Chapter 7
Policies and Practices of Assessment: A Showcase for the Use (and Misuse) of International Large Scale Assessments in Educational Effectiveness Research

Eckhard Klieme

7.1 International Large Scale Assessment (ILSA) and Educational Effectiveness Research (EER)

International Large Scale Assessments (ILSAs) are international assessments of educational topics that target large and representative samples of students and/or teachers, as well as other stakeholders in education such as school principals or parents. They started around 1960 with the first studies of the International Association for The Evaluation of Student Achievement (IEA). Today, the most cited studies include the Progress in International Reading Literacy Study (PIRLS, run by IEA), the Programme for International Student Assessment (PISA, run by OECD) and Trends in International Mathematics and Science (TIMSS, run by IEA).

The overarching and initial goal of such ILSAs is to provide indicators on the effectiveness, equity, and efficiency of educational systems (Bottani & Tuijnman, 1994), to set benchmarks for international comparison, to monitor trends over time and thus inform educational policy on an international, national, regional and even local (school) level, e.g. with regard to innovations in educational governance and curriculum (Klieme & Kuger, 2015). Consequently, ILSAs have attracted much media attention in many countries and have exerted sometimes far-reaching influence on education policy. Bogdandy and Goldmann (2009), scholars in international public law, claim that ILSAs even allow international organizations like OECD to establish a new legal mechanism they call "governance by information". In addition to educational politics, administration and the public, researchers increasingly draw on the results of these assessments to study, on the one hand, the universality and generalizability of certain findings in educational effectiveness and, on the other

E. Klieme (✉)
DIPF| Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany
e-mail: klieme@dipf.de

J. Hall et al. (eds.), *International Perspectives in Educational Effectiveness Research*,
https://doi.org/10.1007/978-3-030-44810-3_7

hand, the respective national, regional, cultural, and other group-specific features that may moderate universal mechanisms.

Scholars repeatedly claimed that studies must take into account theoretical considerations, modeling approaches and research results from Educational Effectiveness Research (EER) to develop a meaningful system of reporting indicators (e.g., Bryk & Hermanson, 1994). However, when the U.S. National Research Council attempted to summarize what had been learned from more than three decades of ILSA up to the year 2000, in the concluding chapter Rowan (2002, p. 338) argued for a deeper analysis of "relationships among school characteristics and student achievement" (p. 338) and "effects of educational practices" (p. 339). Thus, ILSAs had failed to align with EER in the twentieth century. One reason was that ILSA designs had almost exclusively focused on student tests, and most of the administration time had been spent with students working on test items, while only a small number of student, classroom, school or even system-level characteristics were measured in so-called "background questionnaires". Fortunately, this situation has changed over the last decade, especially as PISA introduced "analytical frameworks" (most recent: Klieme & Kuger, 2015) which systematically linked questionnaire development to policy issues, research questions, and scientific constructs informed by the growing knowledge base of EER. In a recent review of ILSA questionnaire design, Jude and Kuger (2018, p. 5) mention three advantages of basing ILSA frameworks on EER:

> (1) EER acknowledges the complexity of educational systems, (2) EER frameworks ultimately aim at explaining student outcomes, and (3) overarching EER theories offer a number of different anchors to relate to other, interdisciplinary theories or frameworks.

Vice versa, ILSAs can as well contribute to the further development of EER (Klieme, 2012) by providing data, triggering new research studies, and providing instruments which work across multiple cultures. First, ILSA data are based on large representative samples assessed in multiple countries, usually with high quality; yet they are easily accessible. Researchers may use these data for fruitful secondary analysis, both within and across countries. Opportunities and limitations of using such data, including the implementation of enhanced designs, will be discussed in the present chapter; for a recent overview see Singer, Braun, and Chudowsky (2018). Second, the rich description of education in different cultures, school systems, and *s*chool contexts that is provided by ILSA studies – although limited by its descriptive nature – can inspire EER to discover new fields of research. National and international patterns or regional peculiarities are easily accessible through screening publicly available ILSA data, and can trigger new research questions that lead to the careful development of smaller, targeted EER studies. For example, PISA findings on disadvantages for migrant students and students from low SES families motivate EER to have a closer look at differential school effectiveness. Third, ILSAs carefully develop research instruments and methodology that may be used in further studies both within and across countries. The high quality standards typically involved in the preparation and implementation of ILSAs, including sophisticated procedures for translation, adaptation, administration, data cleaning, and scaling,

provide EER with high quality, culturally adapted, policy-relevant material in a large number of languages to support innovative EER studies.[1] Therefore, ILSAs offer an unmatched source of ready-to-use instruments for EER that has been developed and refined under strict quality guidelines and discussed by education, policy, questionnaire, and survey method experts.

When using ILSA data in the context of EER, researchers need to cope with typical limitations of ILSA designs, above all with the cross-sectional nature of data and the lack of cross-cultural comparability. On one hand, state-of-the-art Item Response Theory (IRT) methods are applied, missing data are treated in very sophisticated ways, and complex, multi-level models are used for analysis. On the other hand, the cross-sectional nature of the data severely limits any interpretation in terms of the direction of relationships in educational effectiveness or ability to infer causality. Unobserved confounding variables such as prior achievement might explain existing patterns and relationships, and the direction of causality might not be clear. At least, this holds for interpretations of findings made at the student, class, or school level. To overcome these limitations, researchers, especially in Germany, have added follow-up measures to study conditions of student learning over one school year (see Baumert et al., 2010, for a study of teacher effectiveness based on a longitudinal extension of PISA 2003; Kuger, Klieme, Lüdtke, Schiepe-Tiska, & Reiss, 2017, for a study of teaching quality enhancing PISA 2012) and effects of school policies such as internal evaluation and all-day-programmes on long-term change in school climate and school outcomes (Bischof, Hochweber, Hartig, & Klieme, 2013, based on a sample of schools participating both in PISA 2000 and in PISA 2009). Other scholars have been focusing on the country or "system" level, using repeated measures from trend studies such as PISA and TIMSS, analyzed through difference-in-difference estimation (Gustafsson, 2007), or fixed effects techniques (Bergbauer, Hanushek, & Wößmann, 2018).

Whenever analyses are run across countries, they require some level of measurement invariance (van de Vijver & He, 2016). In particular, "metric invariance" typically is required in order to warrant the claim that the construct of interest has the same meaning across countries. Running linear models, e.g. regression analyses, in parallel for a number of countries and comparing effect sizes are legitimate only if metric invariance can be established. Furthermore, a higher level of invariance, namely "scalar invariance", is required to allow mean differences on the construct to be compared across countries. Researchers who have checked invariance most often found that metric invariance can be established for many questionnaire scales, while scalar invariance seems to be a rare exception (He, Buchholz, & Klieme, 2017; He & Kubacka, 2015).

[1]Usually, questionnaires are published in the source language (mostly English) only. For PISA 2015, translated versions from 75 countries, item- and scale-level statistics are available at https://daqs.fachportal-paedagogik.de/search/show/survey/177?language=en. This online depository includes Field Trial material not yet used in the Main Study. For an introduction and conceptual overview, see Kuger, Klieme, Jude and Kaplan (2016).

Because of these limitations, policy making has often been misinformed and misled by shortcut interpretations and too-far-reaching conclusions (Baker, 2009). For example, based on PISA 2006 the OECD reported 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 socio-economic background of students and schools was accounted for" (OECD, 2007, p. 243). They concluded "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), although the reverse causality interpretation seems to be more realistic: schools might tend to publish their results if they have been successful in an assessment. More recently, OECD claimed that extracurricular activities and school climate would increase the proportion of "resilient" students in Germany (OECD & Vodafone Stiftung, 2018), while this finding was due to a neglect of the tracked structure of the German system (Klieme, 2018). There have also been examples of policy makers cherry picking results and making inappropriate claims about dramatic falls in student attainment to justify new and controversial 'education reforms'. For example in England a so called plummeting of student performance in PISA tests in country league tables was used to justify the introduction of the free school and academisation programme in 2010 although the actual performance of England was not statistically different in terms of the country reference group at the time (Jerrim, 2011). In Germany, a massive investment into after school programmes was publicly claimed to be a consequence from PISA 2012, which in fact it was not (Klieme, Jude, Baumert, & Prenzel, 2010). Within the research community, overstatements have been made as well. This includes the present author, who interpreted cross-sectional relationships between perceived teaching quality and student outcomes as indicative of teaching effects in PISA 2000 (Klieme & Rakoczy, 2003). More recently Schmidt, Burroughs, Zoido, and Houang (2015) drew far-reaching conclusions on the effects of opportunity to learn on mathematics achievement using cross-sectional PISA 2012 data and a mis-specified indicator for "applied mathematics" (as can be seen by comparing with OECD, 2014, pp. 56 and 324).

Rather than supporting claims on educational effectiveness proper, i.e. estimating the *effects* of specific policies and practices on student outcomes, ILSA data may be used to inform about the *distribution* of educational opportunities among students, families, schools, and regions. Policies and practices would be treated as dependent variables, while student achievement as well as student and family background would be treated as independent variables. E.g., do migrant students and students from socially disadvantaged families have an equal share of well-trained teachers, engaged school principals, of well-ordered, supportive and challenging classroom environments and out-of-class learning opportunities? Who receives differentiated instruction, supportive feedback and direct guidance from his or her teachers? Which schools report policies for assessment and evaluation, and which don't? Similarly, differential uptake of activities, use of opportunities and engagement in learning may be studied to understand inequity. E.g., does student truancy and attention in classroom differ between subpopulations? While studies on effectiveness typically

require experimental, or at least quasi-experimental designs, cross-sectional ILSAs are well prepared to answer questions about the provision of, differential access to, and differential use of learning opportunities. These kinds of questions may not be considered crucial in EER, but they are highly relevant for policy making and for understanding (in)equity in education.

In an attempt to illustrate the methodological issues raised above, to further explore the opportunities and limitations of ILSA data and to discuss their relevance for EER, the present chapter will use a specific showcase: policies and practices of educational assessment. This topic will be introduced in Sect. 7.2, while Sect. 7.3 will present and critically discuss related findings from PISA 2015. Thus, the chapter integrates three layers of academic discourse: (a) The meta-theoretical issue of how ILSAs relate to EER. (b) The substantive EER research question on how assessment is implemented in education, how it is shaped by systems and schools, and how it relates to student learning. (c) Specific methodological issues in analyzing ILSA data, which are discussed in several excursus spread across the chapter.[2]

7.2 Policies and Practices of Assessment as a Topic in Educational Effectiveness Research[3]

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 (Elacqua, 2016). 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 & Skedsmo, 2016, p. 1). This trend is part of an overarching change in concepts and measures of educational governance (Altrichter & Maag Merki, 2016). New forms of educational governance, such as school performance feedback systems (Visscher & Coe, 2003), systemic approaches to educational evaluation and monitoring (Scheerens, Glas, & Thomas, 2003) and concepts of data-driven school improvement (Coburn & Turner, 2011; Spillane, 2012) have become popular among policy makers. Over the years,

[2]The author wants to thank Anindito Aditomo, Sonja Bayer, Janine Buchholz, Jessica Fischer, Jia He, Nina Jude and Susanne Kuger for collaboration on this topic at the DIPF Department for Research on Educational Quality and Evaluation.

[3]This section is in part based on Bayer, S., Klieme, E. & Jude, N. (2016). Assessment and evaluation in educational contexts. In S. Kuger, E. Klieme, N. Jude & D. Kaplan (Eds.), *Assessing contexts of learning: An international perspective* (pp. 469–488). Cham: Springer.

the assessment/evaluation paradigm has shifted from a focus on measurement towards a focus on efforts to improve learning (Wyatt-Smith, 2014). Formative Assessment and Feedback to students have been shown to be among the most powerful tools teachers can use to boost their students' understanding and achievement (e.g., Bennett, 2011; Hattie, 2009; Kingston & Nash, 2011).

In the following, we broadly discriminate two areas of assessment and evaluation that seem to become increasingly popular around the globe: assessing and evaluating schools on the one hand, assessing and measuring student learning in the classroom on the other hand. In both cases, data may be used either for formative purposes, informing school improvement activities and classroom teaching, respectively, or for summative and accountability purposes, such as ranking schools with regard to national standards and issuing certificates for individual students. It should be noted, however, that test 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 alike.

7.2.1 School Evaluation

The evaluation of schools is an important instrument of educational governance used in decisions and judgments about processes, programmes, 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 to facilitate continuous improvement and organizational learning. The improvement of schools participating in evaluation programmes can be explained by feedback theory, (Visscher & Coe, 2003), or as an effect of stakeholders within school being held accountable for evaluation results (Donaldson, 2004):

- Feedback is a core element of data-driven school development (Scheerens et al., 2003), at best pushed by a combination of internal and external evaluation. Feedback may also be provided by national test programs allowing schools to compare their own performance with national standards. Scheerens et al. assume evaluation to be the fundamental process through 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.
- From an accountability perspective, rewards and penalties are assumed to change the behaviours of stakeholders in ways that improve student achievement (Wößmann, Lüdemann, Schütz, & West, 2009). Strong accountability practices include the public availability of assessment and evaluation results (Scheerens et al., 2003). Such information could be used by parents for school choice, or by local communities for resource allocation. Bergbauer et al. (2018) provide an econometric model based on principal-agent-theory, summarized as follows (p. 6): "By creating outcome information, student assessments provide a mechanism for developing better incentives to elicit increased effort by teachers and students, thereby ultimately raising student achievement levels to better approximate the desires of the parents".

School evaluation and improvement can indeed affect students' outcomes. For instance, Scheerens (2002) and also Creemers and Kyriakides (2008) report 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 et al. (2013) report that schools that had done some internal evaluation improved in terms of both student achievement and school climate.

Different evaluation practices generally coexist and benefit from each other (Ryan, Chandler, & Samuels, 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). In a review of 41 empirical studies on evaluation use, Johnson et al. (2009) found the involvement of stakeholders to be most important condition of effective school evaluations. Engagement, interaction, and communication between evaluation clients and evaluators are critical to the meaningful use of evaluations for improvement purposes.

Common steps of effective evaluation can be identified (e.g., Sanders & Davidson, 2003), yet school evaluation approaches are multifold and vary across educational systems (OECD, 2013). Therefore, it is difficult to report on and compare the effects of evaluation across different evaluation systems and education systems.

7.2.2 Assessment Embedded in Classroom Teaching and Learning

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

Apart from summative assessments, formative assessment plays a key role in classroom learning (e.g., Shepard, 2006; Black & 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 achievement (Abrams, 2007). However, there is large variation in the implementation and in the impacts of formative assessment (e.g., Bennett, 2011; Black & Wiliam, 1998; Hattie & Timperley, 2007; Kingston & Nash, 2011; Shute, 2008). Effects of formative assessment have been shown to be moderated by generic

teaching quality (Decristan et al., 2015) and by students' perception of usefulness (Rakoczy, Klieme, Leiss, & Blum, 2017).

Feedback plays a key role in formative assessment. Hattie and Timperley (2007) have identified four types of feedback provided to students that have differential effects on learning: 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 & 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 guide them to take appropriate steps.

7.2.3 *Using ILSAs to Inform Research on Assessment and Evaluation*

Assessments (ILSAs) like TIMSS, PIRLS and PISA are major 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. At the same time, these international surveys can be used as sources of information *about* assessment, evaluation and accountability practices in cross-national comparison. This is illustrated by the findings which will be presented in the remaining part of this chapter.

ILSA 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. Another advantage is the broad coverage of geographical areas and cultural contexts, which helps widening the scope of insights on this topic beyond the "Western", mostly English speaking world which dominated both policy and research on assessment and evaluation for a long time.

A recent, ground-breaking study using ILSA data from 59 countries is the paper entitled "Testing" by Bergbauer et al. (2018). Combining indicators from PISA 2000 to 2015 (mainly the ones discussed later in the present chapter) as well as from international comparative reviews of assessment policy, they claim to measure how strongly each of four different types of assessments has been implemented in a

country: (1) Standardized External Comparisons such as national tests or central exams. (2) Standardized Monitoring, i.e. using standardized tests for internal purposes without necessarily comparing to external standards, (3) Internal Testing, informing local stakeholders based on any measure of student achievement, and (4) Internal Teacher Monitoring, using any kind of evaluation mechanism to judge teachers. The authors find that only the first type of assessment is associated with improvements in student achievement. This finding is supported by a sophisticated set of models and robustness checks. However, the following interpretation may raise some skepticism: "Internal testing that simply informs or monitors progress without external comparability and internal teacher monitoring including inspectorates have little discernible effect on overall performance" (Bergbauer et al., 2018, p. 2). The problem is that assessment policies and practices are measured on the country level only. Thus, for each of the 59 countries, there are 4 indicators measured up to 6 times (in 2000, 2003, 2006, 2009, 2012, 2015). These indicators describe how strongly different kinds of assessments are implemented on average across a country, but there is no attempt to measure policies and practices at the school- or classroom level. This kind of analysis is appropriate when comparing national evaluation and accountability policies such as centralized exams between countries and studying their impact over time (as Sects 7.3.4, 7.3.5 and 7.3.7 below will do), but it is inappropriate for studying the impact of internal evaluation and classroom assessment on student learning in local contexts. To answer the latter research question, longitudinal and (quasi-)experimental enhancements of PISA would be needed. Once again, the limitations of ILSA designs need to be taken seriously (as discussed in Sect. 7.1).

7.3 A Comparative Analysis of Assessment Policies and Practices, Implemented in PISA 2015

Starting with the first wave in 2000, school questionnaires in all cycles of OECD's Programme for International Student Assessment have addressed policies of evaluation and assessment, and how results were used within countries. 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 and how this has impacted system-level change in student achievement (see Sect. 7.3.7 below). In PISA 2015, the author, in close collaboration with staff at the German Institute for International Educational Research (DIPF) and the International Questionnaire Expert Group, developed a broader set of questions covering details of school evaluation and classroom assessment (Bayer, Klieme & Jude, 2016). These new measures will be presented in Sect. 7.3.1. Using PISA 2015 Main Study data from 55 participating countries (Sect. 7.3.2), empirical analyses will present findings on formative assessment practices in classrooms (Sect. 7.3.3), assessment-related school policies (Sect. 7.3.4), national contexts for assessment and accountability (Sect. 7.3.5) and the

relationships between them (Sect. 7.3.6). In the course of the section, methodological issues related to comparability across countries and interpreting cross-sectional findings will be discussed.

7.3.1 Developing Measures for PISA 2015

The PISA 2015 Context Questionnaires (Kuger et al., 2017; OECD, 2013) allow for studying how often students are assessed through highly standardized tests, teacher-made tests or teachers' judgmental rating (Question SC034 on "General assessment practice", see Table 7.1) and whether certain measures for school improvement, including internal and external evaluations of schools are common practice (SC037). Moreover, the impetus for action is also relevant in order to analyze system policies. Thus, the PISA 2015 questions referring to school improvement policies (SC037) and standardized testing (SC034) distinguish action that is mandatory, i.e. required by educational policies, from action that is based on the school's initiative.

To support the description and analysis of data use by schools, a set of items from previous PISA cycles (2000–2012) was taken up addressing various kinds of usage for student test scores, 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 the use of teacher-developed tests. In addition, three items asking whether schools publish test results, track them over time and/or provide scores to parents have also been taken up from previous cycles of PISA (Use of achievement data for accountability; SC036).

In order to understand the link between assessment, evaluation, and school development, fine-grained information on processes of external evaluation (e.g., Have the data been used to inform specific action for improvement of teaching? Were such measures put into practice promptly?; SC041) and consequences of internal evaluation (Which areas have been affected by change in school policies:

Table 7.1 Assessment-related questions in the PISA 2015 school (SC) and student (ST) questionnaires

Policies for assessment, evaluation and accountability (individual indicators and indices)	
General assessment practice	SC034
Purpose of assessment results	SC035
Use of achievement data for accountability	SC036
Measures for school improvement, including internal and external evaluation	SC037
Consequences of internal evaluation	SC040
Processes of external evaluation	SC041
Formative assessment and feedback (latent scale for classroom practice)	
Perceived feedback	ST104

curriculum, professional development, parental engagement, etc.?, SC040) has been added.

It should be noted that two questions (namely, SC034 and SC035) are referring to student assessment at the "national modal grade for 15 year old students", i.e. the grade which enrolls most of the PISA target population nationwide, while all other questions refer to school assessment policies in general.

All topics mentioned so far have been addressed within the PISA 2015 School Questionnaire, which was meant to be answered by the school's principal. Each item in this part of the School Questionnaire asked whether (all questions), why (SC037 and SC040 only) and how frequently (SC034 only) specific activities had been implemented by the school. Answers were treated as individual indicators, and sometimes indices were computed summing up across several activities. For example, three indices based on SC035 are summing up across different kinds of testing purposes, each of which is presented with a forced-choice ("yes"–"no") response format:

- "Formative use": This index counts how many of the following had been identified as purposes of using standardized tests in school: guiding student learning, adapting teaching to students 'needs, identifying aspects of instruction or curriculum that should be improved, informing parents.
- "Summative use on the student level" counts how many of the following purposes of standardized testing have been identified by the school principal: decision about retention, certification, or grouping students.
- "Use for school evaluation" counts positive answers to the following items: compare test results with other schools, compare to national or district performance, monitor school progress.

We do not expect these three indices to represent any "latent" construct. Rather than measuring dispositional concepts of school evaluation, the indices pragmatically summarize school policies which analytically fall within the three categories. In methodological terms, these indices are treated as "formative" measures, i.e. indices are defined by the items covered, rather than items "reflecting" a latent construct (for the distinction, see Ellwart & Konradt, 2011).

Arguably the most prominent form of assessment studied so far in educational research is formative assessment (see Sect. 6.2.2 above). Since feedback is essential in formative assessment, we assessed this concept in the PISA 2015 student questionnaire (Perceived Feedback; ST 104; see Table 7.2). This scale was developed to assess the frequency of (formative) feedback activities as perceived by each individual student, asking how often the teacher would inform the student about his or her performance (Item 1), identify strengths (Item2), tell the students where (Item 3) and how (Item 4) he or she can improve, and provide advice on how to reach the learning goals (Item 5). Students were asked to respond in reference to one science course they had chosen before. Contrary to individual indicators included in the School Questionnaire, these five items are supposed to reflect an underlying latent dimension of classroom practice, i.e. a view on teaching which is understood in a similar way by students within a given class, school or even system. OECD (2017a,

Table 7.2 Items on "perceived feedback" in the PISA 2015 student questionnaire. (OECD 2017a, p. 315)

Item number	Item	Corrected Item-Total-Correlation (Median across 55 countries)
ST104Q01NA	The teacher tells me how I am performing in this course.	.70
ST104Q02NA	The teacher gives me feedback on my strengths in this <school science> subject.	.78
ST104Q03NA	The teacher tells me in which areas I can still improve.	.81
ST104Q04NA	The teacher tells me how I can improve my performance.	.81
ST104Q05NA	The teacher advises me on how to reach my learning goals.	.79

p. 315) provides scale scores estimated within an Item Response Theory (IRT)-based approach; the respective variable is named "PERFEED" in the PISA 2015 data file.

7.3.2 Data and Methods

As described in the Technical Report (OECD, 2017a), PISA 2015 sampled 546,299 students from 18,817 schools representing 26.9 Mio 15 year old students in 74 countries (35 OECD members plus 39 "partner" countries). The study implemented a two-stage stratified strategy, sampling about 150 schools per country and about 30 students per school. Within schools, students of the target age (15 years) were selected across grade levels and classrooms. Thus, contrary to PIRLS and TIMSS, PISA does not allow for an analysis of classroom-level variation in teaching and learning.

Students worked on cognitive tests for about 2 h, followed by a Student Questionnaire which took about 35 min. Both tests and questionnaires were administered on computer in the vast majority of countries. In addition to students, school principals and – in 19 countries – also teachers were asked to fill in web-based questionnaires. The technology-based administration allows for routing procedures to guide the individual respondent through the questionnaire. E.g., school principals are asked about the purpose of using standardized tests (SC035) if and only if they have said they implement such tests in their school (SC034). This procedure should help avoid invalid responses, but as a side effect, a significant part of the data matrix will be "missing by design". E.g., across 55 countries, data on testing purposes are available from two thirds of the schools only.

The full international data file may be downloaded from the OECD website. The SPSS student-level data file contains about 1000 variables, including 130 measures of student achievement (13 separate domains or sub-domains, each represented by 10 "Plausible Values" to cope with the matrix-design used for testing) and about

160 technical variables (e.g., information on sampling and weights). More than 700 variables are based on the student questionnaire, including several optional add-ons. The School file consists of about 300 variables based on the School Questionnaire.

OECD reports routinely cover all participating countries and "systems". Nevertheless, experts are well aware that data quality varies between countries. For example, the feasibility and appropriateness of the study design for developing countries may be questioned. As a consequence, some researchers only use data from OECD countries. However, for cross-cultural studies, this results in a severe loss of cultural and systemic variation. As a compromise, the analyses reported in this chapter have been run for 55 selected countries, which may (a priori) be grouped into ten categories based on geographical and/or linguistic proximity:

- English speaking: AUS, CAN, IRL, NZL, UK, USA
- German/Dutch speaking: AUT, BEL, CHE, DEU, LUX, NLD
- Roman Europe: ESP, FRA, ITA, PRT
- Nordic States: DNK, FIN, ISL, NOR, SWE
- Baltic States: EST, LTU, LVA
- Central Europe: BGR, CZE, HUN, HRV, POL, ROU, SVK, SVN
- Eastern Europe: GEO, KAZ, MDA, RUS
- Eastern Mediterranean: CYP, GRC, ISR, MLT, TUR
- East Asia: HKG, JPN, KOR, MAC, SGP, TAP and China (representing several industrialized regions from the Eastern part of China, including Beijing and Shanghai),
- South America: ARG, BRA, CHL, COL, MEX, PER, URY

Thus, we include 14,111 schools, i.e. 75% of all schools participating in PISA 2015. Cases were weighted using the so-called "Senate weight" which standardizes all national or system samples to an equal size. Descriptive analyses, including exploration of relationships with student outcomes and other kinds of linear relationships, were run in parallel for all 55 countries using SPSS 22. As we do not use any significance testing, findings in Sects. 7.3.3 through 7.3.5 are not distorted by the clustered sampling. In Sect. 7.3.3.1, we report results from Multi-group Confirmatory Factors Analyses including multiple countries, executed in MPLUS. In Sect. 7.3.6.2 we do report significance testing, as we study policies, practices, and mean achievement on the school level only.

7.3.3 Formative Assessment and Feedback: Studying Teaching Practice from a Comparative Point of View

For our measure of "perceived feedback", students responded on a four-point Likert scale with the categories "never or almost never" (1), "some lessons" (2), "many lessons" (3), and "every lesson or almost every lesson" (4). Table 7.2 shows the item

wording and provides information on item-total correlation. Across all items and all countries, the minimum value for any item-total-correlation was .36. In all 55 countries, these items form a highly coherent scale, with Cronbach's alpha mostly above .80 (Median: .91); the single outlier is Romania with alpha = .74.

The five items cover different, yet related facets of a classroom practice which is co-constructed by students and teachers, based on shared norms and expectations and coherent chains of teacher and student activities. As any kind of social practice (Reckwitz, 2002), and classroom practice specifically, we assume the practice of formative assessment and feedback to be socially constructed within a culturally shaped social, physical, and intellectual space. Integrating this theoretical view (which is rooted in sociological theory and usually studied by qualitative methods) into the quantitative measurement approach, "Perceived Feedback" may be modeled as a latent variable, assuming the latent structure (i.e. dimensionality, factor loadings/discrimination and intercepts/difficulty) to be at least partially culture-specific, and expecting some agreement among students within the same institutional context (classroom, and to some extent school).

The assumption of perceptions being shared within the same school is in fact supported by the decomposition of variance: Within countries, between 6 and 19% of the variance in Perceived Feedback is between-school variance (Median across all 55 countries = 9.8%). This supports the claim that Formative Assessment and Feedback is a social practice shaped by the learning environment, and perceived in a somewhat similar way by students sharing the same environment. We would expect even higher levels of agreement on the classroom level, but unfortunately the PISA data set does not allow to identify that level. The structural assumptions are tested in the following excursus, summarized in Table 7.3.

7.3.3.1 Excursus on Cross-Cultural Measurement

Applying common conventions for acceptable model fit in structural equation modeling (CFI > .90; RMSEA <.08) to the first row in Table 7.3, we conclude that configural invariance holds across countries. I.e., within each of the 55 countries all five items can be assumed to represent a unidimensional latent construct. Based on criteria suggested by Rutkowski and Svetina (2014) for large international data sets, the loss in model fit when assuming metric invariance on top is negligible (ΔCFI < .02, ΔRMSEA < .03). Thus, factor loadings can be assumed to be equal

Table 7.3 Multiple group confirmatory factor analyses for perceived feedback

	Model assuming configural invariance		Model assuming metric invariance		Model assuming scalar invariance	
Countries	RMSEA	CFI	RMSEA	CFI	RMSEA	CFI
55 (10 regions)	.061	.981	.059	.969	.075	.928
6 English speaking	.038	.993	.038	.988	.045	.976
UK, IRL, AUS, NZL	.038	.993	.034	.991	.034	.988

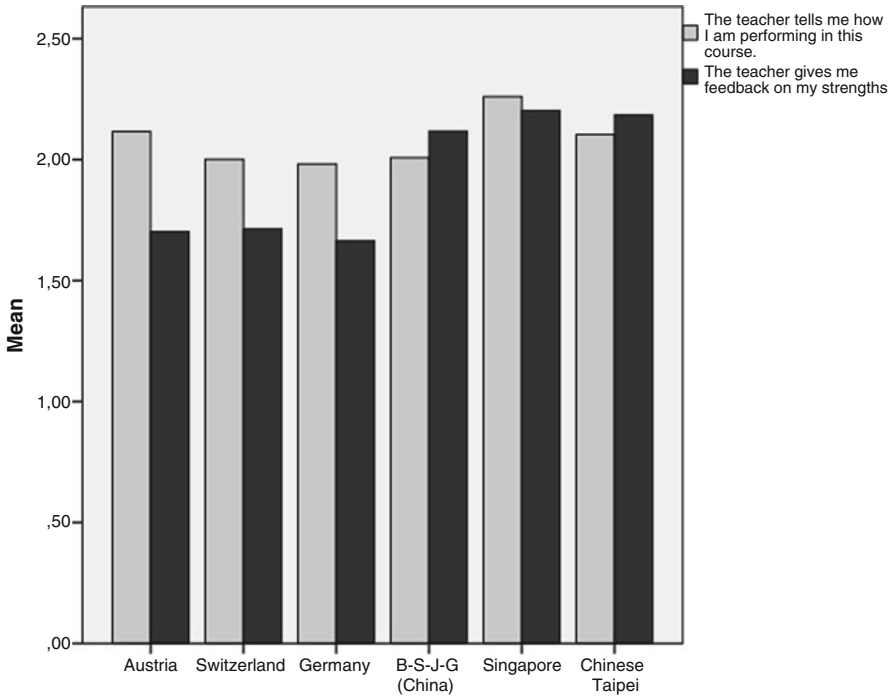

Fig. 7.1 Mean response regarding two feedback items for German-speaking and Chinese-speaking countries

across all 55 countries, meaning that items have similar discriminative power. However, when comparing the model of metric invariance with a third model assuming scalar invariance, the loss in model fit is substantial (neither ΔCFI nor ΔRMSEA are <.01, as requested by Rutkowski & Svetina, 2014). This means that item difficulties vary across countries, an important finding from a substantive point of view.

Figure 7.1 provides an empirical illustration of scalar non-invariance, based on data for items 1 and 2 from three German speaking and three Chinese-speaking countries. First, it should be noted that students report these activities to happen rather seldom; the average score is close to 2 ("some lessons") in all countries and for both items. In the three German-speaking countries participating in PISA, giving feedback on students' strengths is slightly more "difficult" (i.e. mean perceived frequency is lower) than providing feedback on student performance with respect to the course. In the three Chinese-speaking countries shown in this graph, however, the relative "difficulty" of these two items is nearly reversed. What does that mean in terms of "feedback cultures"? Within a Chinese context, teachers relatively often provide feedback based on an individual frame of reference (i.e. interpreting student achievement in relation to individual strengths), while within German culture, teachers seem to prefer a social frame of reference (i.e. interpreting student

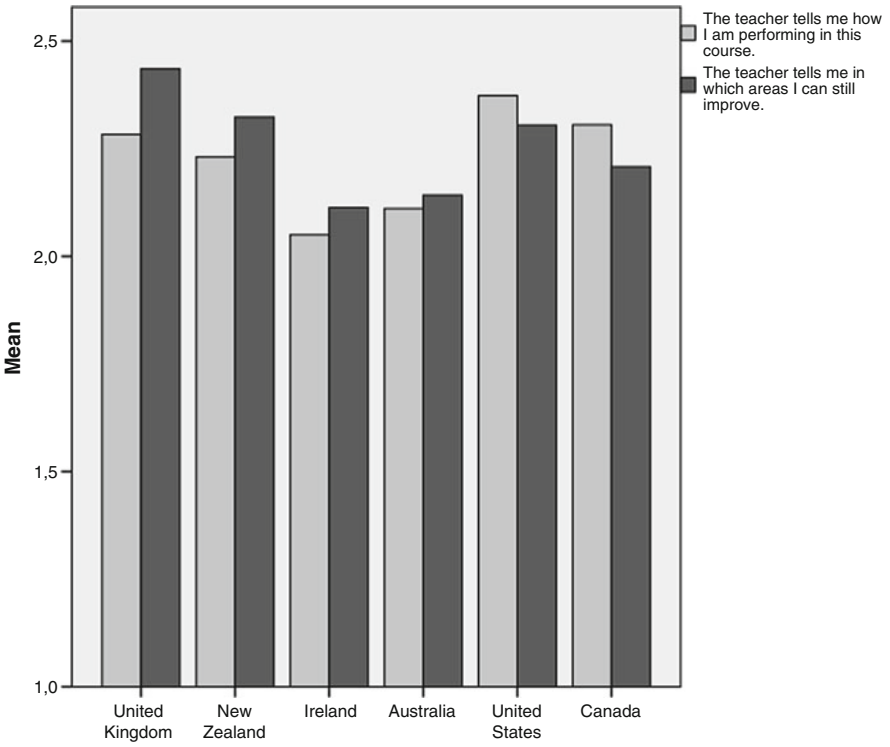

Fig. 7.2 Mean response regarding two feedback items for English-speaking countries

achievement in relation to the class). In statistical terms, the difference in relative "difficulty" implies that these items cannot be assumed to have equal intercepts on a common scale of "Perceived Feedback". Thus, scalar invariance has to be rejected.

The example shown in Fig. 7.1 suggests that linguistic and/or regional proximity might allow for higher levels of comparability. (See Fischer, Klieme & Praetorius, submitted, for similar findings regarding student-reported teaching quality.) This hypothesis is tested in the second row of Table 7.3, checking invariance across all six English-speaking countries included in our analysis. Again, metric invariance holds, but scalar invariance doesn't. Figure 7.2 illustrates, why. Once again, countries differ with respect to the prevalence of applying social comparison within the class vs. an individual frame of reference. Teachers in North America (Canada as well as the US) relatively often provide performance feedback in comparison to the class or course, while in other English-speaking countries, especially in the UK; teachers seem to prefer informing students about areas for individual improvement. As a result, scalar invariance does not hold across all six countries. If however, North America is excluded, in the remaining four countries even scalar invariance holds, as shown in row 3 of Table 7.3.

What does this analysis of measurement invariance imply for ILSA reporting? Still, common practice in policy reports published both by IEA and OECD includes ranking countries on scale means for all kinds of constructs, including measures of classroom practice such as "Perceived Feedback". In volume II of its policy report on PISA 2015, OECD (2017b, p. 67) provides a ranking of mean perceived feedback across countries, with the lowest value for Iceland and high values for the Dominican Republic and (among OECD countries) Mexico. All English-speaking countries are positioned above the OECD-mean, while all German-speaking countries are positioned below the OECD-mean. Based on our analyses of (non-) invariance, this kind of "finding" is not defendable. Rather than providing meaningful and useful information, such OECD ranking produces misleading myths on country-differences in teaching practice, in this case: on the use of formative assessment and feedback.

Strictly speaking, the IRT-based scaling for that set of items provided by OECD is inappropriate. Until PISA 2012, the questionnaire scaling procedures assumed discrimination and difficulty to be the same across all countries, and this assumption had not been tested at all – at least in routine analysis. Following a suggestion from Glas and Jehangir (2014), PISA 2015, for the first time ever, introduced a more sophisticated model, the Generalized Partial Credit Model, checked country- and item-specific misfit as documented in the Technical Report (OECD, 2017a, p. 290), and allowed for some variation in country-specific item parameters.[4] However, the analysis was based on a proprietary software owned by Educational Testing Service, with no prior application to cross-national questionnaire data. Thus, conventions for judging item (mis)fit had to be borrowed from cognitive tests or developed from the scratch. This led to all five items from the "Perceived Feedback" (PERFEED) scale being unflagged; therefore, the IRT scaling was done with common parameters across countries. Using the traditional approach of Multi-Group CFA, the present analysis calls for a revision of those conventions. In general, the methodology of establishing and testing measurement invariance for questionnaire scales in ILSAs is still evolving (van de Vijver, 2018). Organizations such as the OECD and IEA need to be much more hesitant when reporting and statistically comparing country means based on such data. Even within a country, questionnaire scales may lack invariance across different types or tracks of schools (Bayer, 2019), so the caveat applies to EER as well.

7.3.3.2 Restricting Comparison of Scale Means to a Smaller Sample of Countries

Applying rigorous standards of cross-cultural research to the construct of "Perceived Feedback", the discussion summarized in Table 7.3 leads to the decision to restrict

[4]This chapter of the Technical Report was authored by Janine Buchholz from DIPF, who kindly shared findings on the PERFEED scale with the present author. For a review of the scaling method, see Buchholz & Hartig, 2017.

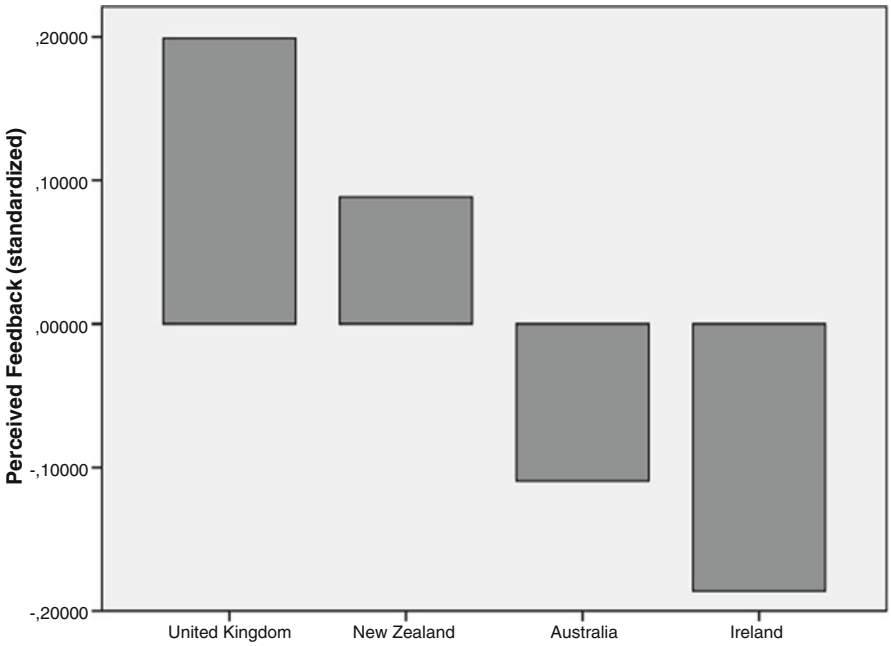

Fig. 7.3 Mean scale score for perceived feedback, standardized across four countries

comparative reports on scale means to four English-speaking countries with proven measurement invariance. Within this group of countries, students from the UK most often report receiving formative feedback (see Fig. 7.3). The difference with New Zealand, Australia, and Ireland roughly amounts to .1, .3, and .4 standard deviations, respectively. This information is trustworthy, and it can be relevant for teacher trainers and policy makers when discussing policies and practices related to formative assessment and feedback. E.g., professionals in the UK may conclude that in their system, promoting *more* use of assessment and feedback in classrooms is less of an issue than addressing the *quality* of assessment. Australian professionals may learn that the strong focus on assessment and feedback established in neighbouring New Zealand has not been implemented in their country (yet?).

7.3.4 The Purpose of Student Testing: Assessment as a School Policy

PISA 2015 provided school principals with an extended list of potential purposes of student testing, asking them to identify those that were relevant for their individual school when using standardized tests. As described in Sect. 7.3.1, three indices were created:

- formative use (e.g., guiding student learning and improving the curriculum),
- summative use on the student level (e.g., certification),
- use for school evaluation (e.g., comparing to national standards).

These indices represent different kinds of school policy towards student testing. Nevertheless, for all three of them, 23% of the overall variance can be identified as between-country-variance, which is quite a large share of the variation in assessment purposes. This proves that preference for any of these purposes is to a large extent driven by national (or system-level) policies, norms, and/or practices. Even the geographical/linguistic grouping of country does have some explanatory power: 10, 11, and 7% of variation in index 1, 2 and 3, respectively can be explained by differentiating between the ten "regions" described above. All in all, English speaking countries, especially the United Kingdom, New Zealand and the US, tend to rank at the top of this "hierarchy", while German-speaking countries are positioned at the bottom (see Fig. 7.4). This shows that traditional differences in how to use (or not to use) standardized student testing still prevail. On the other hand, it is instructive to see variation between countries belonging to the same geographical/linguistic group. E.g., Australia does not fit into the common pattern for English-speaking countries. Within the Nordic group of countries, strong differences can be found with regard to school evaluation purposes; these are rather weak in Finland and Norway, a little stronger in Denmark, and most important in Sweden and Iceland. At the same time, even in Sweden and Iceland, summative use of tests for judging individual students is quite low from an international point of view.

This is just descriptive data, but it may help a lot with public debates on testing. For example, claims that schools suffer from "Testeritis", meaning a move towards extensive testing practices, are quite popular among teacher unions in Germany nowadays,[5] but they can be challenged from a comparative point of view using these empirical data, as Germany has the lowest reported use.

7.3.5 *National Contexts for School Evaluation and Accountability*

While the previous section informed on school-based policies, which nevertheless were shown to be partly driven by national patterns, the following section looks into the wider context, mainly driven by accountability rules established by national, state, or district administration. Two questions in the PISA School Questionnaire which have been used in several cycles may help identify that context:

[5] https://www.focus.de/politik/deutschland/bildung-lehrer-machen-gegen-testeritis-an-schulen-front_id_3819831.html

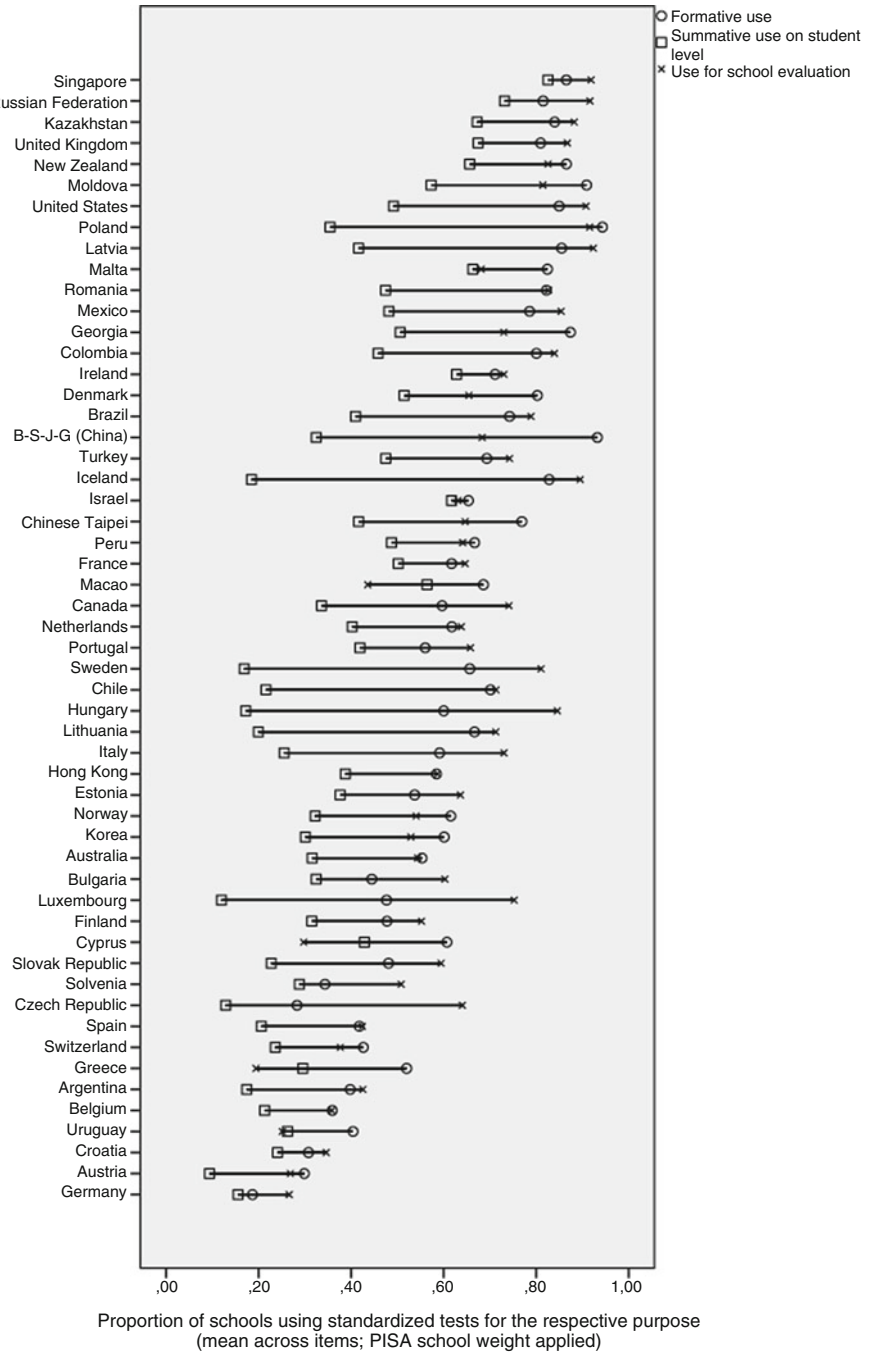

Fig. 7.4 Mean indices of using standardized tests for different purposes across countries. (Data on this question are not available for Japan)

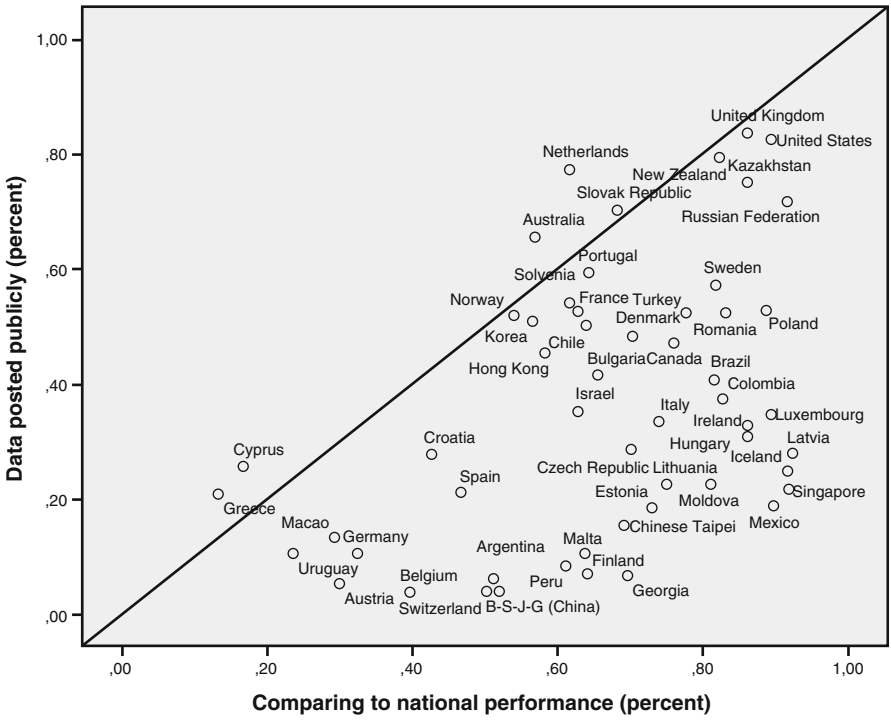

Fig. 7.5 Percentage of school principals reporting evaluation and accountability activities by country

- The question on Use of achievement data for accountability (SC036) asks, among other aspects of accountability, whether achievement data are posted publicly (e.g., in the media) by the school. The response format is "Yes/No".
- One item from the question on Purpose of standardized testing (SC035) asks whether standardized tests are used to compare the school to national, state or district performance. Again, the response format is "Yes/No".

With regard to the mechanisms explaining the impact of school evaluation discussed in Sect. 7.2.1, the first item is indicative of strong accountability practices, while the second item indicates a feedback-based approach to school improvement. Country-level means, as shown in Fig. 7.5, are in line with the patterns reported in the previous section.

- All in all, English speaking countries, especially the United Kingdom, New Zealand and the US, are international leaders in accountability, while German-speaking countries are positioned at the bottom.
- Once again, there are notable differences within country groups. E.g., Ireland and, to some extent (meaning: in some provinces), Canada tend to publish school results less often than other English-speaking countries. Comparing Norway,

Korea and Chile with Iceland, Singapore, and Mexico, respectively, we find complementary patterns within the Nordic, East Asian, and South American region.

- However, this analysis conveys two "global" messages that speak to the theories of school evaluation and accountability (see Sect. 7.2.1). These messages are easily visible by noting that the space below the diagonal is filled with all kinds of combinations of the two variables, while the space above the diagonal is practically void, with the exception of the Netherlands. (1) In practically all countries and systems, comparing test results to national or regional performance is reported more often than making results public. (2) High percentage of comparing is a prerequisite for high percentage of public reporting, but countries may show high prevalence for comparing while differing strongly in their prevalence for publication (see, e.g., Mexico vs. USA). Thus, there seems to be a hierarchy among the two mechanisms of school evaluation mentioned in Sect. 7.2.1. At a lower level, which might be called "soft accountability", schools are expected to compare their own test results to some external standards, probably as a kind of feedback mechanism.[6] At a higher level ("strong accountability"), data are made public to external users, such as parents, establishing a kind of quasi-market for schooling if it is used for the 'choice' of schools. These two kinds, or levels of accountability should not be mixed up.

7.3.6 *Integrating the Picture: How Formative Assessment Practice Relates to Student Composition, Evaluation and Accountability Policies on the School Level*

In this section, we take a closer look at formative assessment and feedback as perceived by students (see Sect. 7.3.3), using the IRT-based estimate of the latent construct named PERFEED in the PISA 2015 database. As PISA does not identify classrooms, analyses will be run on the school level, aggregating data from all students sampled within that school. Thus, we are dealing with formative assessment and feedback as a social practice of science education, as established in a given school, measured through the shared (mean) perception of 15 year-old students. Our research goal is to understand the relationship between this practice on one hand, the school's achievement level in science (represented by the school average of the first plausible value for science literacy), student composition (mean socio-economic status, as measured by the HISEI index of occupational status in PISA) and policies related to assessment, evaluation, or accountability (as measured by the indices discussed in Sects. 7.3.4 and 7.3.5) on the other hand.

[6]Bergbauer, Hanushek & Wößmann (Bergbauer et al., 2018, p. 17) classified this item as an instance of "school-based external comparisons".

7.3.6.1 Excursus on the Methodology of ILSA

Since the core variable used in this study, Perceived Feedback, does have metric, but not scalar invariance across countries, pooling data from all countries is not an option. The approach mostly used by econometricians (e.g. Bergbauer, Hanushek & Wößmann, 2018), introducing a fixed effect for every single country, helps to control for country-specific impact in regression-type models, but it does not take the lack of measurement invariance into account; it is bluntly ignorant with regard to measurement issues. Policy reports from IEA or OECD usually follow a third approach, running analyses in parallel for every country and summarizing findings in a qualitative way, as we did in Sects. 7.3.4 and 7.3.5 when interpreting descriptive findings across countries. However, this approach has severe drawbacks when applied to explanatory rather than descriptive findings. First, sample sizes may be quite small within individual countries. E.g., for studying purposes of student testing (SC035) on the school level, only 149 units would be available in Ireland, and 131 in New Zealand. Thus, the very asset of ILSAs to build a large international database would be lost. Second, summarizing and reporting results from complex models run in more than 70 countries is more an art than a science. Quite often, countries are grouped based on the results of within-country statistical tests, without any conceptual foundation.

When pooling student- or school level data for any analysis across countries, the appropriate approach seems to be the following: checking measurement invariance through rigorous methods (such as Multi-Group CFA, as applied in Sect. 7.3.3) and combining data sets from countries if and only if they meet the criteria of scalar invariance. Thus, in our case, we will work with the combined data from Australia, Ireland, New Zealand and the UK in the following.

We assume measurement invariance for the PISA science literacy test as documented in the Technical Report (OECD, 2017a). The standard measure of socio-economic status in PISA, the ESCS, cannot be assumed to be fully invariant across countries because item parameters from one of its components, the IRT measure of home possessions, have been shown to vary in meaning (Watermann, Maaz, Bayer, & Roczen, 2016). Therefore, we are using the international index of parental occupational status (HISEI) instead which is based on a transnational measurement approach in sociology (OECD, 2017a, p. 298).

For other predictive variables and control variables used in the following, we assume they do not represent latent variables. Rather, as discussed in Sect. 7.3.1, we treat them as "formative" indices, pragmatically summarizing reported activities of a certain kind (teaching and learning activities, professional activities at school, activities related to assessment, evaluation and accountability). Thus, there is no need to check measurement invariance for those variables.

7.3.6.2 Relating Perceived Feedback to Other School-Level Variables: In Search of the Proper *Explanandum*

In the light of conceptual debates on the relationship between ILSA and EER (see Sect. 7.1), it is interesting to note that OECD (2017b) treats Perceived Feedback both as a predictor for student achievement, *and* as an aspect of school practice that can be predicted from other variables.

The latter approach is visible in the following citation from Volume II of the PISA 2015 Policy Report (OECD, 2017b, p. 66): "Students in disadvantaged and rural schools were more likely to report that their teachers provide them with feedback . . . More perceived feedback is also associated with poorer performance in science, probably because low-performing students need and receive more feedback than better-performing students."

The former approach is at least implicitly used just one paragraph after the first citation: "Across OECD countries and after accounting for socio-economic status, students score between 5 and 17 points lower in science when they reported that their teachers use these strategies 'in many lessons' or 'every or almost every lesson' than when they reported that they use them in 'some lessons' or 'never or almost never'" (OECD 2017b, p. 66). Some pages further down (p. 73), the authors even talk about "impact on student performance" when, after controlling for reading and math achievement, they find a very low, but positive correlation between Perceived Feedback and student outcomes in Science Literacy.

Within our school-level analysis of these data, we illustrate both approaches in Table 7.4. Mean school-level achievement (PV1SCI) and mean school-level

Table 7.4 Prediction of mean student achievement vs. prediction of mean Perceived Feedback on the school level. (Data from AUS, IRL, NZL and UK; n = 1276 schools)

	Dependent variable	
Predictor[a]	Mean science achievement	Mean perceived feedback
Perceived feedback (PERFEED)	−.112 ***	
Science achievement (PV1SCIE)		−.181 ***
Socio-economic status (HISEI)	.654 ***	−.099 **
Goal-oriented curricular development	−.048 *	.061 *
Direct instruction (TDTEACH)	.131 ***	.307 ***
Inquiry-based teaching (IBTEACH)	−.037	.170 ***
Purpose of testing: Formative	−.001	.020
Purpose of testing: Summative/student level	.002	.127 ***
Purpose of testing: School level	.049[b]	.084 *
Data posted publicly	.022	.061 *
R^2	.516	.213

[a]Names in capital letters refer to aggregated student variables; all other predictors are based on the School Questionnaire. Parameters are standardized regression coefficients; [b]p < .10, * p < .05, ** p < .01, *** p < .001

Perceived Feedback (PERFEED) are used as *explanandum* (dependent variable) and predictor, respectively, in the left column, and vice versa in the right column. Additional predictors include the school's social composition (mean HISEI) and four indices of school policies in assessment, evaluation and accountability (see Sects. 7.3.4. and 7.3.5). Further, we include three control variables that cover major professional activities at the school level.

- One variable based on the school questionnaire (named "leadcom" in the PISA data base) summarizes the frequency of self-reported professional activities lead by the principal aimed at strengthening goal orientation in the curriculum. Items include "I ensure that teachers work according to the school's educational goals" and "I use student performance results to develop the school's educational goals".
- Two indices cover the frequency of different kinds of teacher activities in science classrooms as perceived by the students (school level average). (1) "Direct instruction" includes four core components of classroom teaching, namely "explaining ideas", "demonstrating an idea", "discussing student questions" and "conducting a whole class discussion". (2) "Inquiry-Based teaching", e.g., includes the following items: "Students are given opportunities to explain their ideas", "Students are allowed to design their own experiments".

Compared to the prediction of mean perceived feedback (right column), mean Achievement (left column) is obviously much easier to predict using this set of predictors; more than 50% of school-level variation can be explained. However, the prediction is mainly due to the relationship between achievement and socio-economic composition. Otherwise, Direct Instruction has a significant "effect" – showing that the achievement level is higher if schools succeed in implementing core activities of teaching across their science courses. However, as stated above, the direction of this relationship remains unclear: Probably, the more knowledgeable students are, the easier it is for teachers to enact core teaching activities. Perceived Feedback and goal-oriented curriculum development are associated with lower achievement – which may be interpreted as a case of reversed causality: both kinds of activities are probably implemented in response to low student outcomes.

Once again, we run into undecidable questions of directionality and causality when using cross-sectional ILSA data in attempts to "explain" the variation in achievement between schools. Regarding the topic of this chapter, assessment-related policies, however, the message is clear: Controlling for student composition and some basic kinds of professional activity, there is no relationship between any of the school policy indices and student achievement. The single index which is close to significance ($p < .10$) is *not*, as put forward by OECD in 2006, public posting of data. Rather, it is the use of data for school-level evaluation purposes such as comparing the school's test results with national standards. Referring to Sect. 7.3.5, we conclude that "soft accountability" tends to be related to a school's achievement level, while "strong accountability" is not.

The complementary research question, "Which schools implement formative feedback, under which conditions?" can be answered more clearly (see right column

in Table 7.4). Only 21% of the school-level variance can be explained, but there are more significant relationships with predictors:

- Schools that provide feedback with high frequency seem to be working 'against the odds', i.e. serving students with relatively low SES and low achievement.
- Also, students report higher intensity of formative assessment and feedback if they report higher levels of other teaching activities as well. Interestingly, the relationship with "Direct instruction" is much stronger than the relationship with "Inquiry-based teaching". In fact, it has been claimed in conceptualizations of teaching quality (e.g., Rosenshine & Stevens, 1986) that "Direct Instruction"--kind of approaches are successful, especially for groups of low achievers, because they provide lots of opportunities for feedback "on the fly" – contrary to inquiry-based learning where students spent more time working on their own.
- Most important, and not trivial at all, is the finding that several school-level policies and activities are related to students' shared perception of receiving feedback. There are relations with goal-oriented professional planning (remember one item referred to the use of performance results!) as well as with evaluation/accountability-related school policies. Both soft accountability policies (purpose of testing: school level) and strong accountability policies (data posted public) show small, but significant regression parameters. Even more relevant is the school's use of standardized testing for decisions about retention, grouping, and certification ("summative" use at the student level). The school's formative use of standardized tests in adaptive teaching and learning, curriculum development and parental involvement, however, seems not be specifically related to feedback approaches as perceived by students.

When choosing the use of formative assessment and feedback to be the *explanandum* (right column in Table 7.4), directionality is easier to decide. There is little sense in assuming that student perceptions of classroom practices in science courses determine school policies reported by principals. Hence, we may interpret any relationship *between* school policies and student-perceived practices as an effect of school evaluation policy *on* assessment and feedback practices embedded in science teaching. Our findings show that school policies do make a difference for everyday classroom practice experienced by students: The more tests are used for summative decision making, school evaluation and accountability, the more students report receiving formative feedback in classrooms. At least in the four English speaking countries covered here, formative and summative assessment seem to be connected rather than being mutually exclusive.

The findings reported in Table 7.4 can inform the development of new research questions. Here are some relatively general hypotheses to be tested in future studies:

1. Formative assessment and feedback is a classroom practice which is closely related to "traditional" (direct) teaching activities. It is still less integrated in "constructivist" teaching activities such as inquiry-based teaching. (Fischer, He and Klieme, submitted, seek to test this hypothesis across cultures.)

2. School policies on assessment, evaluation, and accountability can promote and foster classroom-based assessment if these policies are touching medium to high stakes, e.g. the use of test data in school evaluation or (even more important) the use of tests in decisions on student careers. (Above, we interpreted our data from four English speaking countries in line with this hypothesis. However, we do not know if the statement holds for other systems, especially when the overall level of assessment and evaluation is much lower than in English-speaking countries. Also, intervention studies should be implemented to test causal claims.)
3. (a) Policies regarding assessment, evaluation and accountability are mostly unrelated to student achievement at the school level, if student background and teaching practices are controlled for. (b) Effects on student outcomes may, if any, be expected from assessment data being used more often for (self-) evaluation on the school level. (Again, our cross-sectional findings from English-speaking countries need to be tested under different conditions. Statement (b) will be checked based on country-level trend data in the next section, and it is in line with findings reported by Bergbauer et al., 2018, using different methodology.)

7.3.7 Long-Term Changes in Assessment Strategies

In order to test Hypothesis 3(b) generated in the previous section, school policies on test use and school evaluation would have to be changed under controlled experimental or quasi-experimental conditions. Such studies are very hard to implement. However, as we have seen in Sect. 7.3.4, those school policies seem to be shaped by national (perhaps also state or district) contexts. Thus, instead of implementing (quasi-)experimental treatments, "natural" change in national policies may be studied in relation to changes in student achievement on the national level across different ILSA testing occasions.

ILSAs, in this case PISA, provide trend data that can allow this question to be studied. Bergbauer, Hanushek and Wößmann (Bergbauer et al., 2018, p. 16), mainly using PISA data, observed "a tendency for increased prevalence of the measures of standardized external comparison over time". Teltemann and Klieme (2016) focused on a single indicator which was available for almost all PISA cycles so far[7]: the item "In my school, assessments of 15-year old students[8] are used ... to compare the school to district, state or national[9] performance." In Sects. 7.3.4. and 7.3.5, we classified this item as indicating "use of assessment for school evaluation" and "soft accountability", in line with Bergbauer, Hanushek and Wößmann (Bergbauer et al., 2018, p. 17) who describe it as an instance of "school-based external comparisons". Unfortunately, in PISA 2015 this item was split up, one version addressing

[7]With the exception of PISA 2006.

[8]Later replaced by 'students in national modal grade for 15-year-olds'.

[9]The international school questionnaire allows for national adaptations regarding the level on which comparisons are made.

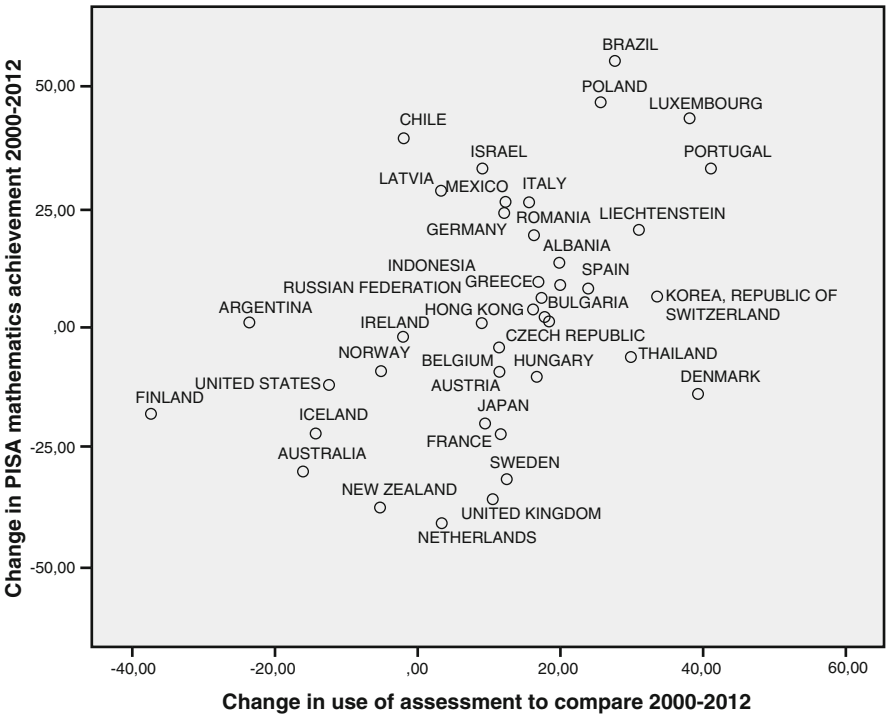

Fig. 7.6 Changes in use of assessment results to compare school performance with national, state, or district data (difference in percentage, as reported by school principals) and changes in mean mathematics achievement, PISA 2000–2012

"standardised tests", one version addressing "teacher-developed tests". Therefore, we restrict the discussion of system level change to the time interval from 2000 to 2012.[10]

As Teltemann and Klieme (2016) documented, from 2000 to 2012 the vast majority of OECD countries increased the use of assessment to compare with national or state/district performance. This global trend is also visible in Fig. 7.6. Furthermore, this figure shows that, on the country level, the change in use of "soft accountability" (horizontal axis) is related to the change in mean student achievement in mathematics (vertical axis). The correlation is $r = .449$ ($p < .01$), computed across all 35 out of the 55 countries (as selected in Sect. 7.3.2) for which both indices are available. Thus, "soft" accountability policy, as indicated by an increased proportion of schools comparing their assessment data to system-wide norms, seems to trigger gains in student achievement on the country level. PISA provides

[10]Changes in background questionnaire wording across cycles of measurement are yet another obstacle against analyzing trend data from ILSA's; cf. Singer et al., 2018, p. 64.

support for the hypothesis put forward at the end of Sect. 7.3.6: using assessment data for self-evaluation of schools has an impact on student outcomes.

How can this directional, even causal claim be justified? Gustafsson (2007) as well as Strietholt, Bos, Gustafsson, and Rosén (2014) provide several examples of such a "longitudinal cross-cohort design". By analyzing change on the country level, this method controls for fixed characteristics of the countries, thus reducing the effects of unobserved (omitted) variables. The method has been criticized by Singer et al. (2018, p. 59, pp. 65–66) mainly because it neglects the multi-level structure of the data, interpreting country level effects (e.g., the impact of mean use of computers at home on mean achievement) on a lower level (e.g., the impact of individual computer use on students' achievement). Similar criticism has been raised above, in Sect. 7.2.3, against the analysis of internal (classroom) testing by Bergbauer et al. (2018). However, accountability policies such as encouraging schools to compare their own performance to national norms are largely decided and implemented at the system level (see Sect. 7.3.4). Thus, when interpreting the finding from Fig. 7.6, we may indeed draw conclusions on the country (system) level.

So far, surprisingly few studies report quantitative analyses of system-level change. (See, however, Aloisi & Tymms, 2017; Lenkeit & Caro, 2014; Strietholt et al., 2014). The more waves of data collection PISA, PIRLS and TIMSS have implemented, the better researchers are prepared to model longitudinal data on the country (system) level. In order to strengthen such research, the meaning of aggregated measures has to be better understood (for examples, see Klieme, 2016; Rozman & Klieme, 2017), new research methods such as Bayesian modeling (Kaplan & Lee, 2018) should be adapted, and, most importantly, testable theories of educational change on the system level (e.g., Sun, Creemers, & de Jong, 2007) need to be further developed.

7.4 Summary and Discussion: Connecting ILSA and EER

The use and misuse of data from International Large Scale Assessments (ILSAs) has been repeatedly discussed both in public and by scholars in educational research (e.g., Rowan, 2002; Singer et al., 2018). Much of this discussion is dealing with how ILSAs may respond to the needs and practical questions of policy makers, professionals, and other stakeholders in education while avoiding overstatements, overgeneralizations or simplifications. While such pitfalls are common both in policy and in research (as shown in the introductory section of this chapter), experts agree that ILSAs should be conceptually based on Educational Effectiveness Research (EER) and adhere to rigorous methodological principles. The methodological foundations and challenges of ILSAs may be grouped into three major areas:

- Design: How to deal with the cross-sectional nature of individual ILSAs; how to use the trend design on the country level; how to specify the proper explanandum (explaining learning opportunities or school process quality rather than searching

for an explanation of student outcomes which is done almost "automatically" by most analysts); being cautious with regard to causality and the direction of effects;
- Sampling and data structure: how to deal with the multi-level nature of data, with clustered sampling, and with missing values (including missing by design): how to avoid ecological fallacies when interpreting country level relationships;
- Measurement: how to discriminate formative vs. reflective constructs, or manifest indices vs. latent (dispositional) measures; how to assess item and scale (mis)fit and establish measurement invariance across cultures.

Some of these issues were discussed and empirically illustrated across the chapter, but most of them could only be touched briefly. For example, issues of multi-level model specification were beyond the limits of the present chapter.

Both the fragile relationship with policy and practice as well as most of the methodological problems are shared features of ILSAs and EER. Therefore, each of the two paradigms of school research can mutually benefit from solutions developed by the other. Even more important seems to be the connection with regard to theoretical foundations and empirical findings. The present chapter explored such connections within one particular area of research and one particular study: the study of policies and practices of assessment and evaluation in PISA 2015.

First (in Sects. 6.2, 6.3.1 and 6.3.2), we showed how related constructs from EER have been taken up and implemented in PISA. Through PISA, national patterns of classroom assessment practices (Sect. 6.3.3), use of student assessment (Sect. 6.3.4), school evaluation and accountability policies (Sect. 6.3.5) have been identified. E.g., it turned out that English-speaking countries are similar in many respects, while full (scalar) invariance could be established for student reports from four countries only: UK, New Zealand, Australia, and Ireland can legitimately be ordered according to their respective prevalence of formative assessment and feedback. On the school level across those countries, formative assessment practices reported by students are related to summative use of assessment reported by principals. Thus, at least in some English-speaking countries, formative and summative assessment are positively connected rather than complementary. Regarding school accountability, the distinction between "soft accountability" (comparing performance with a national standard) and "strong accountability" (making test results public) proved to be informative. On the country level, soft accountability seems to be a necessary, but not sufficient prerequisite for strong accountability. For soft accountability only, a slightly positive relationship with student achievement on the school level (Sect. 6.3.6) and a positive impact on country-level math achievement (Sect. 6.3.7) were found.

From an EER perspective, we conclude that ILSA data help understand the effects of assessment, evaluation, and accountability on student outcomes. Experimental research has proven formative assessment and feedback to be an effective classroom practice (e.g., Kingston & Nash, 2011). Cross-sectional PISA data are not suitable for testing this claim. Nevertheless, they provide additional information on the variation of formative assessment practices within and between countries. In addition, PISA trend data show that summative assessments, such as national

surveys of student achievement, may trigger country-level growth in student outcomes as they provide feedback to schools. Overall, the pattern of our findings is consistent with a theory of school improvement based on "soft accountability", feedback and professional learning as the main mechanisms.

Policies and practices of assessment thus provide a showcase for how EER constructs can inform ILSA design, and how ILSA data in turn can inform the EER knowledge base.

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, NY/London, UK: Teacher College/Columbia University.

Aloisi, C., & Tymms, P. (2017). PISA trends, social changes, and education reforms. *Educational Research and Evaluation, 23*(5–6), 180–220.

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

Baker, D. P. (2009). The invisible hand of world education culture. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 958–968). New York, NY: Routledge.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180. https://doi.org/10.3102/0002831209345157

Bayer, S. (2019). *Alle alles ganz lehren – Aber wie? Mathematikunterricht vergleichend zwischen den Schularten* [Omnes omnia omnino doceantur – But how? Comparing mathematics teaching between school tracks]. Phil. Dissertation. Goethe University, Frankfurt am Main.

Bayer, S., Klieme, E., & Jude, N. (2016). Assessment and evaluation in educational contexts. In S. In Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning. An international perspective* (pp. 469–488). New York, NY: Springer.

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

Bergbauer, A. B., Hanushek, E. A., & Wößmann, L. (2018, July). *Testing* (CESifo working paper no. 7168 7168 2018).

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, IL: University of Chicago Press.

Bogdandy, A. V., & Goldmann, M. (2009). The exercise of international public authority through National Policy Assessment. The PISA study of the OECD as a template for a new international standard legal instrument. *Zeitschrift für ausländisches öffentliches Recht und Völkerrecht, 69*, 51–102.

Bottani, N., & Tuijnman, A. C. (1994). The design of indicator systems. In A. C. Tuijnman & T. N. Postlethwaithe (Eds.), *Monitoring the standards of education* (pp. 47–78). Oxford, UK: Pergamon.

Bryk, A., & Hermanson, K. (1994). Observations on the structure, interpretation and use of education indicator systems. In OECD (Ed.), *Making education count: Developing and using international indicators* (pp. 37–53). Paris, France: OECD.

Buchholz, J. & Hartig, J. (2017). Comparing attitudes across groups: An IRT-based item-fit statistic for the analysis of measurement invariance. *Applied Psychological Measurement*. Advance online publication. https://doi.org/10.1177/0146621617748323.

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

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

Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., et al. (2015). Embedded formative assessment and classroom process quality: How do they interact in promoting students' science understanding? *American Educational Research Journal, 52*(6), 1133–1159.

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, CA: 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, The Netherlands: Springer.

Ellwart, T., & Konradt, U. (2011). Formative versus reflective measurement: An illustration using work-family balance. *Journal of Psychology, 145*(5), 391–417.

Faubert, V. (2009). *School evaluation: Current practices in OECD countries and a literature review* (OECD Education working papers, no. 42). Paris, France: OECD.

Fischer, J., He, J., & Klieme, E.. (Submitted). *The structure of teaching practices across countries: A combination of factor analysis and network analysis*.

Fischer J., Klieme E., & Praetorius A-K.. (Submitted). *The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality*.

Glas, C. A. W., & Jehangir, K. (2014). Modeling country specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large scale assessment* (pp. 97–116). Boca Raton, FL: CRC Press.

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

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

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

He, J., Buchholz, J., & Klieme, E. (2017). Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures. *Journal of Cross-Cultural Psychology, 48*(3), 319–334.

He, J. & Kubacka, K. (2015). *Data comparability in the teaching and learning international survey (TALIS) 2008 and 2013* (OECD education working papers vol. 124). Paris, France: OECD.

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.

Jerrim, J. (2011). *"England's "plummeting" PISA test scores between 2000 and 2009: Is the performance of our secondary school pupils really in relative decline"* (DoQSS working papers 11–09), Department of Quantitative Social Science – UCL Institute of Education, University College London.

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, The Netherlands: Springer.

Jude, N., & Kuger, S. (2018). *Questionnaire development and design for international large-scale assessments (ILSAs)*. Washington, DC: National Academy of Education.

Kaplan, D. & Lee, C. (2018). Optimizing prediction using Bayesian model averaging: Examples using large-scale educational assessments. *Evaluation Review*. Advance online publication. https://doi.org/10.1177/0193841X18761421

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. (2012). The role of large-scale assessments in research on educational effectiveness and school development. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 115–147). Heidelberg, Germany: Springer.

Klieme, E. (2016, December). *TIMSS 2015 and PISA 2015 -How are they related on the country level?* (DIPF working paper). https://pisa.dipf.de/de/pdf-ordner/Klieme_TIMSS2015andPISA2015.pdf

Klieme, E. (2018, February). *Alles schräg* (Biased findings). https://www.zeit.de/2018/07/pisa-studie-oecd-politik-eckhard-klieme.

Klieme, E., Jude, N., Baumert, J., & Prenzel, M. (2010). PISA 2000–2009: Bilanz der Veränderungen im Schulsystem (Making up the balance of changes in the school system). In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Koeller, M. Prenzel, W. Schneider, & P. Stanat (Hrsg.), *PISA 2009. Bilanz nach einem Jahrzehnt (Making up the balance a decade after).* Münster, Germany: Waxmann.

Klieme, E., & Kuger, S. (2015). PISA 2015 context questionnaires framework. In *PISA 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy* (pp. 101–127). Paris, France: OECD.

Klieme, E., & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht (Teaching quality from a student perspective: Culture-specific profiles, regional differences, and relationships with teaching effects). In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann, (Hrsg.). *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland (Hrsg.)* (S. 334–359). Opladen, Germany: Leske + Budrich.

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

Kuger, S., Klieme, E., Lüdtke, O., Schiepe-Tiska, A., & Reiss, K. (2017). Mathematikunterricht von Schülerleistungen in der Sekundarstufe: Zur Validität von Schülerbefragungen in Schulleistungsstudien (Mathematics teaching and student achievement in secondary education: The validity of student surveys in school achievement studies). *Zeitschrift fuer Erziehungswissenschaft, 20*(2), 612. https://doi.org/10.1007/s11618-017-0750-6

Lenkeit, J., & Caro, D. H. (2014). Performance status and change – Measuring education system effectiveness with data from PISA 2000–2009. *Educational Research and Evaluation, 20*(2), 146–174.

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.

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

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

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

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

OECD. (2014). *PISA 2012 technical report*. Paris, France: OECD.

OECD. (2017a). *PISA 2015 technical report*. Paris, France: OECD.

OECD. (2017b). *PISA 2015 Results, Volume II. Policies and practices for successful schools*. Paris, France: OECD.

OECD & Vodafone Stiftung. (2018, January). *Erfolgsfaktor Resilienz* (Success factor resilience). https://www.vodafone-stiftung.de/uploads/tx_newsjson/Vodafone_Stiftung_Erfolgsfaktor_Resilienz_01_02.pdf

Rakoczy, K., Klieme, E., Leiss, D., & Blum, W. (2017). Formative assessment in mathematics instruction: Theoretical considerations and empirical results of the Co^{2}CA project. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education: Research, models and instruments* (pp. 447–467). Cham, Switzerland: Springer.

Reckwitz, A. (2002). Toward a theory of social practices: A development in culturalist theorizing. *European Journal of Social Theory, 5*(2), 243–263.

Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York, NY: Macmillan.

Rowan, B. (2002). Large-scale, cross-National Surveys of educational achievement: Promises, pitfalls, and possibilities. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-National Surveys of educational achievement* (pp. 319–350). Washington, DC: National Academic Press.

Rozman, M., & Klieme, E. (2017). *Exploring cross-national changes in instructional practices: Evidence from four cycles of TIMSS* (Policy brief vol. 13). Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31–57.

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, The Netherlands: Kluwer Academic Publishers.

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, The Netherlands/Oxford, UK: Elsevier Science.

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

Schmidt, W. H., Burroughs, N. A., Zoido, P., & Houang, R. T. (2015). The role of schooling in perpetuating educational inequality: An international perspective. *Educational Researcher, 44*(7), 371–386.

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

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

Singer, J., Braun, H., & Chudowsky, N. (Eds.). (2018). *International education assessments – Cautions, conundrums, and common sense*. Washington, DC: National Academy of Education.

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

Strietholt, R., Bos, W., Gustafsson, J.-E., & Rosén, M. (Eds.). (2014). *Educational policy evaluation through international comparative assessments*. Münster, Germany: Waxmann.

Sun, H., Creemers, B. P. M., & de Jong, R. (2007). Contextual factors and effective school improvement. *School Effectiveness and School Improvement, 18*(1), 93–122.

Teltemann, J., & Klieme, E. (2016). 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, NY: Routledge.

van de Vijver, F. & He, J., (2016), Bias assessment and prevention in non-cognitive outcome measures in PISA questionnaires. In Kuger, S., Klieme, E., Jude, N. & Kaplan, D. (eds.). Assessing contexts of learning world-wide: An international perspective. New York, NY: Springer Science, p. 229–253. 24 p.

van de Vijver, F. J. R. (2018). Towards an integrated framework of Bias in noncognitive assessment in international large-scale studies: Challenges and prospects. *Educational Measurement: Issues and Practices, 37*(4), 49–56. 8p.

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

Watermann, R., Maaz, K., Bayer, S., & Roczen, N. (2016). Social background. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: An international perspective* (Methodology of educational measurement and assessment) (pp. 117–145). Springer. https://doi.org/10.1007/978-3-319-45357-6

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

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