchgeführten Analysen; Interpretation und Diskussion der Ergebnisse mit besonderem Bezug zum deutschen Bildungssystem.

Hamburg,

A handwritten signature in black ink, reading 'Jenny Lenkeit'. The script is cursive and fluid, with the first name 'Jenny' and last name 'Lenkeit' clearly distinguishable.

Jenny Lenkeit