Chapter 7
The Role of Large-Scale Assessments in Research on Educational Effectiveness and School Development

Eckhard Klieme

Goals and Limitations of (International) Large-Scale Assessments

The Role of International Assessments in Educational Policymaking and Effectiveness Research

ILSAs establish a monitoring structure that provides reliable comparative information on education systems, describing system structures as well as the functioning and the productivity (i.e., the gross outcome or "yield") of education systems. The studies also contribute to our *knowledge base on educational effectiveness,* observing patterns of relationships between inputs, processes, and outcomes of education. Thus, they help to understand how educational outcomes are "produced." First, ILSAs allow for a decomposition of variation of student performance by individual, school, and system levels. Moreover, they provide data about multiple factors covering these three levels, which, according to previous research, are expected to impact student performance in specific domains like reading, mathematics, or science. In addition to describing these factors, ILSAs allow us to estimate their direct and indirect relationships to student performance and other outcomes. Statistical models, using multilevel ILSA data, help to reconstruct and understand the complex relationships between input and process factors, and how they interact in "producing" student outcomes. If data on resources and costs are available, ILSAs may also help to understand efficiency, i.e., effectiveness in relation to investments. Large representative samples allow for the generalization of findings both within and across countries.

ILSAs provide a data source for the study of educational contexts in general (e.g., how family, school, and out-of school education interact in the development of life skills). For example, Trends in International Mathematics and Science Study

E. Klieme (✉)
Center for Research on Educational Quality and Evaluation,
German Institute for International Educational Research (DIPF), Goethe University,
Schloßstraße 29, 60486 Frankfurt am Main, Germany
e-mail: klieme@dipf.de

M. von Davier et al. (eds.), *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research,*
DOI 10.1007/978-94-007-4629-9_7, © Springer Science+Business Media Dordrecht 2013

(TIMSS), Progress in International Reading Literacy Study (PIRLS), and PISA data are increasingly used by economists and social scientists to examine broader issues such as the impact of human capital on economic growth (Hanushek and Woessmann 2009, see also the chapter by Hanushek and Woessmann in this volume) or how to predict successful integration of migrant families (Stanat and Christensen 2006). The database will become even more informative once these studies move into further cycles, making trend data available that cover more than a decade.

Thus, ILSAs offer three types of "products": (1) *indicators* that monitor the functioning, productivity, and equity of education systems; (2) *knowledge* on factors that determine educational effectiveness; and, (3) a reliable, sustainable, comparative *database* that allows researchers worldwide to study scientific as well as policy-oriented questions.

Policymakers are mainly interested in No. 1. The policy relevance of this system-monitoring enterprise is based on (a) defining and operationalizing cognitive and noncognitive *outcome measures* that inform the selection and prioritization of educational goals within participating countries, (b) examining and reporting *factors that may be subject to control by policy and professional practice* (so-called malleable factors), and (c) providing *international benchmarks* that allow policymakers to ascertain what they may learn from other countries. The selection of indicators is generally guided by policy demands. Educational policymaking must deal with the functioning of the school system (i.e., operational characteristics such as resources allocated to schools), productivity (such as the gross level of student outcomes) and, last but not least, equity (e.g., how resources are distributed).

For example, several indicators based on PISA context data can be found in recent editions of the OECD's Education at a Glance reports (OECD 2007a, 2008, 2009a), such as:

- Relationship between immigrant background and student performance (2007, indicator A6);
- Profiles of top performing students, including their attitudes and motivation (2009 A4/A5);
- Relationships between resources and outcomes in education (2007 and 2008 B7), especially with regard to class size (2008 D2);
- Outcomes of vocational versus general educational programs (2007 and 2008 C1);
- Use of evaluation and assessment in education systems (2008 D5);
- Relationship between student background and access to (or motivation to participate in) higher education (2007 A4/A7, 2008 A3/A7).

Limitations of Large-Scale Assessments as School Effectiveness Research Tools

Researchers are mainly interested in the "products" described above under items 2 and 3. They tend to perceive ILSAs as multigroup (i.e., multicountry) educational

effectiveness studies. Besides *describing* strengths and challenges with regard to the students' performance and the conditions of teaching and schooling in participating countries, researchers—but to some extent also policymakers—intend to understand *why* students reach certain levels of performance.

Although the analysis of ILSA data can make important contributions to the knowledge base for educational policy and practice (see the section below on "How large-scale assessments may contribute to our knowledge of educational effectiveness and school development" for details and examples), there are limits that have to be taken into account. As Baker (2009) notes, the history of policymaking informed by international comparative studies has seen a number of short-cut conclusions, based on too simple hypotheses as to the causes of performance differences at the system level. Also, econometricians have studied a number of issues in educational productivity, but much of this work remains descriptive, rather than estimating causal effects, because data are cross-sectional, and important explanatory variables—such as cultural factors—remain unmeasured (Hanushek and Woessmann 2010).

For example, PISA is a yield study, assessing literacy and skills that have been accumulated over the lifespan, from early childhood through different levels of schooling until the age of 15 years. PISA does not ascertain how much learning has taken place in the secondary school where a student is presently enrolled. Such an assessment would require that the student's performance level was measured at the time of entering his or her present school and compared with present performance. In so doing, one would obtain a measure of progress or "value-added" in performance associated with educational experiences in the particular school. However, the PISA design does not provide any baseline measure. Teacher quality and its impact on student performance cannot be judged in PISA, either. At least, this is not feasible with the design that has been in place for over a decade. That is because a random sample of 15-year-olds is taken in each school rather than assessing intact classes, precluding the measurement of instructional strategies and effects at the classroom level. Finally, in one out of five countries that participated in PISA 2006, the majority of the students had only recently been allocated to the schools in question, prohibiting direct conclusions on school effects within these countries.

It is extremely difficult to draw causal inferences such as concluding that a particular educational policy or practice has a direct or indirect impact on student performance based on an observational survey and the kind of assessment data collected in ILSAs (Gustafsson 2007). If, for example, links were found between high student performance and rendering school evaluation data accessible to the public (as a school level policy)—as has been the case in PISA 2006—the design of the study would not allow for causal interpretation. This is because data on at least some potentially important factors, such as prior student performance, can hardly be collected in cross-sectional ILSAs. As a consequence, such potentially important factors cannot be included in the analyses. There is no way of assuring statistical control—neither by modeling the factors that predict outcomes, as in Analysis of Co-Variance (ANCOVA), nor by modeling the treatment assignment process, as in propensity score matching. The data needed for those models are simply left unob-

served in current ILSAs. Controlling for student background, such as socioeconomic status (SES), migration status, and gender—as is regularly done in ILSAs—is an inadequate substitute for baseline achievement data. Thus, currently available analyses cannot tell if the policy of making school evaluation data available to the public happens to be applied in high achieving schools, or whether the policy actually results in higher student performance.

The OECD, however, reports that, "Students in schools posting their results publicly performed 14.7 score points better than students in schools that did not, and this association remained positive even after the demographic and socioeconomic background of students and schools was accounted for" (OECD 2007b, p. 243) and concludes "that the impetus provided by external monitoring of standards, rather than relying principally on schools and individual teachers to uphold them, can make a real difference to results" (p. 276). Thus, public posting of achievement data is recommended as a strategy for school improvement. This is just one of many examples of policymakers overinterpreting available data.

The example is noteworthy because it shows that the way out of the dilemmas of causal inference recently proposed by Kröhne (2010) does not help either, at least in this case. Kröhne argued that problems with unobserved predictors arise on the individual level only, e.g., when we want to determine if participation in extracurricular activities has an effect on student learning. However, when analyzing school policies, he considered these policies to be treatments on the school level, introducing propensity score matching on the school level rather than the student level. Based on data from the German national language study DESI (see below), this procedure allowed him to conclude that so-called bilingual instruction (teaching subjects like geography in a foreign language to a certain subgroup of students within the school) had a positive school level effect on students' foreign language competencies (see Fig. 7.1). Had he done the propensity score matching on the individual level, he would have failed to catch the treatment assignment process for individual students within schools because no data were available on student achievement at the time when students were assigned to bilingual instruction. There were, however, good reasons to assume that the implementation of bilingual instruction as a school level policy can be explained from stable variables that we know or can truly estimate, like school type, school size, average parent SES, or percentage of immigrant students. Therefore, in the case of Kröhne's analysis of bilingual instruction, causal inference may be feasible. The same may be true for school policies on truancy and their effect on student absenteeism—an issue that probably will be covered in PISA 2012. However, for many other school level policies, including public reports on evaluation results, the assumption of no relation to prior achievement (both on the individual and on the school level) seems unrealistic.

The main problem with causal inferences in ILSAs is not a statistical or methodological one. The conditions for causal inference from quasiexperimental or survey-type data are well-known, based, e.g., on the Rubin-model of causality. Rather, the problem is substantial. The sociological theory of schooling as well as pedagogical concepts state that student achievement is the core of school education, i.e., the school expects students to strive for achievement, and its main "product" is student

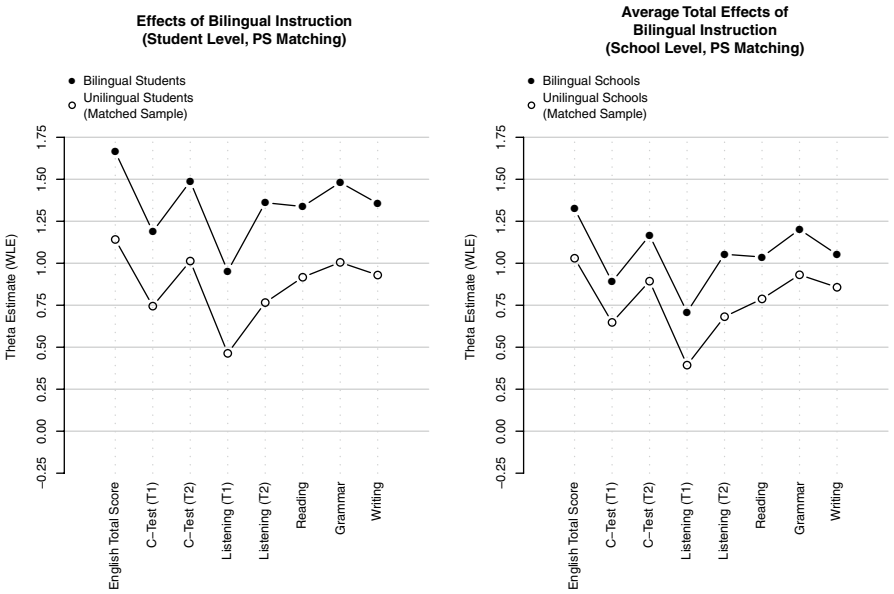

Fig. 7.1 Average total effect of bilingual instruction for eight achievements in English as a Foreign Language, estimated with propensity score matching on the student level (*left*) or on the school level (*right*); data from the German language study DESI, n = ca. 10.000 (taken with permission from Kröhne 2010)

achievement. The process of education (*Bildung* in German) can be defined as finding an appropriate individual pathway to knowledge, competency, and expertise. Pedagogical interventions (*Erziehung* in German) need to adapt to the preconditions of learning, especially to prior achievement. In their daily practice, professional educators need to monitor student achievement and change interventions accordingly. When assigning tasks, forming groups for collaborative learning, giving feedback, deciding on grade retention/promotion, and other aspects of educational careers, teachers will inevitably take students' prior achievement into account. Thus, effects of these interventions can hardly be estimated from cross-sectional data alone without knowing prior achievement, the most salient factor that drives the assignment to interventions. There might be ways to approximate prior achievement, e.g. by asking about prior grades, or more valid quasi-experimental designs, such as exploiting variation between two subjects assessed for the same group of students (Schwerdt and Wuppermann 2011). In general, it is difficult to draw causal inference in education without longitudinal achievement data.

Nevertheless, a productive interplay between ILSAs and effectiveness research may be established. ILSA studies do have an impact on educational research, even if strict causal inference cannot be assumed (see the below section on "How large scale assessments may contribute to our knowledge of educational effectiveness and school development"), and this impact can be greatly increased if the design of ILSA studies is enhanced beyond the traditional cross-sectional survey kind of

Table 7.1 Basic elements of the CIPO model of school effectiveness (adapted from Scheerens and Bosker 1997)

Input	Process	Output
Teacher-student-rate, qualification of teaching staff, student population, parent commitment	Quantity of instruction, school curriculum, leadership, teacher cooperation and collaboration, professional development, cohesion, school culture (norms and values), school climate, internal and external evaluation	*School level*
Students per class, teacher competencies	Instructional quality—opportunity to learn, clear, well-structured classroom management, supportive, student-oriented classroom climate, cognitive activation with challenging content	*Classroom level*
SES, social and cultural capital, family support, gender, language and migration background, general intellectual ability, pre-knowledge	Time invested, self regulation, motivation and interest, self concept, learning strategies	*Individual level*

Context: School structure, curriculum, pedagogical traditions and orientations, teacher education, budgeting and regulation, socio-economic and cultural context

design (see the section on "Examples of enriched (longitudinal) designs integrating LSA and EER"). The inverse impact is even more important, i.e., the impact of educational effectiveness research on ILSAs (see the below section on "How Can Educational Effectiveness Research Inform ILSAs? The PISA Design as an Example"). Before we can elaborate on these links, we have to take a closer look at educational effectiveness studies.

Goals and Research Design of Educational Effectiveness Studies

The Basic Model of School Effectiveness (CIPO-Model) and Instructional Quality

Standard models of school and teaching research conceptualize the school as a system wherein the characteristics of the context, input variables, school and instruction processes interact in "producing" student outcomes. The basic structure of this Context-Input-Process-Outcome (CIPO-) model was developed in the 1960s to inform the design of ILSAs undertaken by the IEA (Purves 1987). Addressing the multilevel-structure of the educational systems, current versions of the framework (see Table 7.1) allocate input, process, and outcome characteristics at respective levels of action (i.e., system level, school level, instruction/class/teacher level, individual level).

The main goal of educational effectiveness research (EER) is to identify "factors in teaching, curriculum, and learning environment at different levels such as the classroom, the school, and the above-school levels (that) can directly or indirectly explain the differences in the outcomes of students, taking into account background characteristics, such as ability, socioeconomic status, and prior attainment" (Creemers and Kyriakides 2008, p. 12).

Taken literally, this definition includes ILSA, as these studies also intend to explain differences in student outcomes, taking into account a broad array of variables from all cells of the CIPO matrix. In fact, "educational effectiveness" has become an umbrella for quite a fuzzy set of studies, from surveys unveiling general characteristics of schools (e.g., leadership, trust (Bryk et al. 2010), and reliability (Teddlie and Stringfield 1993)) to experimental studies identifying effects of specific instructional interventions (e.g., training of self-regulation, peer learning, reading programs). In order to face current challenges (see the section on "Challenges to the EER paradigm" below), more sophisticated designs are needed (see the section on "Examples of enriched (longitudinal) designs integrating LSA and EER" below), including longitudinal data collection and experimental or quasiexperimental assignment to treatments, accompanied by more complex methods (which I will not focus on here) and more substantial theory.

Let me illustrate the need for more sophisticated theoretical and empirical work by just one cell in the CIPO matrix, namely classroom level processes, i.e., instruction—mainly because this cell will play a major role in subsequent examples.

In the early tradition of behaviorist psychology, the *time needed to achieve certain learning goals* was supposed to be a major criterion for instructional effectiveness. Following Carroll (1963), numerous studies have shown that learning time is a major predictor of student outcomes in many subjects. Accordingly, the notion of *opportunity to learn*, introduced by John Carroll in the early 1960s, was initially meant to indicate whether students had sufficient time and received adequate instruction to learn (Carroll 1963; cf. Abedi et al. 2006). *Quality of instruction* was operationalized as the reduction of learning time reached in a specific instructional setting, compared to a standard setting.

The notion of opportunity to learn (OTL) has since become an important concept in international student assessments (Husén 1967, 1974; Schmidt and McKnight 1995; Schmidt et al. 2001), and it was shown to be strongly associated with student performance, especially in cross-country comparisons (Schmidt and Maier 2009, pp. 552–556). At the same time, the construct received a much broader meaning. Stevens (1993, pp. 233–234) already identified four kinds of OTL variables most prevalent in research:

- Content coverage variables: These measure whether instruction covers the curriculum for a particular grade level or subject matter.
- Content exposure variables: These take into consideration the time allowed for and devoted to instruction (time on task) and the depth of the teaching provided.
- Content emphasis variables: These describe which topics within the curriculum are selected for emphasis and which students are selected to receive instruction

emphasizing lower order skills (i.e., rote memorization) or higher order skills (i.e., critical problem solving).
- Quality of instructional delivery variables: These reveal how classroom teaching practices (i.e., structuring of lessons) affect students' academic performance.

Thus, for certain authors, OTL has become more or less a synonym for the quality of instruction experienced by the student. Schmidt and Maier (2009), however, in their review argue that OTL is a rather uncomplicated concept: "What students learn in school is related to what is taught" (p. 541), and they intentionally focus on OTL "in the narrowest sense: Student's content exposure" (p. 542).

Schmidt and Maier acknowledge that although OTL may be a straightforward construct, it is difficult to measure. In order to explain differences in the achieved curriculum, teachers and/or students have traditionally been asked whether and how certain curricular content has been realized in instruction (the implemented curriculum), sometimes using logs (Rowan et al. 2004). In addition, curriculum experts have been asked whether and how content elements have been covered within curricular documents like syllabuses, textbooks, and standards (the intended curriculum). From these raw data, various indicators have been extracted. In many cases, the content taught has been judged twofold, in terms of topic and level of demand, while at the system level, indices for coherence, rigor, and focus have been derived (Schmidt and Maier 2009).

In addition to OTL as described above, a number of other processes at the classroom level have been found to be relevant for educational effectiveness (Creemers and Kyriakides 2008; Harris and Chrispeels 2006; Hopkins 2005; Scheerens and Bosker 1997). Well-structured lessons with close monitoring, adequate pacing and classroom management, clarity of presentation, and informative and encouraging feedback (i.e., the key aspects of "direct instruction") are positively linked to student performance. These components help create an orderly classroom environment and maximize effective learning time. Yet student motivation and noncognitive outcomes benefit from additional characteristics of instructional quality, such as a classroom climate and teacher–student relations that support student autonomy, competency and social relatedness (Deci and Ryan 1985). Furthermore, in order to foster conceptual understanding, instruction has to use challenging content (Brown 1994). Also, different student subpopulations may benefit from different instructional practices. Thus, teachers have to orchestrate learning activities in a way that serves the needs of their specific class. Klieme et al. (2009) condensed this knowledge into a framework of three "basic dimensions of instructional quality": (a) clear, well-structured classroom management, (b) supportive, student-oriented classroom climate, and (c) cognitive activation with challenging content. Several independent studies of secondary school mathematics education have since confirmed this triarchic structure of instructional quality and given some support for the cognitive and motivational impact that was hypothesized (*TIMSS-Video*: Klieme et al. 2001; *COACTIV*: Baumert et al. 2009; *Pythagoras*: Lipowsky et al. 2009). Klieme and Rakoczy (2003) as well as Kunter et al. (2008) identified similar structures within national extensions to PISA. The triarchic model is also revealed in observational

data on elementary and primary education in the United States (Pianta and Hamre 2009) as well as in the Ohio teacher efficacy scales (OSTES) developed by Tschannen-Moran and Woolfolk Hoy (2001).

Challenges to the EER Paradigm

The paradigm of EER faces a number of severe theoretical and empirical challenges when conceptualizing and operationalizing the general model in more detail. The main challenges seem to be:

(a) The adaptive nature of educational processes: Practices may neither be equally effective for all students within a school nor for all education systems, local contexts, and schools. Moreover, depending on the kind of outcomes emphasized, different conclusions may be drawn (Kyriakides and Tsangaridou 2004). Hence, modern research into educational effectiveness also takes interactions with input into account and examines differential effectiveness and adaptive practices. A considerable amount of research has been carried out in this field (e.g., Creemers and Kyriakides 2008; Scheerens 2000; Teddlie and Reynolds 2000).

(b) The dynamic nature of educational processes: When turned into a dynamic model of school effectiveness (see Creemers and Kyriakides 2008), outcomes become inputs for further development. Mathematics anxiety, for example, can be an outcome of schooling as well as an input—impacting, for instance, students' homework activities. Moreover, inputs may have reciprocal mutual effects. For example, a school's socioeconomic composition in many education systems is correlated with funding, parental involvement, or even teacher quality. This, in turn, allows for other (better) teaching-learning environments to be offered, which attract students (or, rather, parents) from higher socioeconomic backgrounds, so that, in the end, social stratification, resources, and process quality are mixed and are difficult to disentangle (see "Examples of enriched (longitudinal) designs integrating LSA and EER" below for empirical results on that topic).

(c) The complexity of mediating processes: It is reasonable to assume that not all effects on student outcomes are direct. Comparatively weaker effects on student outcomes are often found for policies at the school and system level, as compared to student background variables and classroom processes (e.g., Wang et al. 1993). This may, in part, be because the former variables do not exert a direct effect on students, but are rather related to school or classroom processes, which in turn have an effect on student performance. Moreover, school level variables such as school climate, shared values and norms, or procedures to deal with behavioral problems, may have a direct effect on noncognitive outcomes (e.g., learning motivation, academic aspirations) and student behavior (e.g., truancy, violence), while school effects on student performance and other subject-related outcomes (e.g., interest and self-efficacy beliefs) probably will be mediated by teaching and learning within classrooms.

(d) The importance of moderating variables: Based on a constructivist understanding of student learning, current educational theory assumes that student learning is largely dependent on self-regulated processes, which are moderated by school, classroom, and teacher factors. Modeling such differences requires the examination of interaction/moderation effects. Contemporary research findings indicate that the relevance of school characteristics does not remain consistent across subjects and classes, and varies according to the constellation of a student population (Ditton and Krecker 1995; Luyten and de Jong 1998; Sammons et al. 1997; Scheerens and Bosker 1997). In line with the theory of differential effectiveness (e.g., Kyriakides and Tsangaridou 2004), it is important to acknowledge that relationships between variables may not be similar in different subgroups. For example, there is some evidence that students from diverse social backgrounds may benefit from different instructional techniques (e.g., Brophy 1992; Walberg 1986).

(e) The weakness of distal effects, especially school effects: Within the multilevel CIPO model, "effects" are usually supposed to cascade from the upper to the lower levels. However, meta-analyses of school and instructional effectiveness that are grounded in this model (Hattie 2009; Seidel and Shavelson 2007; Scheerens and Bosker 1997; Wang et al. 1993) force us to acknowledge that prerequisites of learning and individual activities bear more significance to the students' learning results than the characteristics and processes of instruction, whereas instruction and teacher competencies, in turn, bear more significance to student outcomes than school level factors. School effectiveness research thus concludes that learning conditions, norms, and practices at the school level do provide a framework for learning and teaching processes, but they are more distant to the students' learning achievement and thus bear less predictive power than the teaching and learning activities in the classroom (Creemers 1994; Ditton 2000, 2007; Fend 1998; Sammons 1999; Slavin 1996; Stringfield 1994). This view is supported by cognitive models of learning and teaching, which do not define instruction as an "immediately effective" measure, but rather as social interactions and learning opportunities that the students use for acquiring competencies, pursuant to their individual abilities and preexisting knowledge. Hence, individual learning activities are considered more meaningful for acquiring competencies than classroom instructional characteristics, and even more so compared to school process characteristics (Seidel and Shavelson 2007).

(f) The fundamental difference between status (at a given moment) and individual growth or organizational change (over time): Individual growth and organizational change (i.e., longitudinal outcomes of education) have to be studied distinctively, because explaining and predicting change is quite different from explaining and predicting levels of outcomes in cross-sectional comparison.

(g) While a vast body of evidence exists from English-speaking countries and the Netherlands regarding characteristics of effective schools, which have been retrospectively gained from analyzing high achieving schools (see Sammons et al. 1997) and school effectiveness studies (Scheerens and Bosker 1997), sound assessments of school developments are lacking from a longitudinal perspec-

tive. International surveys on school improvement research have been published in recent years (Lee and Williams 2006; Hopkins 2005; Harris and Chrispeels 2006) but they can merely report on case study effects or repeat the well-known meta-analyses of school effectiveness research; a longitudinal assessment involving schools as units of observation, objective criteria measures, and reliable sample sizes scarcely has been realized so far.

American school research in the 1970s and early 1980s brought processes of school development forward to large questionnaire-based studies, the Rand Change Agent Study and the DESSI study (Dissemination Efforts Supporting School Improvement) and thus highlighted these processes without being able to evaluate their effectiveness. It thus became clear that it is impossible to plan and predict school development in a harmonized way, but that it is locally adjusted, with "ownership" of the staff (which is an important condition for sustainable change) resulting from experiencing practical success (Teddlie and Stringfield 2006, p. 26 f.). From the late 1980s onwards, the principle of treating individual schools as units of action ("site-based management") also brought changes to research: for instance, Teddlie and Stringfield (1993) observed 16 schools over a period of 10 years, developing the concept of reliable schools. At present, longitudinal analyses are conducted on effects of "comprehensive school reform" (see the overview in Borman et al. 2003). These analyses are mainly based on school statistical data and standardized achievement tests. It is thus possible to determine whether schools participating in specific reform programs differ from other schools regarding the development of achievement. In some cases, recognizable effect sizes are reported. However, this line of research reveals hardly anything about processes and conditioning factors of school development.

(h) The incoherence and instability of effect sizes: According to Scheerens and Bosker (1997, p. 81), stability and consistency of school effects are "one of the most fundamental issues in school effectiveness research," but one that has been widely neglected. Current accountability policies are based on strong but questionable assumptions: that student achievement can legitimately be attributed to school (as opposed to teacher or department, for example) effects; that we can measure progress on the school level in a reliable manner; and that *change* in school-level effects is an indicator for successful school improvement, not an artifact due to unreliability and instability (see Goldschmidt et al. 2004 for statistical models that allow testing of these assumptions). Most recently, Bryk et al. (2010) took the analysis of school development a huge step forward, presenting complex data records from the evaluation of the Chicago school reform. But in this latter case, the indicators were rather descriptive.

(i) Early work by Willms and Raudenbush (1989) indicated that the overall achievement level of a school is remarkably consistent, has been challenged by British researchers. For example, Thomas et al. 1997, p. 194, state that "only a minority of schools performed both consistently (across subjects) and with consistency (over time) and … these schools are at the extremes of the effectiveness range (i.e., strongly positive or strongly negative)." But those cases of well-performing or failing schools, these authors argued, can be understood considering our

school effectiveness knowledge base. They hypothesized that high achievement expectations, a shared vision, strong and flexible leadership, high quality instruction, and strong parental involvement are among the factors that support positive school development in a longitude (Sammons et al. 1995, p. 93). These hypotheses were confirmed in part by a followup study using interviews with school headmasters.

Problems (f) and (g) will be illustrated in the following section with an example from a German school survey.

Comparing Value-Added Status, Growth, and Change Indicators for Schools: An Empirical Study

Klieme et al. (2010b) evaluated extracurricular activities in some 230 lower secondary schools all over Germany, using a multicohort longitudinal design. As a global measure of language competency, standard vocabulary tests were administered three times, in 2005, 2007, and 2009. Each time, students from grades 5, 7, and 9 participated, allowing for identification of individual growth over a two- or even four-year period for most of the students. All data were standardized within age groups. Also, student background information (gender, socioeconomic status, and migration status) is available. Thus, on the school level, different indicators for school quality can be derived:

(a) Based on data from the most recent wave of measurement, 2009, achievement scores can be calculated and adjusted for student background variables. The adjusted test score, aggregated for the school, can be used as a proximal indicator for the school's added value. This indicator represents the kind of data that would be available in a purely cross-sectional survey such as traditional ILSAs. This indicator turns out to possess stability—calculated as the correlation coefficient for $n=232$ schools—of about 70, which indicates the school results are relatively stable in Germany.
(b) For those students who were observed twice, a difference score can be calculated, describing the relative gain (or loss) in achievement between 2007 and 2009 relative to the respective age groups. Aggregated on the school level, this indicator measures "achievement gain" over two years.
A similar indicator can be derived for the period between 2005 and 2007. Both indicators correlate significantly, but only at $r=0.305$, indicating limited stability of this effect.
(c) An even more complex indicator can be calculated as the difference between the mean growth rate 2005–2007 and the mean growth rate 2007–2009. We consider this as an indicator for change in value-added of individual schools, i.e., as a statistical aspect that may reflect effects of organizational change.

As can be seen from the plots in Fig. 7.2, (a) and (b) correlate moderately ($r=0.39$, $p<0.001$), while the change indicator (c) is uncorrelated to both (a) and (b).

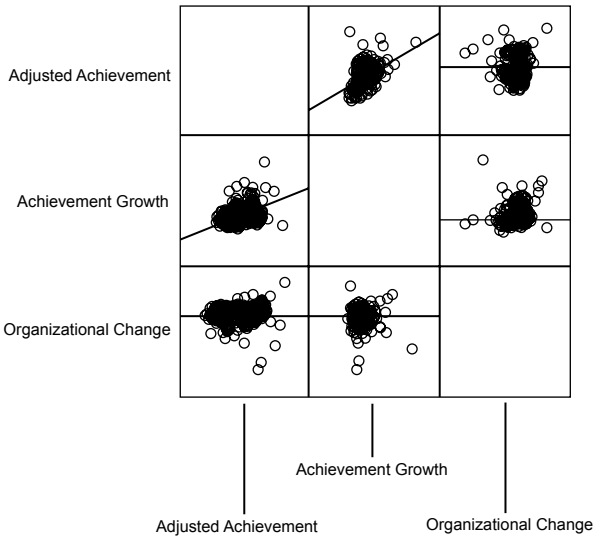

Fig. 7.2 Relationships between three indicators of school effects in $n = 232$ German secondary schools

This finding illustrates that studies on school effects that can only use cross-sectional data will trigger misinterpretations—even if student background characteristics are controlled for. Cross-sectional estimates of "value-added" are weakly correlated or even uncorrelated to indicators that actually cover growth and change. Findings from cross-sectional studies should not be interpreted as explanations of school development.

How Can Educational Effectiveness Research Inform ILSAs? The PISA Design as an Example

One of the consequences of the issues raised here is that rather than being a sound foundation for educational effectiveness research, cross-sectional ILSAs depend on input from EER studies and theories. Factors that have been demonstrated to be relevant for educational effectiveness or efficiency in the research literature are premier candidates for continuous monitoring within ILSAs and for incorporation into the broader system of educational indicators.

For example, a recent version of the CIPO model, as shown in Table 7.2, covers practically all constructs that have been suggested for inclusion in the design of background questionnaires in the PISA 2012 study (Klieme et al. 2010a). The first column displays four levels: students, classrooms, schools, and countries. The three production phases are then given in the remaining columns, i.e., inputs, processes and outcomes, respectively. As the major achievement domain in PISA 2012 will be mathematics, the table focuses on student outcomes in this domain.

The choice of constructs in ILSAs is based on a combination of policy priorities and research evidence. Policymakers on the PISA Governing Board decide upon

Table 7.2 Overview of constructs covered by PISA 2012 (from Klieme et al. 2010a)

	Input	Processes	Outcomes
Students	Gender, grade level, socioeconomic status	Attendance/truancy	Mathematical literacy
	Educational career, grades	Outside-class activities—e.g., participation in after school programs	Mathematics-related attitudes, beliefs and motivation
	Immigration background Family environment and support	Motivation, engagement	General school-related attitudes and behavior, e.g., commitment, truancy
	ICT experience, attitudes, skills	Learning and thinking strategies, test taking strategies	Learning motivation, educational aspirations
	Openness, problem solving styles	Learning time (including homework and private tuition)	
Classrooms	Class size, socio-economic background and ethnic composition	Quality of instruction: structure, support, challenge	Aggregated student variables
	Teacher education/training, expertise	Opportunity to learn: implemented curriculum, assigned tasks, mathematics-related activities	
		Instructional time, grouping, assessment and feedback	
Schools	Socioeconomic background and ethnic composition Affluence of the community	Achievement orientation, shared norms, leadership, teacher morale and cooperation, professional development	Aggregated student variables
	School funding, public vs. private	Admission and recruitment policies, tracking, course offerings/school curriculum, evaluation	Promotion/retention and graduation rates
	School size		
	Parental involvement	Teacher-student relations, supportive environment	Attendance
Countries (Systems)	Economic wealth Social (in)equality	School funding, tracking and allocation, policies for professional teacher development, support for special needs and language minority students, hiring and certification policies	Aggregated student variables
	Diversity policies	Accountability and evaluation policies, locus of decision-making	Average graduation level

the goals and research questions, while experts, building on extensive knowledge in EER, choose the appropriate constructs, instruments, and variables. For example, the definition of "mathematical literacy" as the most important outcome variable, and the decision to include mathematics-related attitudes and relations as outcome variables, are both based on policy decisions, reflecting general curriculum goals, and goals of the educational system shared by most participating countries. The constructs we use, however, and how these are operationalized, mainly reflect insights gained from research literature. Also, input and process variables are included if there is strong research evidence that they have an impact on the outcomes.

Some input factors are fairly stable and difficult to change, while others can be shaped by school development activities or policy decisions. Processes are usually more malleable, at least indirectly (e.g., by teacher education and professional development), and outcomes reflect the effects of the inputs and processes. Note, however, that the discrimination between the three strands of variables is by no means clear cut: Outcomes from one educational setting become an input for the next, while some process aspects (e.g., learning strategies) may well be treated as either input or outcome, depending on a given theoretical perspective, research design, or practical considerations.

As PISA is a trend study, assessing the same set of achievement domains every three years, it is crucial to define a core of variables that will be kept constant. Only if trend variables are kept unchanged—or moderately edited, leaving at least some anchor items unchanged—can policymakers and researchers be informed about change on the system level. Once again, the selection of constructs and variables is based on a combination of policy arguments and input from research studies. The PISA 2012 Questionnaire Framework (Klieme et al. 2010a) suggests the following design structure:

1. **General (i.e., domain-independent) trend variables**

   General input variables:

   - Student level inputs (grade; gender; socioeconomic background: parental education and occupation/family wealth/educational resources/cultural possessions; migration data: immigration status/heritage language/age on arrival in country; family support)
   - School level contexts and inputs (community size, resources, qualifications of teaching staff)

   General process variables:

   - School level processes (decision-making, admission policies, assessment and evaluation policies, professional development, student-teacher-relations, parental involvement)
   - Instructional processes (learning time, disciplinary climate, teacher support)

   General outcome variables:

   - General noncognitive outcomes—Commitment to learning (behavioral: truancy; personal goal: educational aspirations; motivational: learning engagement, affective: sense of belonging)

2. **Domain-specific trend variables**
   - Domain-specific cognitive outcomes (math, science, reading literacy)
   - Domain-specific noncognitive outcome variables (strategies and metacognition, domain-related beliefs, self-related beliefs, motivation)
   - Domain-specific process variables (opportunity to learn, instructional quality, system and school level support)
3. **Thematic extension variables (extensions within individual cycles)**
   - International options (e.g., in PISA 2012, educational career/second language learners; information and computer technology (ICT) literacy)
   - Context variables for additional domains (e.g., ICT-related experiences relevant for computer-based problem solving)
   - Descriptive and explanatory variables for specific reports (e.g., in PISA 2012: mathematics-related motivations and intentions)
   - Malleable variables at the school level (e.g., in PISA 2012: truancy policies) that are specifically selected for descriptive purposes or for causal inference
4. **System level data, gained from the OECD's international system of indicators, or from a system-level questionnaire**
   - Output of educational institutions (e.g., certificates)
   - Financial and human resources invested into education
   - Access to and participation in education
   - Learning environment and organization of schools

How ILSAs May Contribute to Our Knowledge of Educational Effectiveness and School Development

Much of the value of ILSAs is based on a constant interplay between assessments such as PISA as a monitoring survey and more rigorous kinds of effectiveness research done elsewhere. As shown before, factors that have been demonstrated to be relevant for educational effectiveness or efficiency in the research literature are prime candidates for continuous monitoring and for incorporation into the OECD system of educational indicators. In the following, the inverse kind of link will be discussed. Even while causal inferences are not warranted, ILSA data can be put to substantial use for gaining insights in educational effectiveness: (1) Correlational and other exploratory results from ILSAs may lead to hypotheses that can subsequently be tested in more robust designs, namely longitudinal, experimental, or intervention studies. As an example, the next section discusses the German TIMSS video study, which led to the formulation of the triarchic theory of instructional quality; (2) Hypotheses from EER can be tested in ILSAs, making use of broad, representative samples, high participation rates, and good measurement quality. In presenting results of such tests, our theory of instructional quality is again referred

to. (3) Last but not least, ILSAs allow for checking the cross-cultural and cross-national validity of EER findings.

ILSAs as a Means of Exploration and Hypothesis Generation: Findings from TIMSS and PISA

The TIMSS 1995 video study, an add-on to the international ILSA in grade 8, had a huge impact on instructional research in the United States (Stigler and Hiebert 1999) and in Germany (Baumert et al. 1997, Kunter et al. 2006), the two countries that participated along with Japan. Compared to Japan, with its strong focus on high level thinking, especially in the areas of geometry, open-ended problem solving, and a choreography that included extended seat work and group work as well as teacher lecturing, instruction in both Germany and the United States looked rather narrow. The instructional "script" found in Japanese classrooms was understood by many to be the cause for the high level of mathematics achievement that TIMSS as well as previous IEA studies and—later—the OECD PISA studies found in that country. However, as there was no overlap between the TIMSS video samples and the TIMSS assessment samples in Japan and the United States, this hypothesis could not be tested within the video study itself. Later, the 1998 TIMSS video study, which included another five high-achieving countries (Korea, the Czech Republic, the Netherlands, Switzerland, and Australia), would show that high achieving countries had quite different profiles in teaching practices, devaluating any attempt at directly linking student achievement to teaching practices on a national level (Hiebert et al. 2003; Pauli and Reusser 2006).

Within country, between-classrooms differences could be studied in depth for Germany, where TIMSS achievement tests had been implemented in the 1995 video sample, and a broad range of student and teacher questionnaire scales had been added, including a longitudinal followup one year later. Also, a number of high-inference video ratings were performed (Clausen 2002). Three basic (second-order) dimensions of instructional quality were identified in these ratings and shown to have specific effects on the classroom level, as seen in Table 7.3: (1) student-oriented, supportive climate and practices were related to positive development of student motivation; (2) so-called cognitive activation (e.g., Socratic deep-level questioning, use of complex problems) was related to achievement growth; (3) efficient classroom management with low level of disruptive student behavior seemed to underlie both (Klieme et al. 2001). Effects were quite small, but in subsequent ILSAs, namely PISA 2000 (Klieme and Rakoczy 2003) and PISA 2003 (Kunter et al. 2006), the basic pattern could be reproduced. Thus, ILSA studies served as the foundation for theory development, which was of course later augmented with arguments from educational and psychological research (see Klieme et al. 2009).

Hypotheses generated from ILSA data may later be tested in (quasi)experimental and/or longitudinal designs, as has been the case in the "Pythagoras" study on instructional quality (Klieme et al. 2010). This study, conducted in 2003/2004 in

Table 7.3 Second order factors of classroom practice based on high-inference video-ratings (TIMSS-Video 1994 Germany: national sample, 100 lessons; see Klieme et al. 2001)

Classroom management	Supportive climate	Cognitive activation
Effective treatment of interruptions: "teacher intervenes immediately, before disturbance may evolve"	Social orientation: "teacher takes care of her/his students' problems"	Teacher's ability to motivate students: "can present even abstract content in an interesting manner"
Clarity of rules • Interruptions (–) • Waste of time (–) • Monitoring • Time on task • Teacher unreliability (–)	Teachers' diagnostic competency with regard to social behavior	
Clarity and structuredness of the Instruction	*Individual reference norm in evaluation* • Rate of interaction (–) • Pressure on students (–)	*Errors as opportunities* *Demanding tasks* • Practicing by repetition (–)

Switzerland and Germany, adapted many design elements, techniques, and procedures from the TIMSS video studies. However, the content of the lessons to be videotaped was controlled for: all participating classes were filmed during their first three lessons of introduction to the Pythagorean Theorem. Instructional approaches were controlled to some extent, too: teachers were asked to do a proof (of any kind) during the lessons. The content focus set by design could be used to develop and implement tailored assessments and questionnaires that directly addressed teaching and learning within the lessons that had been taped.

The sample consisted of 20 Swiss and 20 German classes from two secondary school types. Because participation was voluntary, the sample is not representative. The analyses draw on data from a maximum of 1,015 students in the ninth grade (Germany) or the eighth grade (German-speaking part of Switzerland).

In addition to video ratings, student ratings of instructional quality were implemented to test the triarchic theory. In fact, all three dimensions of instructional quality could be assessed by student questionnaires and were shown to be highly predictive of general achievement growth over the school year. Student ratings for (a) structure, (b) teacher support, and (c) process-oriented approach to homework, as an indicator of cognitive activation, all correlate highly (0.47–0.52) with changes in achievement on the class level.

ILSAs as a Means of Testing Hypotheses from EER: Findings from PISA

Our theory of instructional quality predicts that classroom management has a strong positive correlation with student achievement, while supportive climate would be related to student motivation. These hypotheses were tested with the international

PISA 2000 data set. A three-level hierarchical regression model was specified, involving individual, school, and country level predictors. The International Socio-Economic Index of Occupational Status (ISEI)—more precisely, the maximum of mothers ISEI and fathers ISEI called HISEI—was used as a control variable, and two scales from the PISA student questionnaire were used as predictors at the individual level, while their aggregated analogues were used as predictors on the school and the system level. The model was run twice: once with reading literacy, a cognitive variable, as the dependent, and once with interest, an affective variable, as the dependent variable (Fig. 7.3).

ILSAs as a Means of Understanding the Systemic and Cultural Context of Education and How It Moderates EER Results: Findings from TALIS and PISA

A behavior-oriented version of the triarchic model of instructional quality was implemented in the OECD TALIS study by asking teachers how often they implemented each of 13 given practices in their teaching:

- Structuring practices (5 items): e.g., "I explicitly state learning goals." Other items include summary of former lessons, homework review, checking the exercise book, and checking student understanding during classroom talk by questioning students.
- Student-oriented practices (4 items): e.g., "Students work in small groups to come up with a joint solution to a problem or task." Other items include ability grouping, student self-evaluation, and student participation in classroom planning.
- Enhanced activities (4 items): e.g., "Students work on projects that require at least one week to complete." Other items include making a product, writing an essay, and debating arguments.

Based on TALIS main study data from 23 countries, it has been shown that (a) the three dimensions can be differentiated across countries (i.e., the triarchic model has some cross-cultural validity), (b) structuring practices, as hypothesized, are associated with higher levels of classroom discipline (as perceived by teachers), and (c) participation in professional development as well as teaching high-ability classes is correlated with a higher frequency of using these practices. Mathematics and science teachers report less student orientation and less frequent use of enhanced activities than teachers of other subjects (Klieme and Vieluf 2009).

Quite often, questionnaire scales show strange behavior when individual, school, and country level relations are compared. Especially for self-reported Likert-type questions, a number of negative correlations with student achievement have been found on the country level, although on the individual level, the correlation is positive. This kind of reversion of a correlation, when considering the aggregated level of states rather than the familiar individual level, can often be found in ILSA data records. Explanations so far mostly refer to culture-specific

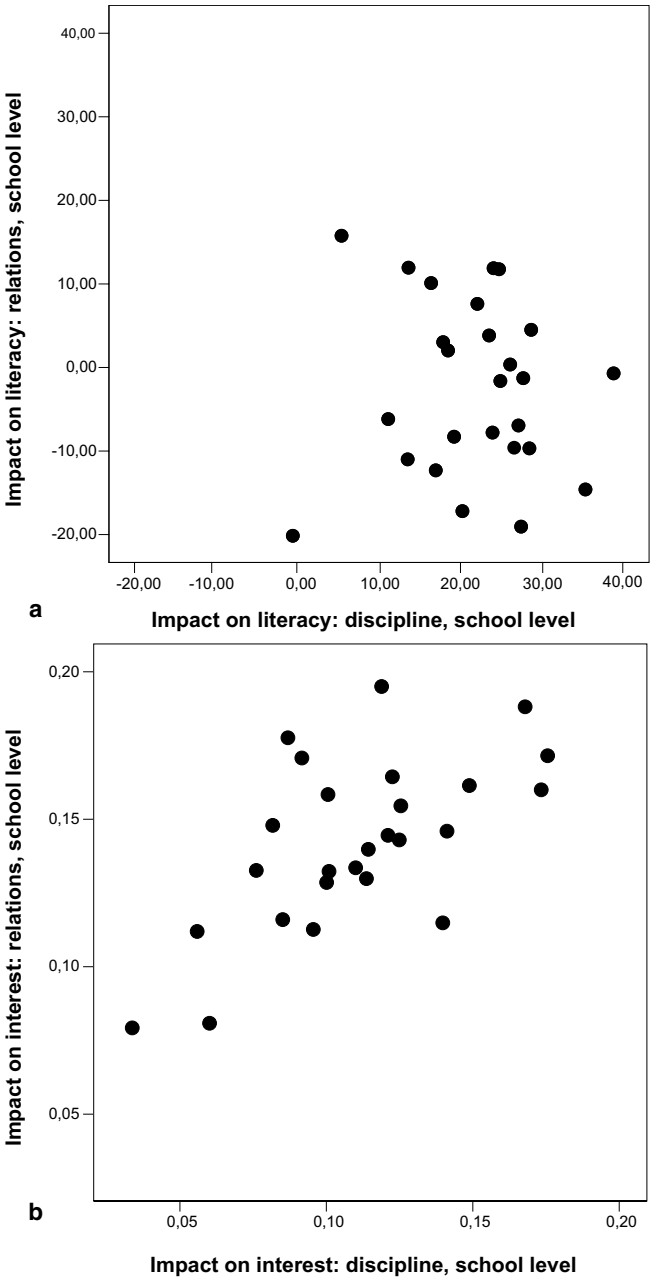

Fig. 7.3 **a** Effects of perceived classroom discipline and perceived quality of teacher-students-relation on reading literacy. *Each dot* represents one participating country. For each country, the graph shows the country-specific school level parameters. Apparently, effect sizes are larger for disciplinary climate than teacher-student-relations, as predicted. **b** Effects of perceived classroom discipline and perceived quality of teacher-students-relation on reading interest. *Each dot* represents one participating country. For each country, the graph shows the country-specific school level parameters. Apparently, effect sizes are larger for teacher-student-relations than for disciplinary climate

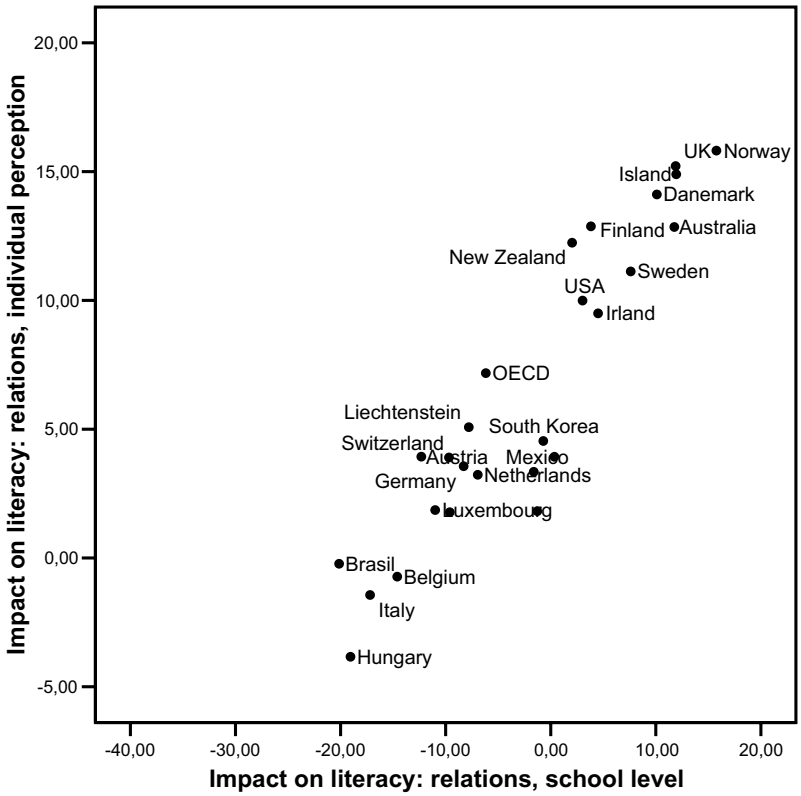

Fig. 7.4 School level effect (*horizontal*) and individual level effect (*vertical*) of perceived teacher-student-relations on reading literacy for various countries; data from a three-level hierarchical regression analysis

response styles. Figure 7.4 shows that in some countries this phenomenon occurs on the school level as well.

The countries depicted in Fig. 7.4 clearly fall into two distinctive categories: In systems with strong and early tracking, such as the German-speaking countries as well as Hungary, Italy, and some non-European countries, the effect is negative on the between-school level, and close to zero on the individual, within-school level. Both effects can be interpreted referring to selection and framing processes that typically operate within those tracked systems: Students are allocated to secondary school types according to their overall achievement level. In low track schools, teachers tend to be more supportive and less demanding. This is clearly reflected in student perceptions, causing the negative correlation on the school level. Within schools, however, variation in student perceptions is quite small because of the selection process; therefore, correlation is about zero.

In systems with less tracking, however, as in Nordic and Anglo-Saxon countries, both between-school and within-school parameters are clearly positive. In those

Table 7.4 Design enhancements to PISA (national options) in Germany

Level of analysis	Cross-section	Longitude
Students	Individual competency level	*Individual learning progress (PISA Germany 2003/2004)*
Classroom	*Competency levels of school classes*	*Instructional effectiveness (PISA Germany 2003/2004)*
Schools	Mean competency level of schools	*School development (PISA Germany school panel 2000/2009)*
School systems (states)	"Output" of educational systems	State trends

systems, schools are more equal, so that each school has a relatively wide range of achievement levels as well as a wide range of student perceptions, allowing for higher correlations.

The example shows how effects measured between and within schools are shaped by system characteristics.

Examples of Enriched (Longitudinal) Designs Integrating ILSAs and EER

Furthermore, the ILSA design may be enhanced by oversampling as well as adding additional instruments, allowing for quasiexperimental add-ons and for longitudinal studies on the school and/or the individual level. For example, several such enhancements have been implemented as national options for the PISA studies in Germany (See Table 7.4).

Two examples of such enhancements will be presented: (1) a national large-scale assessment (NLSA) study on language competencies in Germany (DESI) reassessed students one year after the first NLSA allowed studying the impact of school level factors on classroom instruction and student growth. (2) PISA/Germany 2009 reassessed schools nine years after their first participation in that NLSA, allowing the study of school development over nearly a decade.

Longitudinal Design on the Individual Level: The German National Study on Language Learning, DESI

This section reports on a representative study of language development in ninth grade, $n = 209$ schools, 1,579 teachers, 9,980 students. Hierarchical linear modeling (HLM) specifies the impact of school level (achievement orientation, strength of competency goals, cooperation among German language teachers) and classroom level factors (structure, teacher support, cognitive challenge, frequency of opportunities for language learning). Drawing on a school achievement study that is rep-

resentative for Germany regarding the subjects of German and English (DESI; see Klieme et al. 2008), we assess how far differences in the development of achievement in terms of language awareness and learning motivation in the subject of German in ninth-year students can be explained by differences in the school norms and teaching practices and by differences in the norms and practices in German instruction. We also assess what pattern of relations can be identified between the school and classroom instructional characteristics. In each school, two classes from grade 9 were assessed. Data were analyzed with a series of three-level models, allowing for an analytical dissection of school, class, and individual levels.[1]

For teaching German to ninth-year students, we intend to assess how far differences in the development of achievement in the area of "language awareness" and motivation to learn can be explained by school norms and practices among the teaching staff and by characteristics of German instruction quality. The DESI subtest on language awareness is used here because it bears the best measurement characteristics of all tests applied to German lessons, and because the pertinent demands on German lessons can be measured comparatively well by surveying student perceptions.

At the level of (German) lessons, we once again identified three basic dimensions of instructional quality:

1. clear, well-structured teaching, (structuredness)
2. a supportive learning climate that is oriented towards the students (teacher support), and
3. challenging, cognitively activating demands (cognitive challenges)

However, only student self report of perceptions during lessons were available. In contrast to expert coded videotaped lessons, student self reports are limited particularly in assessing the third quality dimension. We administered a questionnaire scale regarding the perceived importance of correct language use, which should be able to model high demands on the achievement criterion of "language awareness." We also took into account a fourth scale for questions regarding the frequency of language-related learning opportunities in the classroom.

Following the learning and teaching theory assumptions outlined above, we expect supportive teacher behavior to be crucial to the development of motivation, while cognitively challenging lesson design is important for achievement development. Both of these criteria are likely to be positively influenced by well-structured instruction. Contrary to the three basic dimensions, the fourth scale pertinent to the frequency of learning opportunities in the field of language awareness constitutes a "surface characteristic" of methodological-didactic design, and we do not expect this scale to bear an effect on learning development.

The following predictors are applied at the school level: achievement expectations of the German teachers, norms that are shared among the German teaching staff (here: the relevance of language competency goals) as well as cooperation among the German teaching staff. These aspects of professional action among

[1] These analyses have first been published in German by Klieme et al. 2010c.

colleagues are generally assumed to influence the quality of instruction and also cognitive and motivational learning processes. We can specifically anticipate high expectations of achievement and respective competency goals of teachers to lead to more challenging lessons, thus mediating the improved development in achievement.

First, we are looking for effects of school level processes on instructional quality, as perceived by students (Table 7.5). Considering the model with control variables (model II), the following picture emerges: Explanation of perceived instruction is least successful for the surface characteristic of "learning opportunities." For the three deep level dimensions of instructional quality—i.e., structuredness, support and challenge—we can, however, state that the school type has significant impact, because all three dimensions of quality were assessed more positively in the educational track of Hauptschule (general secondary school) than at schools from the Gymnasium (grammar school) or Realschule (intermediate secondary school) tracks. Moreover, the aspects of professional work we assessed among the teaching staff (cooperation, competency goals, and expectations of achievement) do not reveal any significant effects, thus they do not contribute to the students' perceptions of instructional quality beyond the control variables we considered.

In a final analytical step, the effects of the school and instruction level on the increase in learning and motivation are assessed (see Table 7.6). Regarding our main research question, we can establish that none of the three characteristics of professional work at the school level impacts upon achievement and motivation. This applies when simultaneously taking control variables into account (in each case, models II), but also when only looking at the school characteristics as such (models I).

However, the findings outlined in Table 7.6 support our model of instructional quality. The indicator of cognitively challenging lessons used here, "demand on correct language use," bears a significant and also sizable effect on the increase in achievement, at both the individual and the classroom levels. This implies that a high cognitive challenge, as commonly perceived by the students, influences achievement development in a positive way; moreover, within a class, those students who perceive this aspect of instructional quality in a more positive light than their peers are distinguished by an even higher increase in achievement. Pursuant to our assumptions, teacher support is particularly important for the development of motivation.

Thus, the theoretical assumption that school quality, and more precisely the professional norms and cooperation among teaching staff, mediated by instructional quality, influences the development of students, receives no support from the DESI data on German lessons. Contradicting the assumptions of school research, and even more contrary to the expectations of school development researchers, our study does show indications of an effect of school characteristics on the development of learning and motivation in German lessons.

At the level of classroom instruction, however, an effect can be ascertained for cognitively demanding lesson designs (promoting achievement development) as

Table 7.5 Three-level model for explaining instructional characteristics

Predictors	Structuredness		Support		Demand language use		Learning opportunity	
	I	II	I	II	I	II	I	II
Student level								
Social status		−0.034*		−0.053*		−0.047*		−0.013
Basic cognitive abilities		−0.030*		0.027		0.065*		−0.051*
Gender female		0.072*		0.206*		0.365*		0.023
German as first language		-0.033		0.082*		0.029		0.051
Classroom level								
Social composition		0.062		0.234		0.298*		−0.046
Cognitive composition		0.027		−0.003		0.152**		−0.002
Proportion of girls		−0.339*		−0.140		0.186		−0.151
Proportion of first language German		−0.347*		−0.370*		−0.413*		−0.266
School level								
Expected achievement	−0.016	0.020	0.006	0.019	0.045	0.014	<0.001	0.048
Cooperation	0.054*	0.015	−0.017	−0.037	−0.036	0.019	0.054*	0.026
Competency goals	−0.046*	−0.002	−0.032	−0.010	0.020	−0.011	−0.082*	−0.033
School with Hauptschul-educational track		0.343*		0.325*		0.176*		0.168*
Grammar school		0.079		−0.028		−0.049		−0.138

*$p<0.05$; **$p<0.07$

Table 7.6 Three-level model for explaining achievement gains and motivational development

Predictors	Achievement gains			Increase in motivation		
	I	II	III	I	II	III
Student level						
Social status		0.001	0.001		0.009	0.009
Basic cognitive abilities		0.123*	0.122*		−0.075*	−0.076*
Gender female		0.077	0.077		0.190*	0.191*
German as first language		0.006	0.006		0.025	0.027
Demand language use		0.085*	0.085*		0.055	0.055
Structuredness of lessons		0.039	0.039		0.064*	0.064*
Learning opportunities		−0.017	−0.017		0.010	0.010
Teacher support		0.029	0.029		0.205*	0.204*
Classroom level						
Social composition		−0.151	−0.001		−0.040	0.018
Cognitive composition		0.204	0.274*		0.132	0.180
Proportion of girls		0.056	0.063		−0.207	−0.200
Proportion German first language		−0.221	−0.322		−0.547*	−0.559*
Demand language use		0.332*	0.335*		0.126	0.120
Structuredness of lessons		−0.141	−0.139		0.087	0.074
Learning opportunities		−0.147	−0.166		0.110	0.099
Teacher support		0.147	0.135		0.232*	0.216*
School level						
Expected achievement	−0.030	−0.034	−0.037	−0.049	−0.020	−0.017
Cooperation	−0.008	0.040	0.034	0.050	0.049	0.035
Competency goals	−0.008	−0.033	−0.024	−0.006	0.007	0.025
School with Hauptschule educational track			0.106			0.150
Grammar school			−0.205			−0.043

*$p < 0.05$; **$p < 0.07$

well as supportive teacher behavior (promoting motivation). DESI thus supports the teaching quality model that assumes three basic dimensions, with cognitive challenges presenting the most important predictor of achievement, whereas teacher support determines motivation development. Both are described as "deep level characteristics" in instructional research. Corresponding to theoretical assumptions, the frequency of learning opportunities in terms of a "surface characteristic" does not correlate with learning and motivation development.

We can summarize this pattern of findings as follows: Basic dimensions of instructional quality prove to be effective in the subject of German, while it is impossible to match the professional instructional activities with the professional actions at school level (i.e., cooperation, expected achievement, and competency goals among staff). The school level factors influence instructions particularly regarding characteristics of the diversity of learning opportunities, which other than supportive measures and cognitive challenges, do not render any significant contribution to instructional effectiveness.

Enhanced ILSA Designs Allow for Testing Organizational Change: Longitudinal Studies on the School Level—The German PISA School Panel

The mainstream of school improvement research is still largely grounded in case studies (cf. Hopkins 2005; Lee and Williams 2006). Large-scale international student assessments like TIMSS and PISA can provide new insights into the mechanisms of school change, because they offer high quality achievement data and a broad array of context and process data (including school policies, curricular and extracurricular opportunities, school climate, and many more). However, from a school effectiveness point of view, these studies have limited explanatory power because they are all cross-sectional. Effects caused by school policies and school-level processes cannot be separated from selection bias.

"The often-heard plea for more longitudinal research in school effectiveness can only be repeated here. Not only effects should be measured at more than one point in time, but also input and process variables" (Scheerens and Bosker 1997, p. 315).

A national enhancement to the PISA studies luckily provides longitudinal information for hundreds of schools. The results presented here are built on national enhancements to the OECD PISA studies that were administered in 2000, 2003, and 2006.[2] In each of those years, the "international" PISA sample in Germany, which consisted of about 200 schools, has been enhanced by a national sampling scheme, which applies PISA tests and background questionnaires to 1,500 schools all over Germany, allowing for a comparison between federal states.

Within those very large data sets, 506 schools could be found that had been assessed at least twice in 2000 and in 2003. Most of those schools are located in small federal states, so the sample is by no means representative for Germany. However, it can help to study stability in school variables.

We applied hierarchical linear modeling, with students both from 2000 and 2003 included in a virtual sample, and membership in one of the two cohorts as a level 1-indicator (see Fig. 7.5 and Table 7.7 for the associated parameter estimates).

With a 0.93 correlation between mean school achievement in 2000 and in 2003, this variable shows high stability. However, the extremely high parameter also results from the stability of school *type* (track) differences. When looking at the lowest track (Hauptschule) only, the stability of achievement over three years decreases to 0.84; while for grammar schools only, it is down to 0.57. Thus within the German school system, there is some instability of school results. Schools move up or down, and we might try to explain those changes by changes in school input and school processes.

As we had assumed, there is a complex interplay between school composition (i.e., mean student SES) and student achievement (Fig. 7.6). Schools with a com-

[2] This research has been initiated by Klieme and Steinert 2008; the findings cited here have first been presented by Hochweber et al. (2010).

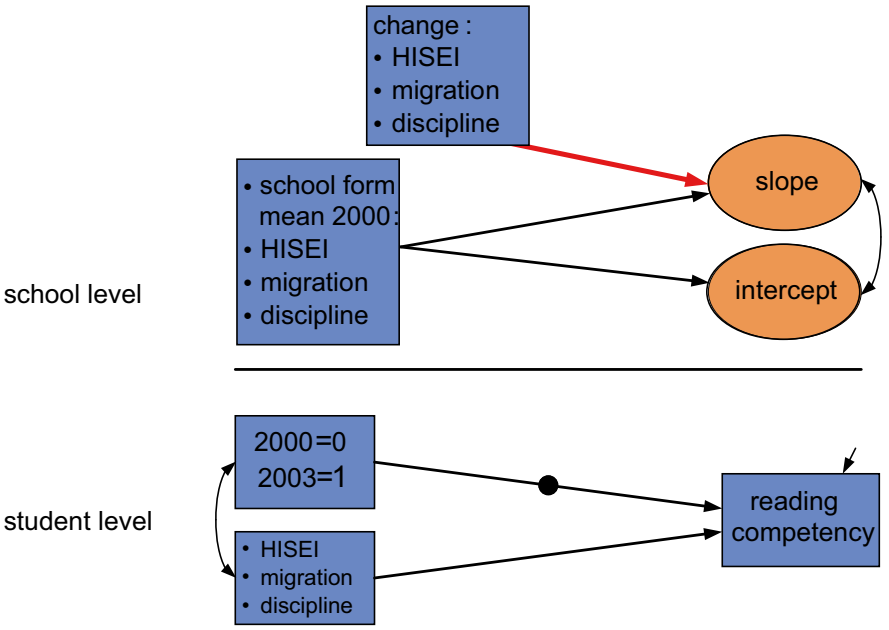

Fig. 7.5 Multilevel model for the analysis of organizational longitudinal data

Table 7.7 Parameter estimation and significance test for the model in Fig. 7.5

	I	II
Dependent: 2000 reading level		
School form: HS educational track	−32.4	−32.2
School form: GY	36.2	36.6
Proportion of migrant students	−12.3	−12.0
Mean HISEI	31.8	31.5
Discipline, school climate	−11.5	−12.8
Dependent: change in reading 2000–2003		
School form: HS educational track	−0.5	1.7
School form: GY	0.9	−4.5
Proportion of migrant students	*−3.4*	−1.7
Mean HISEI	−3.2	0.1
Discipline, school climate	*4.2*	*3.7*
Difference migration		*−4.0*
Difference HISEI		*3.4*
Difference school climate		*3.9*

*$p < 0.05$

paratively high achievement can maintain or improve their social composition. This finding leads to a better understanding of the relation between student composition and school outcome. Traditionally, only the impact of individual SES and student composition on student learning and outcomes has been considered.

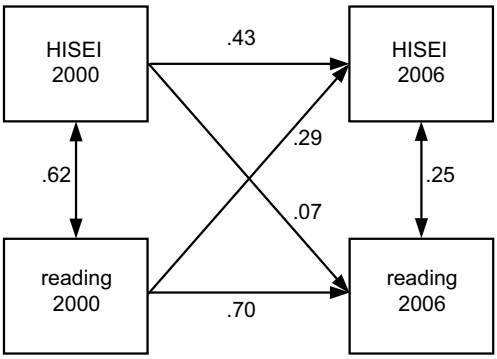

Fig. 7.6 Cross-lagged panel analysis (school level only) of the interrelation between reading and SES background

What about school level input and processes? Results indicate that classroom discipline, mean SES, and proportion of migrant students explain the (aggregated) achievement status on school level. The better the disciplinary climate, and the lower the proportion of migrant students, the better reading competency develops over three years. And, finally, schools that succeed in increasing disciplinary climate, attracting students from higher SES backgrounds, and a reduced proportion of migrant students will show higher gains in reading achievement.

Summary and Conclusion

The main purpose of international large-scale assessment is to provide indicators for continuous monitoring of educational systems. Compared with the complexity and theoretical challenges of Educational Effectiveness Research, however, ILSAs show severe limitations—the most important being the absence of longitudinal data, especially baseline information on prior achievement. Without longitudinal designs, it is practically impossible to build adequate, complex explanatory models or to draw causal inferences.

However, there are ways to enhance the design of national as well as international large-scale assessments to allow for stronger explanatory power. This chapter reported on two such enhancements implemented in Germany: adding a short-term longitudinal assessment on the student and classroom level covering one school year (implemented within the language assessment study DESI), and resampling schools to study school development as an organizational process (implemented within national extensions to PISA). We recommend that ILSA studies move in those directions to increase validity as well as policy relevance.

Nevertheless, useful links already exist between ILSAs and EER. First, ILSAs need constructs and instruments and theoretical insight from Educational Research, including EER, to design and analyze the studies. Thus, designing advanced ILSA studies is a challenge to Effectiveness Research, which may even initiate new developments in theory and empirical work, as has been the case with the notion of

"opportunity to learn." Second, ILSAs can help foster EER by allowing explorative analyses and generating hypotheses, by testing research hypotheses, and by studying the intercultural biases and culture-specific context factors that shape the functioning of educational systems.

Hopefully the future will see further advancements in the interaction between Educational Effectiveness Research and international large-scale assessments. To reach this goal, the research community has to gain support of policymakers.

Acknowledgment The present author gratefully acknowledges contributions from Johannes Hartig, Jan Hochweber, Nina Jude, Ulf Kröhne, Katrin Rakoczy, Brigitte Steinert, Svenja Vieluf, all from DIPF, Center for Research on Educational Quality and Evaluation. Martina Kenk and Gwen Schulte have been especially helpful in editing this chapter.

References

Abedi, J., M. Courtney, S. Leon, J. Kao, and T. Azzam. 2006. *English language learners and math achievement: A study of opportunity to learn and language accommodation* (*CSE Report 702, 2006*). Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Baker, D. P. 2009. The invisible hand of world education culture. In *Handbook of education policy research,* eds. G. Sykes, B. Schneider and D. N. Plank, 958–968. New York: Routledge.

Baumert, J., R. Lehmann, M. Lehrke, B. Schmitz, M. Clausen, and I. Hosenfeld, et al. 1997. *TIMSS—Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde*. Opladen: Leske+Budrich.

Baumert, J., M. Kunter, W. Blum, M. Brunner, T. Voss, and A. Jordan, et al. 2009. Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal* 47:133–180.

Borman, G. D., G. M. Hewes, L. T. Overman, and S. Brown. 2003. Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research* 73:125–230.

Brophy, J., ed. 1992. *Planning and managing learning tasks and activities: Advances in research on teaching* (*Vol. 3*). Greenwich: JAI Press.

Brown, A. L. 1994. The advancement of learning. *Educational Researcher* 23(8):4–12.

Bryk, A. S., P. B. Sebring, E. Allensworth, S. Luppescu, and J. Q. Easton. 2010. *Organizing schools for improvement. Lessons from Chicago*. Chicago: University of Chicago Press.

Carroll, J.B. 1963. A model of school learning. *Teachers College Record* 64:723–733.

Clausen, M. 2002. *Unterrichtsqualität: Eine Frage der Perspektive*? Münster: Waxmann.

Creemers, B. P. M. 1994. *The effective classroom*. London: Cassell.

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

Deci, E. L., and R. M. Ryan. 1985. *Intrinsic motivation and self-determination in human behavior*. New York: Plenum Press.

Ditton, H. 2000. Qualitätskontrolle und Qualitätssicherung in Schule und Unterricht. *Zeitschrift für Pädagogik* 41(Beiheft):73–92.

Ditton, H. 2007. Schulqualität – Modelle zwischen Konstruktion, empirischen Befunden und Implementierung. In *Qualität von Schule,* eds. J. van Buer and C. Wagner, 83–92. Frankfurt a. M.: Lang.

Ditton, H., and L. Krecker. 1995. Qualität von Schule und Unterricht. Empirische Befunde zu Fragestellungen und Aufgaben der Forschung. *Zeitschrift für Pädagogik* 41(4):507–531.

Fend, H. 1998. *Qualität im Bildungswesen. Schulforschung zu Systembedingungen, Schulprofilen und Lehrerleistung*. Weinheim: Juventa Verlag.

Goldschmidt, P., K. Choi, and F. Martinez. 2004. *Using hierarchical growth models to monitor school performance over time: Comparing NCE to scale score results. CES Report 618*. Los Angeles: Center for the study of evaluation, University of California.

Gustafsson, J.-E. 2007. Understanding causal influences on educational achievement through analysis of differences over time within countries. In *Lessons learned: What international assessments tell us about math achievement,* ed. T. Loveless, 37–63. Washington, D.C.: The Brookings Institution.

Hanushek, E. A., and L. Woessmann. 2009. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. NBER Working Paper No. 14633, National Bureau of Economic Research, Cambridge. http://www.nber.org/papers/w14633.pdf. Accessed 3 Nov 2011.

Hanushek, E. A., and L. Wossmann. 2010. The economics of international differences in educational achievement. NBER Working Papers 15949, National Bureau of Economic Research, Cambridge. http://www.nber.org/papers/w15949.pdf. Accessed 3 Nov 2011.

Harris, A., and J. H. Chrispeels., eds. 2006. *Improving schools and educational systems: International perspectives*. London: Routledge.

Hattie, J. 2009. *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hiebert, J. R., H. Gallimore, K. B. Garnier, H. Givven, J. Hollingsworth, and A. M.-Y. Jacobs, et al. 2003. Teaching mathematics in seven countries: Results from the TIMSS 1999 video study. US Department of Education, National Center for Education Statistics, Washington, D.C.

Hochweber, J., B. Steinert, J. Gomolka, and E. Klieme. 2010. Schulentwicklung 2000–2003–2006: Befunde auf Basis der PISA-Studien. Paper presented at the 2010 Congress of the German Society for Educational Science (DGFE), Mainz, Germany.

Hopkins, D., ed. 2005. *The practice and theory of school improvement: International handbook of educational change*. Springer: Dordrecht.

Husén, T. 1967. *International study of achievement in mathematics* (*Vol. 2*). New York: Wiley.

Husén, T. 1974. Introduction to the reviews of three studies of the International Association for the Evaluation of Educational Achievement (IEA). *American Educational Research Journal* 11(4):407–408.

Klieme, E., and K. Rakoczy. 2003. Unterrichtsqualität aus Schülerperspektive. In *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland,* eds. J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele et al., 333–359. Opladen: Leske+Budrich.

Klieme, E., and B. Steinert. 2008. Schulentwicklung im Längsschnitt. Ein Forschungsprogramm und erste explorative Analysen. *Zeitschrift für Erziehungswissenschaft* 10(Sonderheft):221–238.

Klieme, E., and S. Vieluf. 2009. Teaching practices, teachers' beliefs and attitudes. In *Creating effective teaching and learning environments. First results from talis,* ed. OECD. Paris: Organisation for Economic Co-operation and Development. http://www.oecd.org/dataoecd/17/51/43023606.pdf.

Klieme, E., G. Schümer, and S. Knoll. 2001. Mathematikunterricht in der Sekundarstufe I: Aufgabenkultur und Unterrichtsgestaltung. In *Bundesministerium für Bildung und Forschung (BMBF)* ed. TIMSS—Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente, 43–57. München: Medienhaus Biering.

Klieme, E., A. Helmke, R. Lehmann, G. Nold, H.-G. Rolff, K. Schröder, et al. eds. 2008. *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie*. Weinheim: Beltz.

Klieme, E., C. Pauli, and K. Reusser. 2009. The Pythagoras study: Investigating effects of teaching and learning in Swiss and German classrooms. In *The Power of Video Studies in Investigating Teaching and Learning in the Classroom*, eds. T. Janik and T. Seidel, 137–160. Münster: Waxmann Verlag.

Klieme, E., E. Backhoff, W. Blum, J. Buckley, Y. Hong, D. Kaplan, H. Levin, J. Scheerens, W. Schmidt, F. van de Vijver, and S. Vieluf. 2010a. *Designing PISA as a sustainable database for educational policy and research: The PISA 2012 Context Questionnaire Framework*. Paris: OECD.

Klieme, E., N. Fischer, H. G. Holtappels, T. Rauschenbach, and L. Stecher. 2010b. *Ganztagsschule – Entwicklung und Wirkungen*. Frankfurt a. M.: DIPF.

Klieme, E., B. Steinert, and J. Hochweber. 2010c. Zur Bedeutung der Schulqualität für Unterricht und Lernergebnisse. In *Schulische Lerngelegenheiten und Kompetenzentwicklung. Festschrift für Jürgen Baumert,* eds. W. Bos, E. Klieme and O. Köller, 231–255. Münster: Waxmann.

Kröhne, U. 2010. Comparison of quasi-experimental methods for large-scale assessments: Estimating the effect of bilingual instruction based on a subsample of the DESI study. Invited presentation at the 2010 Symposium on Causality, Jena University, Germany.

Kunter, M., T. Dubberke, J. Baumert, W. Blum, M. Brunner, A. Jordan, et al. 2006. Mathematikunterricht in den PISA-Klassen 2004: Rahmenbedingungen, Formen und Lehr-Lernprozesse. In *PISA 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres,* eds. M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand et al., 161–194. Münster: Waxmann.

Kunter, M., Y.-M. Tsai, U. Klusmann, M. Brunner, S. Krauss, and J. Baumert. 2008. Students' and mathematics teachers' perception of teacher enthusiasm and instruction. *Learning and Instruction* 18:468–482.

Kyriakides, L., and N. Tsangaridou. 2004. School effectiveness and teacher effectiveness in physical education. Paper presented at the 85 Annual AERA Meeting, American Educational Research Association, Chicago, USA.

Lee, J. C., and M. Williams. 2006. *School improvement: International perspectives*. New York: Nova Science Publishers.

Lipowsky, F., K. Rakoczy, C. Pauli, B. Drollinger-Vetter, E. Klieme and K. Reusser. 2009. Quality of Geometry instruction and its short-term impact on students' understanding of the Pythagorean theorem. *Learning and Instruction* 19:527–537.

Luyten, H., and R. de Jong. 1998. Parallel classes: Differences and similarities. Teacher effects and school effects in secondary schools. *School Effectiveness and School Improvement* 9(4):437–473.

Organisation for Economic Co-operation and Development. ed. 2007a. *Education at a glance*. Paris: Organisation for Economic Co-operation and Development.

Organisation for Economic Co-operation and Development. ed. 2007b. *PISA 2006. Science competencies for tomorrow's world*. Paris: Organisation for Economic Co-operation and Development.

Organisation for Economic Co-operation and Development. ed. 2008. *Education at a glance*. Paris: Organisation for Economic Co-operation and Development.

Organisation for Economic Co-operation and Development. ed. 2009a. *Education at a glance*. Paris: Organisation for Economic Co-operation and Development.

Organisation for Economic Co-operation and Development. ed. 2009b. *PISA 2009 Assessment framework key competencies in reading, mathematics and science*. Paris: OECD.

Pianta, R. C., and B. K. Hamre. 2009. Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher* 38(2):109–119.

Purves, A. C. 1987. The evolution of the IEA: A memoir. *Comparative Education Review* 31(1):10–28.

Rowan, B., E. Camburn, and R. Correnti. 2004. Using teacher logs to measure the enacted curriculum in large-scale surveys: Insights from the study of instructional improvement. *Elementary School Journal* 105:75–102.

Sammons, P. 1999. *School effectiveness: Coming of age in the 21st century*. Lisse: Swets & Zeitlinger.

Sammons, P., D. Nuttall, P. Cuttance, and S. Thomas. 1995. Continuity of school effects: A longitudinal analysis of primary and secondary school effects on GCSE performance. *School Effectiveness and School Improvement* 6(4):285–307.

Sammons, P., S. Thomas, and P. Mortimore. 1997. *Forging links: Effective schools and effective departments*. London: Paul Chapman Publishing Ltd.

Scheerens, J. 2000. *Improving school effectiveness. Fundamentals of educational planning series, IIEP, Vol. 68*. Paris: United Nations Educational, Scientific, and Cultural Organization (UNESCO).

Scheerens, J., and R. J. Bosker. 1997. *The foundations of educational effectiveness*. Oxford: Pergamon Press.

Schmidt, W. H., and A. Maier. 2009. Opportunity to learn. In *Handbook of Education Policy Research,* eds. G. Sykes, B. Schneider and D. N. Plank, 541–559. New York: Routledge.

Schmidt, W. H., and C. McKnight. 1995. Surveying educational opportunity in mathematics and science: An international perspective. *Educational Evaluation and Policy Analysis* 17(3):337–353.

Schwerdt, G., and A. C. Wuppermann. 2011. Is traditional teaching really all that bad? A within-student between-subject approach. *Economics of Education Review* 30:365–379.

Seidel, T., and R. J. Shavelson. 2007. Teaching effectiveness research in the last decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research* 77:454–499.

Slavin, R. E. 1996. *Education for all*. Lisse: Swets & Zeitlinger.

Stanat, P., and G. Christensen. 2006. *Where immigrant students succeed—a comparative review of performance and engagement in PISA 2003*. Paris: Organisation for Economic Co-operation and Development.

Stevens, F. 1993. Applying an opportunity-to-learn conceptual framework to the investigation of the effects of teaching practices via secondary analyses of multiple-case-study summary data. *Journal of Negro Education* 62(3):232–248.

Stigler, J. W., and J. Hiebert. 1999. *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: Free Press.

Stringfield, S. 1994. A model of elementary school effects. In *Advances in school effectiveness research and practice,* eds. D. Reynolds, B. P. M. Creemers, P. S. Nesselrodt, C. Teddlie, E. C. Shaffer and S. Stringfield, 153–187. Oxford: Pergamon.

Teddlie, C., and D. Reynolds. eds. 2000. *The international handbook of school effectiveness research*. New York: Routledge.

Teddlie, C., and S. Stringfield. 1993. *Schools make a difference. Lessons learned from a 10-year study of school effects*. New York: Teachers College Press.

Teddlie, C., and S. Stringfield. 2006. A brief history of school improvement research in the USA. In *Improving schools and educational systems: international perspectives,* eds. A. Harris and J. Chrispeels, 131–166. London: Routledge.

Thomas, S., P. Sammons, P. Mortimore, and R. Smees. 1997. Stability and consistency in secondary schools' effects on students' GCSE outcomes over three years. *School effectiveness and school improvement* 8(2):169–197.

Tschannen-Moran, M., and A. Woolfolk Hoy. 2001. Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education* 17(7):783–805.

Walberg, H. J. 1986. Syntheses of research on teaching. In *Handbook of research on teaching* (*3rd edn.*)*,* ed. M. C. Wittrock, 214–229. New York: Macmillan.

Walker, D. A. 1976. *The IEA six-subject survey: An empirical study of education in twenty-one countries. International studies in evaluation*. New York: Wiley.

Wang, M. C., G. D. Haertel, and H. D. Walberg. 1993. Toward a knowledge base for school learning. *Review of Educational Research* 63(3):249–294.

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