Chapter 19
Assessment and Evaluation in Educational Contexts

Sonja Bayer, Eckhard Klieme, and Nina Jude

Contents

Abstract For at least the past three decades, assessment, evaluation, and accountability have been major strands of educational policy and practice internationally. However, the available data on how exactly assessment- and evaluation-based policies are framed and implemented, or how they shape practices within schools, are still limited. This chapter addresses these issues with a broad focus that takes into account several perspectives on school evaluation and student assessment, together with everyday practices of teacher judgment and grading. First, we address assessment and evaluation practices for the purpose of educational system monitoring. Second, school evaluation practices, as well as the use of assessment and evaluation results at the school level, are discussed. A third perspective focuses on practices of teacher evaluation. Finally, practices of student assessment within schools and classrooms are examined. The instruments described and recommended in this chapter have implications for international research, as well as national studies.

S. Bayer (✉) • E. Klieme • N. Jude
Department for Educational Quality and Evaluation, German Institute for International Educational Research (DIPF), Frankfurt, Germany
e-mail: bayer@dipf.de; klieme@dipf.de; jude@dipf.de

S. Kuger et al. (eds.), *Assessing Contexts of Learning*, Methodology of Educational Measurement and Assessment, DOI 10.1007/978-3-319-45357-6_19

19.1 Introduction

For at least three decades, assessment and evaluation have been major strands of educational policy and practice internationally. In recent years, there has been growing interest in the use of assessment and evaluation results through feedback to students, parents, teachers, and schools, as one of the most powerful tools for quality management and improvement. Reporting and sharing data from assessments and evaluations with different stakeholders provides multiple opportunities for monitoring both individual learning and institutional development, for certification and accountability (Elaqua 2016, Chap. 15 in this volume). The volume *Schools and Quality*, published by OECD in 1989, marked the initiation of a global trend that is still ongoing: "educational assessment, evaluation, and accountability are still evident in educational practice and policy making in virtually every country" (Huber and Skedsmo 2016, p. 1). This trend is part of an overarching change in concepts and measures of educational governance (Altrichter and Maag Merki 2016). New forms of educational governance, such as school performance feedback systems (Visscher and Coe 2003), systemic approaches to educational evaluation and monitoring (Scheerens et al. 2003) and concepts of data-driven school improvement (Coburn and Turner 2011; Spillane 2012) have become popular among policy makers. Research sets out to understand their functionality and effectiveness (e.g., Altrichter and Maag Merki 2016; Torrance 2013). However, there is still limited knowledge on how exactly assessment- and evaluation-based policies are framed and implemented, or how they shape practices within schools.

This chapter reflects this debate, but it also expands the focus to several layers of evaluation and assessment. Our conceptual framework, elaborated in the next section, addresses four levels of the educational system: the system in general, schools, classrooms, and the individual. First, we describe how the idea of system monitoring in the educational context evolved, and describe current developments and practices in the monitoring and governance of educational systems (Sect. 19.2.1). Second, evaluation practices and processes at the school level are the subject of discussion (Sect. 19.2.2). The results of school evaluations and student assessment may be used for evidence-based management within schools, e.g., to guide the allocation of resources, the promotion and retention of students, or the professional development of the teaching staff. Third, discussion focuses on the evaluation of teachers (Sect. 19.2.3). Finally, the practices of student assessment within schools and classrooms are the objects of interest in Sect. 19.2.4, which takes into account grading, certification, and formative feedback using various assessment instruments.

International Large-scale Student Assessments (ILSAs) like TIMSS, PIRLS and PISA are major instruments of, and driving factors for, system-level monitoring. They provide complex techniques to be used *for* assessment, evaluation, and accountability at all levels of the educational system, as this volume as a whole shows. At the same time, these international surveys can be used as sources of information *about* assessment, evaluation and accountability practices in cross-national

comparison, as demonstrated in the present chapter. The intention of this chapter is to support the "*assessment of assessment"* through instruments that help document and analyze all layers of the evaluation and monitoring system. Thus, empirical data may inform critical debates on assessment, evaluation, and accountability systems in the public sphere, in policy and pedagogy, and overcome the purely ideological debates that oftentimes dominate this discourse (for an outline of the aims and objectives of PISA see also Kuger and Klieme 2016 and Jude 2016 Chaps. 1 and 2 in this volume).

In order to discuss these matters, we have to integrate theories from several perspectives, including educational effectiveness theories, governance theories, organizational theories, and theories on teaching and learning. Following the conceptual framework, we summarize the most relevant concepts and discuss their benefits and feasibility in national and international large-scale assessments. Most of the concepts we propose in this conceptual framework of assessment and evaluation were realized in the PISA 2015 field trial (see Table 19.1 in Sect. 19.3). The instruments described and recommended in this chapter may thus be used for international research, as well as for national studies.[1]

19.2 Conceptual Framework

Over the years, the assessment/evaluation paradigm has shifted from a focus on measurement towards a focus on efforts to improve learning (Wyatt-Smith 2014). In an international review undertaken by the OECD, experts from 28 countries agreed that the ultimate objective of assessment and evaluation is to improve the quality of education in countries and, as a consequence, raise student outcomes (OECD 2013). Nevertheless, different stakeholders make decisions for different levels of the educational systems, and they support their decisions using data drawn from the educational system. In line with the OECD (2013) review, we identify and define four main areas of assessment and evaluation that, while related to each other, differ with respect to the unit of judgment: Monitoring the educational system as a whole, school evaluation, teacher evaluation, and student assessment.

Educational system monitoring sometimes also called education system evaluation, concerns the evaluation of an education system to provide accountability information to the public, and to inform policies aiming to improve educational processes and outcomes. The unit of evaluation can be either a national education system or a subnational education system. In the present chapter, we focus on systematic and regular system evaluation, such as indicator-based reports, and therefore use the term "monitoring", which emphasizes the ongoing observation of educational systems.

[1] This chapter expands on a technical paper that was presented to the PISA 2015 Questionnaire Expert Group (QEG) in May 2012 (Doc. QEG 2012–05 Doc 08).

School evaluation refers to judgments on the quality and effectiveness of schools. The evaluation may be implemented by a school inspectorate, any other administrative body, or the school itself. School evaluation concentrates on key processes within school, often in association with an analysis of student outcomes. It also takes into account input variables such as infrastructure, funding or characteristics of the school staff.

Teacher evaluation also known as teacher appraisal, refers to judgments on the performance of teachers. The evaluation of teachers is subject to two alternative procedures: (1) The formative approach typically includes regular appraisal to gain and maintain registration and accreditation to teach, and for promotion as part of a school's performance management processes. (2) The accountability approach intends to identify a select number of high-performing teachers, to reward and acknowledge their teaching competence and performance, while underperforming teachers may be required to participate in professional development, their salary may be reduced, or they even may be fired. These formal schemes are often complemented with more informal school-level practices of feedback to teachers.

Student assessment refers to judgments on individual student progress and achievement of learning goals. It covers classroom-based assessments, including grading by teachers, as well as large-scale external assessments and examinations.

It should be noted, however, that measures may be used across areas. For instance, student outcomes, aggregated to the appropriate level, may be used to judge educational systems, individual schools, and teachers. International Large-scale Assessments, for example, do assess individual students, although their goal is monitoring educational systems.

In the following, we discuss the main developments, concepts and practices for each of these four areas separately.

19.2.1 Educational System Monitoring

Educational system monitoring contributes to the building of national and international evidence bases that offer the prospect of allowing us to analyse and compare structures and processes in educational systems. This in turn can enhance our understanding of education-related decisions. Across the world, growing interest in student assessment and educational comparability studies has led to the establishment of national and international assessment associations since the late 1950s. This is associated with a focus on output-driven models of governance. This change in governance perspective also reflects "the rise of a profound skepticism about the possibilities of hierarchical control of complex social systems" (Boer et al. 2007, p. 137).

One of the early key findings of research in this area concerns the relationships between centralization and decentralization, and student achievement. Decentralization of various educational functions is said to be positively related to performance (Blöchliger 2013). Based on PISA data, Hanushek et al. (2013) have shown that decentralization (autonomy on key operations of a school) has a positive impact on student achievement in developed countries only, whereas the impact is negative in developing countries. In many western and eastern countries the idea of decentralization has become national educational policy, combined with systems of evaluation and monitoring (Scheerens et al. 2003). Thus, policies based on school autonomy and decentralization also require quality assurance through strong, transparent monitoring mechanisms, including, for example, national standards, centralized exams, and large-scale assessments. There is evidence that the combination of school autonomy with standard setting and accountability measures may be an effective reform strategy, at least in developed countries (Wößmann 2003).

Relevant monitoring indicators on a national level are commonly set by central educational authorities, chief inspectorates or departments within ministries or education authorities (Faubert 2009). Educational policy making 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; Klieme 2013). Outcome indicators are oftentimes measured with regard to national educational standards defining the skills that students should possess in primary, secondary and tertiary education, and the knowledge that they are expected to know at a specific stage of their education (Koeppen et al. 2008; Shepard 2006). The results of large-scale assessments based on national educational standards are often used for system monitoring, but also for school evaluation.

In many countries, international comparison of educational achievement is an essential part of long-term system monitoring policies. While national standards are hardly comparable between countries, international studies like TIMSS, PIRLS, and PISA aim at addressing comparable educational indicators. In addition to educational outcomes, in the sense of literacy assessed by tests, these studies also focus on context indicators such as inputs, processes, and non-cognitive outcomes (see Kuger and Klieme 2016, Chap. 1 in this volume). These data, as well as conclusions drawn from international comparisons, can then be used in national educational policy making.

Overall, large-scale assessments allow for national and international comparisons of educational systems. This spreads accountability to the system level. The European Union for instance, sets benchmarks for education, which are monitored regularly (European Commission 2011). By taking into account national and international reports, central educational authorities are able to evaluate and monitor system policies, their implementations and value.

19.2.2 School Evaluation

The evaluation of schools is an instrument of educational governance that becomes even more important with the switch to more decentralized educational systems. It is also used in decisions and judgments about processes, programs, reforms, and educational resources (Faubert 2009). Moreover, the evaluation of schools can help school leaders to make better decisions about processes, build knowledge and skills, or facilitate continuous improvement and organizational learning. The improvement of schools participating in evaluation programs can be explained by feedback theory (Visscher and Coe 2003), or as an effect of stakeholders within school being held accountable for evaluation results (Donaldson 2004). Scheerens et al. (2003) elaborate on the notion of data-driven school development, pushed by a combination of internal and external evaluation. They assume evaluation to be the fundamental process in which a school becomes a learning organization, and they believe evaluation- and feedback-based school improvement to be more effective than any forward-planning strategy.

School evaluation and improvement can in turn also affect students' outcomes. For instance, Scheerens (2002), and also Creemers and Kyriakides (2008), found some evidence that systematic school evaluation can positively impact students' outcomes. On the basis of a school panel added to the PISA 2000 and 2009 samples in Germany, Bischof and colleagues (2013) report that schools who had done some internal evaluation improved in terms of student achievement and school climate. Likewise, Hofman and colleagues (2009) identified factors of internal evaluation (self-evaluation) that contribute to student achievement. However, studies over the past decades have shown that non-profit organizations, like most kinds of schools, oftentimes do not use evaluation effectively (Donaldson 2004); some challenges need to be overcome.

In a review of 41 empirical studies on evaluation use, Johnson and colleagues (2009) found the involvement of stakeholders to be most important for effective school evaluations. Engagement, interaction, and communication between evaluation clients and evaluators are critical to the meaningful use of evaluations. This is in accordance with the utilization-focused evaluation theory (Patton 1997), which emphasizes the involvement and engagement of users in the evaluation processes of designing, judging, and decision-making (Alkin and Christie 2004). Other categories related to the use of evaluation are detailed, actionable, evidence-based recommendations, and decision characteristics (Johnson et al. 2009). Scheerens et al. (2003) claim that effective school evaluation needs to combine outcome- and process-related indicators. Consequently, common steps of effective evaluation can be identified (e.g., Sanders and Davidson 2003), yet school evaluation approaches are multifold, spanning, for instance, empowerment evaluation, utilization-focused evaluation, inclusive evaluation, or theory-driven evaluation, to name just some of the most popular (Donaldson 2004; see also Alkin and Christie 2004, who have developed a different scheme for classifying evaluation theories).

Thus, it is hardly surprising that evaluation approaches vary across educational systems (OECD 2013) and that it is difficult to report on and compare the effects of evaluation across different evaluation systems and education systems. Even though evaluation instruments and approaches differ across educational systems, at least two broad categories of evaluation can be identified: internal evaluation and external evaluation. Evaluations are external when contractors and evaluators or test administrators do not belong to the school that is being evaluated. If the evaluator or test administrator is a member of the same organization, but not part of the unit that is evaluated, evaluation or assessment is internal. Self-evaluation is a special form of internal evaluation. Here, the evaluators are part of the unit that is being evaluated (Scheerens 2002; Berkemeyer and Müller 2010). The different evaluation practices generally coexist and benefit from each other (Ryan et al. 2007). External evaluation can expand the scope of internal evaluation, and also validate results and implement standards or goals. Internal evaluation can improve the interpretation of external evaluation results (Nevo 2002).

According to one review of evaluation use (Johnson et al. 2009), there seems to be a lack of research addressing the processes of evaluation. Nevertheless, certain topics and types of evaluation can be discerned. While in early days school evaluations—especially in English speaking countries—mainly and sometimes only, focused on students' outcomes (Nevo 1998), evaluations nowadays seem to address various components or subcomponents of the school environment (Donaldson 2004). For instance, evaluation frameworks across countries address educational practices (OECD 2013). A comprehensive framework for guiding school evaluation processes is the context-input-process-outcome (CIPO) model (Stufflebeam 2003). Each type of evaluation has its own focus: needs, strategies, implementations or outcomes (Alkin and Christie 2004). In the context of educational effectiveness research that aims to explain differences between schools, the CIPO model allocates input, process and outcome characteristics at the appropriate levels of action (Scheerens and Bosker 1997). Within this framework, relevant foci of evaluation might, for instance, be the school's resources or the proportion of at-risk student sub-groups (input). Processes addressed in evaluations may be teacher collaboration or parental involvement. The most common output addressed is the cognitive performance of students, but also socio-emotional outcomes or equity within the school might be relevant aspects (Faubert 2009). The results of evaluations may be used in a formative way, guiding school improvement, or in a more summative way—e.g., making schools accountable for their students' outcomes (Alkin 1972). Formative school evaluation aims at teaching and school-based processes. Summative evaluations have a strong but not exclusive focus on student outcomes, and encourage schools to meet specific externally-defined standards.

Some educational systems hold schools accountable for their outcomes. This approach is linked to market-oriented reforms and is designed to improve programs and society (Alkin and Christie 2004). For instance, rewards and penalties are considered to change the behaviors of stakeholders in ways that improve student achievement (Wößmann et al. 2009). In addition, accountability of schools is likely

to be desirable to taxpayers and other stakeholders (Scheerens et al. 2003; Simons 2002). Accountability practices may also refer to the public availability of assessment and evaluation results (Scheerens et al. 2003). Such information could be used by parents for school choice (Kellaghan and Stufflebeam 2003), or by local communities for resource allocation.

In some countries, the evaluation of teachers and holding them accountable is a common practise (Faubert 2009; Santiago and Benavides 2009), and this has become an important field of research (Hallinger et al. 2014). Thus, we address this kind of evaluation in more detail in the following section.

19.2.3 Teacher Evaluation

Barber and Mourshed (2007) analyzed 25 educational systems in order to examine commonalities among the highest performing school systems. They concluded that teacher quality made the largest difference in student achievement, but it was not tied to the teachers' qualifications. Instead, there are hints that rigorous evaluation programs enhance teacher effectiveness and student performance (Taylor and Tyler 2011). Such findings strengthen the international move towards teacher evaluation policies. However, the effects of teacher evaluation on student achievements are not so clear. In a synthesis of several research studies Goe (2007) found that empirical research leads to different results. Only for the subject of mathematics did the evaluation of teaching and teachers show a clear and positive relationship with student outcomes. More recently, the review published by Hallinger et al. (2014) uncovered a large gap between policy logic and empirical evidence, concluding that teacher evaluation may actually be one of the less-efficient strategies for school improvement.

The policy logic of teacher evaluation assumes that teachers need feedback on their performance to help them identify how to better shape and improve their teaching practice. Holding teachers accountable for student learning outcomes and providing different kinds of appraisal is expected to promote improvement, or alternatively to support the laying off of "ineffective" teachers. Teacher evaluation policies may also provide a mechanism to recognize and reward high-quality teaching and to manage teacher career advancement (Mead et al. 2012). Accountability policies vary widely across educational and cultural systems; from centralized national systems to informal approaches developed at the discretion of individual schools, and from informal recommendations (e.g., Ireland, Iceland) to financial sanctions or rewards (e.g., Czech Republic, Flemish Community of Belgium; OECD 2013). Earlier research has shown that effective teacher evaluation is related to a collaborative and supportive environment, evaluation purposes having been agreed to by all stakeholders, strong educational leaders, and the use of multiple sources to gather data (Colby et al. 2002), as well as to teachers' involvement in their evaluation processes (Papanastasiou 1999).

Some systems incorporate student growth on test scores in ways that aim to capture the contribution teachers make toward student achievement—often referred to as teacher value-added (Glazermann et al. 2011). In the US, there is an ongoing debate on whether and how effective teaching can be measured (Kane et al. 2013; Whitcomb 2014). Effectiveness is mostly conceptualized as an attribute of the individual teacher that may be assessed by measures of teacher qualifications (assessment of teacher knowledge), process measures (observer or student ratings) and product measures (value-added student test scores). The research debate focuses on technical issues, such as how multiple measures should be integrated (Kane and Staiger 2012), how value-added measures should be defined (Goldhaber et al. 2013), how reliable and valid these measures are (Haertel 2013). From a policy perspective, side effects on teacher motivation and professionalism, as well as local strategies undermining the validity of the data, need to be monitored and discussed carefully.

19.2.4 Student Assessment

Several skills are relevant in student learning. Non-cognitive outcomes like motivation, self-effort and collaboration seem to be connected to student achievement, and these have been increasingly focused on in recent years. However, the assessment of such non-cognitive outcomes is a challenge, especially when transparency and comparability standards must be met. Thus, student achievement is still the core business of student assessment (Guskey 2012).

There are several ways to assess students' knowledge and progress. Figure 19.1 provides a rough outline of the most common forms of assessment along the dimensions of standardization and purpose, bearing deviations in mind. In addition, teachers may combine several assessment methods to gather evidence about their students' ideas and skills. A good description of different forms of assessment has been provided by Harlen (2007).

In its summarizing function, assessment takes place in order to grade, certify or record progress. A summative assessment, whether external or internal, therefore indicates and monitors standards, but it may also raise standards by causing students, as well as teachers and schools, to invest more effort in their work (Harlen and Deakin Crick 2002). On the other hand, summative assessment might lead to lower self-esteem and diminished effort in students at risk, which will increase the gap between lower- and higher-achieving students (Black and Wiliam 2004). Another side effect can emerge if teachers neglect skills and knowledge development in opting rather to train their students in test-taking strategies (Harlen and Deakin Crick 2002).

Apart from summative assessments, formative assessment plays a key role in classroom learning (e.g., Shepard 2006; Black and Wiliam 2004; McMillan 2007; OECD 2005). Several meta-analyses indicate that formative assessment is a significant source of improvement in student learning processes. In particular, low achievers benefit from formative assessment, which can lead to sizable gains in student

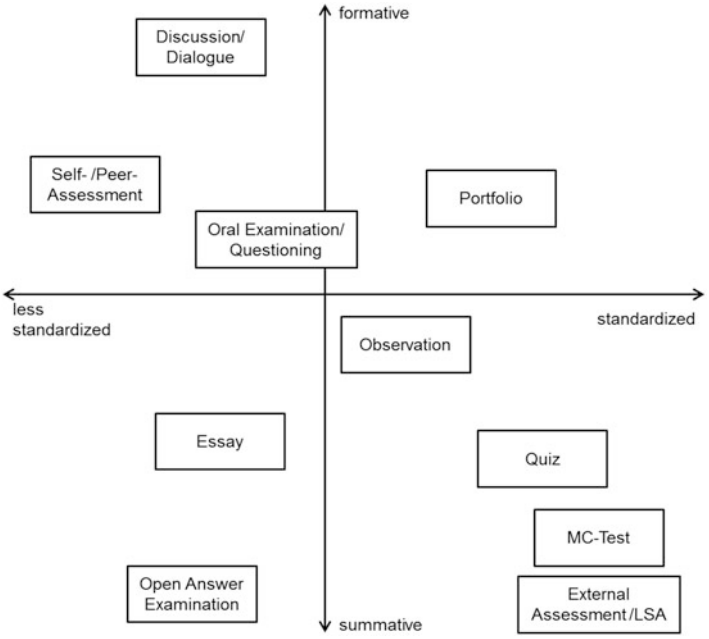

Fig. 19.1 Forms of assessment

achievement (Abrams 2007). However, there is large variation in the implementation and in the impacts of formative assessment (e.g., Bennett 2011; Kingston and Nash 2011; Shute 2008; Hattie and Timperley 2007; Black and Wiliam 1998).

Feedback plays a key role in formative assessment. Hattie and Timperley (2007) have identified four types of feedback that have differential effects on student learning. Accordingly, feedback may refer to (1) the student, evaluating him or her on a personal level, (2) task performance, (3) task processing and (4) self-regulation (see also Kingston and Nash 2011). Most commonly, feedback is given about task performance (2; also called corrective feedback). This feedback can be useful if the recipient uses it to reconsider and if necessary adapt their strategies or to enhance self-regulation. Otherwise, feedback can explicitly refer to processes to solve a specific kind of task (3) or to non task-specific strategies (4): for example, how to learn, or how to structure a learning process. The latter two types of feedback have been shown to be the most effective, but learners need to know how to incorporate the feedback into their thinking. Feedback on a personal level (1; e.g., "you are a nice student") is less effective. In general, feedback to students needs to be simply coded, and suggestions need to be realistic (Sadler 1989). Feedback that meets these conditions will allow students to understand the gap between the intended learning goal and what they have achieved so far, and take appropriate steps. In addition, formative assessment and reciprocal feedback might be useful for teachers as well, helping them to adapt their instruction to their students' needs (Black and Wiliam 2004). When teachers gather evidence about students' knowledge and understanding, they

are simultaneously considering which teaching practices would work and what new strategies are needed (Shepard 2006).

A tool in the context of classroom assessment that is suitable for both formative and summative feedback, is grading. The summative role of grading becomes evident when each individual grade a student might earn in a class has a marked impact on a student's educational career. Grades, as represented by the grade point average, usually play a critical role in promotion, allocation or selection (Guskey 2007). On the other hand, an effective marking system provides the individual student with formative information that directly relates to the progress made in relation to the objectives that are to be learned (Haptonstall 2010). In addition, teachers may use grades to motivate students by rewarding certain behaviors and signaling what attitudes, behaviors and habits are valued in school (OECD 2012). Research has revealed a lack of validity in the assignment of grades: The judgment of student outcomes can be subjective and can be influenced by different aspects, depending on teachers' perceptions of what grading is about (Guskey 2012), and grading can be biased by prejudices relating to a student's sex, past performance or social background (Archer and McCarthy 1988). However, aspects of teacher competence (assessment literacy: DeLuca et al. 2015) and teaching quality (classroom management: Hochweber et al. 2014) have been shown to increase judgment accuracy and diminish bias in grading. Mixing several aspects of attitude, effort and achievement in grading may even increase predictive validity (Cross and Frary 1999; Brookhart 2004; Rakoczy et al. 2008). Variability in grading concerns comparability both across and within educational systems (Haptonstall 2010).

19.3 Measuring the Policies and Practices of System Monitoring, School Evaluation, and Student Assessment

The conceptual background of system monitoring, school evaluation, teacher evaluation, and student assessment has been furnished in the conceptual framework above. The research findings referred to are mostly national or experimental, with some data being based on international comparative studies. In line with the goal of this book, we encourage researchers to use large-scale surveys, especially international student assessment systems, to document and understand policies and practices in the field of assessment and evaluation.

In former PISA cycles (2000–2012), the school questionnaires already addressed policies of evaluation and assessment, and how results have been used within countries (see Table 19.1). Thus, existing PISA trend data helps us understand how the use of student assessments has widened over the past 15 years in almost all OECD countries (Teltemann and Klieme in press). In the PISA 2015 field trial, the authors, in close collaboration with the International Questionnaire Expert Group developed and implemented a broader set of questions, also covering details of school

evaluation and classroom assessment. The new and more systematic set of measures was informed by the research reviewed above.

The following sections provide an overview of the measures that were implemented in the PISA 2015 field trial. In doing so, we refer to the list of constructs that is included in this chapter (Table 19.1). This overview is arranged in a similar way, as follows: (a) *Policies* of assessment and evaluation, (b) *use* of assessment and evaluation results, and *practices* regarding (c) school evaluation and (d) classroom assessment with a special focus on (e) formative assessment. Several constructs have been measured from different perspectives in the field trial: i.e., the school leaders' as well as the teachers' perspectives for evaluation measures, and the teachers' (TC) as well as the students' (ST) perspectives on classroom assessment prac-

Table 19.1 List of constructs included in the PISA 2015 field trial to assess assessment and evaluation in educational contexts

Theoretical relation	Name of construct	PISA 2015 ID	Included in PISA 2015 main survey
Policies	Teacher evaluation	SC032	YES
	General assessment practice	SC034	YES
	Measures for school improvement, including internal and external evaluation	SC037	YES
	Existence of internal evaluation	TC063	NO
	Teacher evaluation	TC067	NO
Use of assessment	Teacher incentives	SC033	NO
	Purpose of assessment results	SC035	YES
	Use of achievement data for accountability	SC036	YES
	Teacher incentives	TC068	NO
School evaluation practices	Foci of internal evaluation	SC038	NO
	Processes of internal evaluation	SC039	NO
	Consequences of internal evaluation	SC040	YES
	Processes of external evaluation	SC041	YES
	Foci of internal evaluation	TC064	NO
	Processes of internal evaluation	TC065	NO
	Consequences of internal evaluation	TC066	NO
Classroom assessment practices	Classroom assessment instruments	TC054	YES
	Teachers' grading practices	TC055	YES
Formative assessment	Perceived feedback	ST104	YES
	Source of feedback	ST105	NO
	Use of feedback to guide learning	ST106	NO
	Adaptation of instruction	ST107	YES
	Adaptation of instruction	TC038	NO

For detailed documentation see: https://doi.org/10.7477/150:174:1

Note. ID coded ST for student questionnaire; SC for school questionnaire; TC for teacher questionnaire; EC for educational career questionnaire; IC for ICT familiarity questionnaire; PA for parent questionnaire

tices. Below, we summarize our recommendations for the preferred source as well as our thoughts on which measures might be practicable in further national and international large-scale assessments.

19.3.1 Policies of Assessment and Evaluation

Evaluating, monitoring and comparing educational systems requires descriptive information on overall school evaluation policies and student assessment policies. It is essential to know whether certain measures for school improvement, including internal and external evaluations of schools (SC037) are common practices, and how often students are assessed through highly standardized tests, teacher-made tests or through teachers' judgmental rating (General assessment practice, SC034). Moreover, the impetus for action is also relevant, in order to analyze system policies. Thus, the PISA 2015 items referring to school improvement policies (SC037) or standardized testing (SC034) distinguish action that is mandatory, required by educational policies, and action that is based on the school's initiative.

Countries show high variation in evaluation and assessment activities (OECD 2007, 2010), but as these are mostly determined by national or state policies, less variation is to be expected within countries. In contrast, methods used for teacher evaluation (SC032) differ across countries and even vary within countries, as we know from TALIS (OECD 2014). Thus, the assessment of teacher evaluation policies is relevant at both national and international levels.

For all three questions mentioned so far, partially comparable data are available from previous PISA cycles (PISA 2012 for SC037 and SC03; PISA 2000–2009 for SC034). In parallel to the school questionnaire, items on internal school evaluation and teacher evaluation were also implemented in the teacher questionnaire (TC067 and TC063).

19.3.2 Use of Assessment and Evaluation Results

The way student assessment and school evaluation results are used differs across educational systems, and is subject to change. To support the description and analysis of data use, we took up a set of items from previous PISA cycles (2000–2012) addressing various kinds of usage for student assessment results, such as informing parents, deciding upon student promotion, or comparing the school with other schools (Purpose of assessment results; SC035). Some items on formative use (e.g., guiding student learning and adapting teaching) were newly added, and the response format was changed, with the intention to discriminate the use of standardized tests from use of teacher-developed tests. However, field trial results showed that missing rates increased, suggesting that the definition of standardized vs. teacher-developed tests might not always be applicable to all countries. Another question, on the use of

assessment results, refers directly to different accountability strategies (use of achievement data for accountability; SC036). This question has been used since PISA 2006. However, some items have changed over time. The items used in the PISA 2015 field trial address the debate whether student achievement results should be published, tracked over time and/or provided to parents directly.

Finally, a question on Teacher incentives (SC033) was taken over from TALIS 2013. Its items address formative and improvement strategies, as well as summative and accountability purposes. It was complemented by a question to teachers, also from TALIS 2013, asking about consequences of teacher feedback (TC068).

Policies on accountability, especially sanctions or rewards to teachers, differ strongly across countries, as does a country's tendency to use students' outcomes in a formative or summative manner. Further research on accountability policies and data use, based on the questions introduced here, may enrich the debate regarding the positive and negative effects of accountability systems. For example, the ongoing debate on formative and summative use of data would benefit from longitudinal studies.

19.3.3 School Evaluation Practices

Evaluation practices differ in respect of the initiators and enacting agents, of responsibilities and instruments, across and within countries (Faubert 2009). In order to describe evaluation systems more precisely and enrich the interpretation of student achievements, fine-grained information on foci, processes and consequences of evaluation should be assessed in national and international studies. Processes clearly depend on the evaluation purpose and the initiator. Therefore, it is indispensable to explore internal and external evaluation processes separately. Altogether, the constructs mentioned were covered in the school questionnaire through questions SC038 (foci), SC039 (processes) and SC40 (consequences) for internal evaluation, and SC041 (processes) for external evaluation; all newly developed. The questions on internal evaluation were paralleled in the teacher questionnaire (TC064 to TC066).

In the PISA 2015 field trial, question SC038 included a rather long list of possible topics that an internal evaluation could focus on, ranging from school resources through the quality of teaching and teacher cooperation, to equity in school. School principals reported high proportions of coverage across all items, which may or may not reflect the degree of social desirability in their answers. In contrast, when exactly the same list of topics was used to ask whether specific measures in any of these areas had been implemented as a consequence of internal evaluation (SC040), responses were more differentiated..

The two questions on change processes associated with internal evaluation (SC039) and external evaluation (SC041), respectively, were largely parallel. Items included statements such as "The results led to changes in school policies", as well as more negative reports: for example, "The impetus triggered by the evaluation

'disappeared' very quickly at our school". For internal validation, once again ceiling effects were observed, whereas for external evaluation, approval rates were lower and the items turned out to be valid and relevant.

19.3.4 General Classroom Assessment Practices

According to the model of planned behavior (Ajzen 2005), there is strong evidence that beliefs about the nature and the purpose of assessment influence assessment techniques and practices (Brown 2012). Consequently, a full model of assessment should take teachers' assessment beliefs into account. In a cross-cultural comparison of teacher conceptions of assessment, Brown (2012) found evidence that teacher belief systems differ between cultures, while they seem to be consistent within a culture. However, those kinds of questions are prone to be non-equivalent across cultural groups. If the measure is biased against one or some cultural groups, individual differences within a cultural population and across cultural populations are not measured at the same scale (Van de Vijver 1998). From the onset of our preparation, we found cultural differences in teachers' understanding of items proposed for the measurement of teacher beliefs. Addressing these sensitive constructs across a large number of countries presents a challenge, and careful and thorough testing is required. Thus, PISA 2015 did not include a measure of teachers' assessment beliefs—either in the main survey or in the field trial.

Another construct that is sensitive to the cultural context concerns teachers' grading practices (TC055). Teachers were asked to self-report on the criteria they apply, and the sources of evidence they use in marking and grading students' work. For international use in the PISA 2015 main survey, we proposed a reduced scale for the subdimensions of individual judgment and criteria-based judgment. National studies or studies with a reduced set of countries, however, may implement wider aspects and make use of the full range of items.

The cultural background of respondents probably plays a less restrictive role regarding items on classroom assessment instruments (TC054, taken over from TALIS). The attainment of learning goals and educational standards needs to be monitored. To this end, teachers use several assessment methods—often in combination—to gather evidence about their students' knowledge and skills in relation to the learning goals. The PISA 2015 field trial indicated relevant variation across countries.

19.3.5 Formative Assessment in the Classroom

Arguably the most prominent form of classroom assessment covered in empirical research is formative assessment (see theoretical and conceptual background above). Since feedback is essential in formative assessment, we tested several facets

of this concept in the PISA 2015 field trial student questionnaire. First of all, we assessed whether students perceived (formative) feedback at all (ST104), asking how often the teacher would tell the student about his or her strengths and weaknesses, how often he or she would receive advice on how to reach the learning goals, etc.

In addition, we asked whether the frame of reference used in giving the feedback was criterion-oriented, social-comparative, or individual (source of feedback, ST105). Research predicted that individual feedback would support student learning and motivation best. Furthermore, we wanted to learn more about two types of use of feedback: students' use of feedback to guide learning (ST106), and teachers' adaptation of instruction (ST107). Among these constructs, only adaptivity of teaching—which is also an important indicator of teaching quality—was kept for the main survey.

We also intended to implement questions on the level of feedback (whether it addresses the student's character and behavior, task performance, task processing or self-regulation) identified by Hattie and Timperley (2007) as impacting on students' improvement. For national studies, this construct could be a relevant predictor of students' effort.

19.4 Conclusion

In the PISA 2015 field trial, an attempt was made to expand the framework of assessment and evaluation measures and to address concepts beyond the perspective of system monitoring and educational effectiveness. Furthermore, exploration of different kinds of concepts, item formats and perspectives was shown to be possible; however, the scope of the material had to be significantly reduced for the main survey. Time constraints allow for the consideration of just one perspective on any single construct. However, even higher-quality measures were not implemented in the main study, due to the reduced assessment time; policy relevance and the measure's reference to theoretical models were the criteria for selection. Table 19.1 above provides an overview of the measures realized in the PISA 2015 field trial and the PISA 2015 main survey.

References

Abrams, L. M. (2007). Implications of high-stakes testing for the use of formative classroom assessment. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 79–98). New York/London: Teacher College, Columbia University.

Ajzen, I. (2005). *Attitudes, personality, and behavior* (2nd ed.). Maidenhead/New York: Open University Press.

Alkin, M. (1972). Evaluation theory development. In C. Weiss (Ed.), *Evaluation action programs* (pp. 105–117). Boston: Allyn and Bacon.

Alkin, M., & Christie, C. A. (2004). An evaluation theory tree. In M. Alkin (Ed.), *Evaluation roots tracing theorists' views and influences* (pp. 12–65). Thousand Oaks: Sage.

Altrichter, H., & Maag Merki, K. (2016). *Handbuch Neue Steuerung im Schulsystem* (2nd ed.). Wiesbaden: Springer VS.

Archer, J., & McCarthy, B. (1988). Personal biases in student assessment. *Educational Research, 30*(2), 142–145.

Barber, M., & Mourshed, M. (2007). *How the world's best-performing school systems come out on top*. New York: McKinsey and Co.

Bennett, R. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25.

Berkemeyer, N., & Müller, S. (2010). Schulinterne evaluation: Nur ein Instrument zur Selbststeuerung von Schulen? [Internal school-based evaluation: Only a tool for self-management?]. In H. Altrichter & K. Maag Merki (Eds.), *Handbuch Neue Steuerung im Schulsystem* (1st ed., pp. 195–218). Wiesbaden: Springer VS.

Bischof, L. M., Hochweber, J., Hartig, J., & Klieme, E. (2013). Schulentwicklung im Verlauf eines Jahrzehnts: Erste Ergebnisse des PISA-Schulpanels [School improvement throughout one decade: First results of the PISA school panel study]. *Zeitschrift für Pädagogik, special issue, 59*, 172–199.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74.

Black, P., & Wiliam, D. (2004). The formative purpose. Assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: 103rd yearbook of the national society for the study of education, Part II* (pp. 20–50). Chicago: University of Chicago Press.

Blöchliger, H. (2013). *Decentralisation and economic growth—part 1: How fiscal federalism affects long-term development* (OECD working papers on fiscal federalism, No. 14). Paris: OECD Publishing.

Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record, 106*(3), 429–458.

Brown, G. T. L. (2012). Prospective teachers' conceptions of assessment: A cross-cultural comparison. *The Spanish Journal of Psychology, 15*(1), 75–89.

Coburn, C., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research and Practice, 9*(4), 173–206.

Colby, S. A., Bradshaw, L. K., & Joyner, R. L. (2002). *Teacher evaluation: A review of literature*. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

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

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*(1), 53–72.

de Boer, H., Enders, J., & Schimank, U. (2007). On the way towards new public management? The governance of university systems in England, the Netherlands, Austria and Germany. In D. Jansen (Ed.), *New forms of governance in research organizations* (pp. 137–152). Dordrecht: Springer.

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2015). Teacher assessment literacy: a review of international standards and measures. *Educational Assessment, Evaluation and Accountability, 28*, 1–22. doi:10.1007/s11092-015-9233-6.

Donaldson, S. I. (2004). Using professional evaluation to improve the effectiveness of nonprofit organizations. In R. E. Riggo & S. S. Orr (Eds.), *Improving leadership in nonprofit organizations* (pp. 234–251). San Francisco: Wiley.

Elacqua, G. (2016). Building more effective education systems. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: An international perspective*. Dordrecht: Springer.

European Commission. (2011). *Progress towards the common European objectives in education and training: Indicators and benchmarks 2010/2011* (Commission staff working document based on document SEC(2011)526)). Luxembourg: European Union.

Faubert, V. (2009). *School evaluation: Current practices in OECD countries and a literature review* (OECD Education Working Papers, No. 42). Paris: OECD Publishing.

Glazermann, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D., & Whitehurst, G. J. (2011). *Passing muster: Evaluating teacher evaluation systems*. Washington, DC: The Brookings Brown Center Task Group on Teacher Quality.

Goe, L. (2007). The link between teacher quality and student outcomes: A research synthesis. Washington, DC: National Comprehensive Center for Teacher Quality. http://www.gtlcenter.org/sites/default/files/docs/LinkBetweenTQandStudentOutcomes.pdf. Accessed 17 June 2016.

Goldhaber, D. D., Goldschmidt, P., & Tseng, F. (2013). Teacher value-added at the high-school level. Different models, different answers? *Educational Evaluation and Policy Analysis, 35*(2), 220–236.

Guskey, T. R. (2007). Multiple sources of evidence. An analysis of stakeholders' perceptions of various indicators of student learning. *Educational Measurement: Issues and Practice, 26*(1), 19–27.

Guskey, T. R. (2012). Defining students' achievement. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement. Educational psychology handbook series* (pp. 3–6). New York/London: Routledge.

Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores*. Princeton: Education Testing Service. https://www.ets.org/Media/Research/pdf/PICANG14.pdf. Accessed 17 June 2016.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability, 26*(1), 5–28.

Hanushek, E. A., Link, S., & Wößmann, L. (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. *Journal of Development Economics, 104*, 212–232.

Haptonstall, K. G. (2010). *An analysis of the correlation between standards-based, non-standards-based grading systems and achievement as measured by the Colorado Student Assessment Program (CSAP)* (Doctoral dissertation). Colorado: ProQuest, UMI Dissertation Publishing.

Harlen, W. (2007). Formative classroom assessment in science and mathematics. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 116–135). New York/London: Teachers College Press, Columbia University.

Harlen, W., & Deakin Crick, R. (2002). *A systematic review of the impact of summative assessment and tests on students' motivation for learning* (EPPI-Centre Review, version 1.1*). London: EPPI-Centre. https://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/ass_rv1.pdf?ver=2006-02-24-112939-763. Accessed 17 June 2016.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Hochweber, J., Hosenfeld, I., & Klieme, E. (2014). Classroom composition, classroom management, and the relationship between student attributes and grades. *Journal of Educational Psychology, 106*(1), 289–300.

Hofman, R. H., Dijkstra, N. J., & Hofman, W. H. A. (2009). School self-evaluation and student achievement. *School Effectiveness and School Improvement, 20*(1), 47–68.

Huber, S. G., & Skedsmo, G. (2016). Editorial: Data use—a key to improve teaching and learning. *Educational Assessment, Evaluation and Accountability, 28*(1), 1–3.

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation, 30*(3), 377–410.

Jude, N. (2016). The assessment of learning contexts in PISA. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: An international perspective*. Dordrecht: Springer.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: combining high-quality observations with student surveys and achievement gains* (Research paper, MET Project). Seattle: Bill & Melinda Gates Foundation. http://files.eric.ed.gov/fulltext/ED540960.pdf. Accessed 17 June 2016.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment* (Research paper, MET Project). Seattle: Bill & Melinda Gates Foundation. http://www.hec.ca/iea/seminaires/140401_staiger_douglas.pdf. Accessed 17 June 2016.

Kellaghan, T., & Stufflebeam, D. L. (Eds.). (2003). *International handbook of educational evaluation. Part one: Perspectives/part two: Practice*. Dordrecht: Kluwer Academic Publishers.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37.

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

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie/Journal of Psychology, 216*(2), 61–73.

Kuger, S., & Klieme, E. (2016). Dimensions of context assessment. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: An international perspective*. Dordrecht: Springer.

McMillan, J. H. (2007). Formative classroom assessment: The key to improving student achievement. In J. H. McMillan (Ed.), *Formative classroom assessment. Theory into practice* (pp. 1–7). New York/London: Teacher College, Columbia University.

Mead, S., Rotherham, A., & Brown, R. (2012). *The hangover: Thinking about the unintended consequences of the nation's teacher evaluation binge. Teacher Quality 2.0, Special Report 2*. Washington, DC: American Enterprise Institute. http://bellwethereducation.org/sites/default/files/legacy/2012/09/Teacher-Quality-Mead-Rotherham-Brown.pdf. Accessed 17 June 2016.

Nevo, D. (1998). Dialogue evaluation: A possible contribution of evaluation to school improvement. *Prospects, 28*(1), 77–89.

Nevo, D. (2002). Dialogue evaluation: Combining internal and external evaluation. In D. Nevo (Ed.), *School-based evaluation: An international perspective* (pp. 3–16). Amsterdam/Oxford: Elsevier Science.

OECD. (1989). *Schools and quality: An international report*. Paris: OECD.

OECD. (2005). *Formative assessment: Improving learning in secondary classrooms*. Paris: OECD.

OECD. (2007). *PISA 2006: Science competencies for tomorrow's world* (Vol. 1). Paris: OECD.

OECD. (2010). *PISA 2009 results: What students know and can do*. Paris: OECD.

OECD. (2012). *Grade expectations: How marks and education policies shape students' ambitions*. PISA. Paris: OECD.

OECD. (2013). *Synergies for better learning. An international perspective on evaluation and assessment. OECD reviews of evaluation and assessment in education*. Paris: OECD.

OECD. (2014). *TALIS 2013 results: An international perspective on teaching and learning* (Revised version). TALIS.

Papanastasiou, E. C. (1999). *Teacher evaluation: Theories and practices*. ERIC. http://files.eric.ed.gov/fulltext/ED439157.pdf. Accessed 17 June 2016.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks: Sage.

Rakoczy, K., Klieme, E., Bürgermeister, A., & Harks, B. (2008). The interplay between student evaluation and instruction. *Zeitschrift für Psychologie, 2*, 111–124.

Ryan, K. E., Chandler, M., & Samuels, M. (2007). What should school-based evaluation look like? *Studies in Educational Evaluation, 33*(3–4), 197–212.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.

Sanders, J. R., & Davidson, E. J. (2003). A model for school evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation. Part one: Perspectives/ part two: Practice* (pp. 807–826). Dordrecht: Kluwer Academic Publishers.

Santiago, P., & Benavides, F. (2009). *Teacher evaluation: A conceptual framework and examples of country practices*. Paris: OECD.

Scheerens, J. (2002). School self-evaluation: Origins, definitions, approaches, methods and implementation. In D. Nevo (Ed.), *School-based evaluation: An international perspective* (pp. 35–69). Amsterdam/Oxford: Elsevier Science.

Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford: Emerald.

Scheerens, J., Glas, C. A., & Thomas, S. M. (2003). *Educational evaluation, assessment, and monitoring. A systemic approach*. Lisse/Exton: Swets & Zeitlinger.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (pp. 623–646). Westport: Rowman and Littlefield Publishers.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Simons, H. (2002). School self-evaluation in a democracy. In D. Nevo (Ed.), *School-based evaluation: An international perspective* (pp. 17–34). Amsterdam/Oxford: Elsevier Science.

Spillane, J. P. (2012). Data in practice: Conceptualizing the data-based decision-making phenomena. *American Journal of Education, 118*(2), 113–141.

Stufflebeam, D. L. (2003). The CIPP model for evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation. Part one: Perspectives/part two: Practice* (pp. 31–62). Dordrecht: Kluwer Academic Publishers.

Taylor, E. S., & Tyler, J. (2011). *The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers*. NBER Working Paper 16877. Cambridge, MA.

Teltemann, J., & Klieme, E. (in press). The impact of international testing projects on policy and practice. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 369–386). New York: Routledge.

Torrance, H. (Ed.). (2013). *Educational assessment and evaluation: Major themes in education*. New York: Routledge.

Van de Vijver, F. J. (1998). Towards a theory of bias and equivalence. *Zuma Nachrichten Spezial, 3*, 41–65.

Visscher, A. J., & Coe, R. (2003). School performance feedback systems: Conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement, 14*(3), 321–349.

Whitcomb, J. (2014). *Review of "Fixing classroom observations"*. Boulder: National Education Policy Center. http://nepc.colorado.edu/thinktank/review-fixing-classroom-observations. Accessed 17 June 2016.

Wößmann, L. (2003). Schooling resources, educational institutions, and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics, 65*(2), 117–170.

Wößmann, L., Lüdemann, E., Schütz, G., & West, M. R. (2009). *School accountability, autonomy and choice around the world*. Cheltenham: Edward Elgar.

Wyatt-Smith, C. (2014). *Designing assessment for quality learning: The enabling power of assessment*. Heidelberg: Springer.