**Unclassified** **EDU/WKP(2011)9**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **02-Dec-2011**

**English - Or. English**

**DIRECTORATE FOR EDUCATION**

EDU/WKP(2011)9
**Unclassified**

**Cancels & replaces the same document of 05 October 2011**

# ALIGNMENT IN COMPLEX EDUCATION SYSTEMS: ACHIEVING BALANCE AND COHERENCE

**OECD Education Working Paper No. 64**

**by Janet W. Looney**

*This paper was commissioned to Janet Looney, an independent consultant specialising in programme design, evaluation and learning. The paper forms part of the work undertaken by the OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes and includes revisions in light of the discussion of an earlier version [EDU/EDPC/EA(2010)3] at the 2nd meeting of the Group of National Experts on Evaluation and Assessment (9-10 September 2010).*

*The OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes is designed to respond to the strong interest in evaluation and assessment issues evident at national and international levels. The overall purpose is to explore how systems of evaluation and assessment can be used to improve the quality, equity and efficiency of school education. The Review looks at the various components of assessment and evaluation frameworks that countries use with the objective of improving student outcomes. These include student assessment, teacher appraisal, school assessment and system evaluation. More information is available at www.oecd.org/edu/evaluationpolicy.*

Contact: Mr. Paulo Santiago [Tel: +33(0) 1 45 24 84 19; e-mail: paulo.santiago@oecd.org]
and Ms. Claire Shewbridge [Tel: +33(0) 1 45 24 99 63; e-mail: claire.shewbridge@oecd.org].

**JT03312703**



**English - Or. English**

**OECD DIRECTORATE FOR EDUCATION**

**OECD EDUCATION WORKING PAPERS SERIES**

This series is designed to make available to a wider readership selected studies drawing on the work of the OECD Directorate for Education. Authorship is usually collective, but principal writers are named. The papers are generally available only in their original language (English or French) with a short summary available in the other.

Comment on the series is welcome, and should be sent to either edu.contact@oecd.org or the Directorate for Education, 2 rue André Pascal, 75775 Paris CEDEX 16, France.



---

www.oecd.org/edu/workingpapers

---

## ABSTRACT

The majority of OECD countries now implement one form or another of standards-based assessment and evaluation. The core logic of standards-based systems rests upon the alignment of three key elements: *standards* defining the knowledge and skills – *or competences* – students are expected to have attained at different stages of their education; *curricula*, which cover the objectives identified in standards; and *student assessments and school evaluations* which measure attainment of standards. If systems are misaligned, it is impossible to draw valid conclusions about the success of student learning or to develop effective strategies for school improvement. Yet, no system can achieve perfect alignment. This report proposes that rather than thinking of alignment literally, as a lining up of the various elements and actors across systems, it may be more appropriate to approach it as a matter of balance and coherence. The discussion touches on both the technical and social dimensions of alignment.[1]



## RÉSUMÉ

La majorité des pays de l'OCDE met désormais en œuvre un système d'évaluation fondé sur des normes, quelle que soit la forme de ce système. La logique de base des systèmes d'évaluation fondés sur des normes repose sur l'alignement de trois éléments clés : des *normes* définissant les connaissances et les compétences que les élèves sont censés avoir acquis à différents stades de leur éducation; des *programmes* qui couvrent les objectifs identifiés dans les normes ; et des *évaluations des étudiants et des écoles*, qui mesurent le niveau des normes. Si les éléments clés de ces systèmes sont mal alignés, il est impossible de tirer des conclusions valables sur la réussite de l'apprentissage des élèves ou de développer des stratégies efficaces pour l'amélioration des écoles. Cependant, aucun système ne peut parvenir à un alignement parfait. Ce rapport propose qu'au lieu de penser l'alignement de manière littérale, à savoir une succession de divers éléments et d'acteurs au travers des systèmes, il serait plus approprié de l'aborder en termes d'équilibre et de cohérence. La discussion porte sur les dimensions techniques et sociales de l'alignement.


[1] ***Janet Looney***, an American national, is an independent consultant specialising in programme design, evaluation, and learning. Between 2002 and 2008, Ms. Looney was the project lead for the What Works in Innovation in Education programme at the OECD's Centre for Educational Research (CERI). She led the development of two major international synthesis reports: *Formative Assessment: Improving Learning in Secondary Classrooms* (2005), and *Teaching, Learning and Assessment for Adults: Improving Foundation Skills* (2008). Prior to her work with the OECD, Ms. Looney was Assistant Director of the Institute for Public Policy and Management at the University of Washington (1996-2002), where she was involved in evaluation of community development programmes, urban education reforms, and state-level implementation of federal welfare. Between 1994 and 1996, she was a Programme Examiner in the Education Branch of the U.S. Office of Management and Budget. She received her Master of Public Administration and Master of Arts in International Studies degrees from the University of Washington in 1993.

## TABLE OF CONTENTS



**Boxes**

## SECTION 1. THE KEY ROLE OF ALIGNMENT IN STANDARDS-BASED ASSESSMENT AND EVALUATION

1. The majority of OECD countries now implement one form or another of standards-based assessment and evaluation. In standards-based systems, governments set standards for student attainment, promoting both quality and equity of student outcomes. Standards define the knowledge and skills – or competences – students are expected to have attained at different stages of their education. The curriculum covers the objectives identified in standards, and student assessments and school evaluations focus on attainment of standards. Most standards-based systems also include some kind of incentives – for example, publication of assessment and evaluations results, rewards and or sanctions for schools – to motivate improvements in instruction.

2. The core logic of standards-based systems rests upon alignment of these key elements. If systems are misaligned, it is impossible to draw valid conclusions about the success of student learning or to develop effective strategies for school improvement. The cost – in terms of money, time and lost opportunities – is potentially enormous.

3. Yet, no system can achieve perfect alignment. A number of commentators have pointed to the complexity of systems as a barrier to tight alignment (Baker, 2004; Hargreaves, 2003; O'Day, 2002; Weick, 1976). School systems include multiple layers and links, operate in diverse contexts, and employ teachers and school leaders with a range of experiences and capabilities. Learning is in and of itself a complex process. In addition, a central and persistent concern of the research on standards-based assessment and evaluation is in regard to tensions between external and internal school accountability – and how this, in turn, affects the quality of information gathered and uses to which it is put. Given this complexity, it is very difficult to establish clear relationships across standards, curriculum, incentives and assessments and evaluations.

4. This report proposes that rather than thinking of alignment literally, as a lining up of the various elements and actors across systems, it may be more appropriate to approach it as a matter of balance and coherence. The discussion touches on both the technical and social dimensions of alignment, and addresses the following questions:

- How can systems most effectively balance goals for school accountability and for improvement? What are the most important features of standards, curriculum and assessment?

- What is the best mix of incentives to motivate and support change in schools and classrooms? Which actors and institutions are best placed to influence improvements in school management and instruction? Is it possible to achieve an appropriate balance between bureaucratic needs for accountability and a strong role for teachers as professionals?

- How might standards-based systems achieve overall coherence, ensuring that assessment and evaluation meet both policy makers' and practitioners' needs for information? A complementary question is how systems can avoid placing too much of an emphasis on coherence – so that there is room for innovation and support for the "softer" and less measureable goals of education, such as moral and ethical values?

5. The following sections explore these issues in more depth. The next section (Section 2) discusses countries' motivations for developing standards-based systems, and provides a broad overview of different approaches. Section 3 turns to an examination of the technical issues related to alignment in complex systems, while Section 4 describes research on the impact of social alignment in schools and districts. Section 5 explores how systems might create policy frameworks to support balance and coherence across systems. The sixth and final section of the report briefly sets out broad policy principles for achieving balance and coherence in complex education systems.

### 1.1 A note on the international research base

6. Typically, research on standards-based systems addresses either technical or social issues of alignment, but rarely both. A few researchers have started to address this gap, and their work informs this report. But there is still a need to strengthen the theoretical framework and to build the evidence base. This report highlights some of the important gaps in the international research.

7. It should also be noted that standards-based assessments are very new in many OECD countries. As a result, much of the research informing this report is from countries with a longer history with these approaches – Australia, Canada, the United Kingdom, and the United States. Research on external school inspections and internal school self-evaluations, by contrast, is more representative of experiences in European countries, where these approaches have a longer tradition. Certainly, the research from different countries is not always easily transferable across different policy contexts and traditions. Nevertheless, all countries can learn from the experiences of others as they seek to achieve balance and coherence in educational policy, practice and research.

### 1.2 A note on the terminology used in this report

8. Every discipline uses terms in a specific way in order to communicate and clarify important concepts – and of course, these also vary across languages. A few of the key terms important for describing standards-based approaches in the English language, and as used in this report, are presented in Box 1.

**Box 1. Terminology used in this report**

***Standards*** – refer to descriptions of what students should know (content standards) and be able to do (performance standards) at different stages of the learning process. The standards may be set out in a separate document, or may be embedded in curriculum.

***Competences*** – refers to the proven ability to use knowledge, skills and attitudes (*e.g*. personal and social skills) in work or learning environments and for professional and personal development. The notion of key competences is generally independent of subject-based competences. An individual applies his/her competences to specific tasks or problems, and is also able to transfer knowledge and skills to different situations and contexts. Note, however, that definitions of competence vary across countries. There is no standard definition in English or across the range of European languages and systems (Gordon *et al.*, 2009).

***Assessment*** – refers to judgments of individual student performance and achievement of learning goals. It includes classroom-based assessments as well as large-scale, external tests and examinations.

***Evaluation*** – refers to judgments on the effectiveness of policies, schools and school systems, and/or targeted learning programmes. It encompasses school inspections, school self-evaluations and targeted programme evaluations (*e.g.* of a new reading programme, or a system-wide intervention in early childhood education).

Note that, while the terms assessment and evaluation are used interchangeably in the English language, education specialists often make careful distinctions between the two terms to clarify their different roles.

***Teacher appraisal*** – judgments of individual teacher performance.

***Curriculum frameworks*** – provide a blueprint for implementing content and performance standards. Countries and regions take different approaches to how they design curricula, but in general, they establish broad guidelines, leaving room for teachers to decide upon methods and materials.

***External vs. internal evaluation*** – the distinction here is in regard to *who* conducts the assessment or evaluation. Internal evaluations (school self-evaluations) are conducted by project or school staff. External evaluations (*e.g.* school inspections and targeted project evaluations) are conducted by an individual or team who are not part of the school staff.

## SECTION 2. STANDARDS-BASED ASSESSMENT AND EVALUATION ACROSS OECD COUNTRIES: COMMON MOTIVATIONS, DIFFERENT APPROACHES

9. Standards-based approaches represent a fundamental shift in educational governance and goals. These approaches have taken hold as the majority of OECD countries have decentralised education systems. It is believed that schools with greater autonomy will have more freedom to innovate, to tailor education to the local context, and to meet the needs of diverse students. Schools are given more freedom to decide upon the content and methods they will use.

10. At the same time, schools are held accountable for helping *all* students to meet centrally defined standards for learning. Education systems have long been charged with providing equitable opportunities for students to learn. But schools have also traditionally played a sorting role, guiding students toward different tracks based on academic performance in primary and lower secondary school years. With the introduction of standards-based approaches, schools are now charged not only with providing access and opportunity, but also ensuring equity of outcomes. Schools are thus expected to lay the basis for lifelong learning, helping students to develop the sophisticated skills they will need to navigate economic and social changes.

11. Standards-based approaches also aim at improving the quality and delivery of education across systems. Data gathered in assessments and evaluations are used to determine where systems are performing well, and where they need to make improvements. Education authorities at national, regional and local levels align resources and coordinate efforts to address needs. Data gathered through external monitoring help to highlight any existing inequities within systems, as well as progress made in closing those gaps. Policy makers are better able to develop coherent, coordinated and strategic responses to needs.

12. The OECD's (2010) Programme for International Student Assessment (PISA) has found that external standards-based assessments are positively associated with higher performance of school systems and that performance differences between schools with students of different social backgrounds are, on average, lower in countries where more schools use standardised tests. In addition, several studies on "opportunity to learn" (OTL) provide significant evidence that the focus, content coverage and flow, and cognitive demands in curricula have a strong and direct impact on student achievement (see Gamoran *et al.*, 1997; Porter and Smithson, 2001; Smithson and Collares, 2007 cited in Schmidt and Maier, 2009).

13. Standards-based approaches have spread rapidly across OECD countries. At the same time, countries have adapted the key elements of standards-based systems to their own educational contexts and cultures – how they define standards, how they balance incentives and support, how they measure school and student performance[2].

[2] The information for different countries has been brought together from country reports to UNESCO's World Data on Education database (*www.ibe.unesco.org/Countries/WDE/2006/index.html*) and, for European countries, reports to the Eurydice's Eurybase for 2008-2009 (*http://eacea.ec.europa.eu/education/eurydice/eurybase_en.php*). This overview is not intended to provide an exhaustive review of each country's standards-based system.

- Several countries set out standards for learning and competence development in central and/or regional documents (Austria, Australia, Canada, Germany, Italy, New Zealand, Switzerland, Turkey, the United States and the United Kingdom). In other OECD countries, standards are embedded in curricula and/or framework documents (the Czech Republic, Denmark, Estonia, Finland, France, Greece, Hungary, Iceland, Ireland, Israel, Japan, Korea, Luxembourg, the Netherlands, Norway, Poland, Portugal, Slovak Republic, Sweden, Spain) (UNESCO, 2006/07). Box 2 provides an overview of key competences in selected European countries.

- All countries stress academic achievement in core subjects, including language, mathematics, science, history and social sciences. A few countries have developed standards and teaching guidelines for technology (the Flemish Community of Belgium, the Czech Republic, New Zealand and Scotland) (UNESCO, 2006/07).

- Several countries also incorporate goals for cross-curricular competences, such as learning-to-learn (the Flemish Community of Belgium and Japan), problem solving (the Czech Republic and New Zealand), social skills, including skills for cooperation (the Flemish Community of Belgium and New Zealand), and individual development (Canada, the French-speaking community of Belgium, Finland and New Zealand) (UNESCO, 2006/07).

- More than half of OECD countries administer periodic national assessments of student performance [Australia, Belgium (French and Flemish communities), Canada, England, Estonia, Finland, France, Hungary, Italy, Korea, Luxembourg, Mexico, Norway, Scotland, Spain, Sweden, Turkey, and the United States]. In Germany, the *Länder* are currently cooperating in the development of external examinations that will provide comparable information on student performance (OECD, 2009a; UNESCO, 2006/07).

- National assessments are administered at different points in the academic year. The French-speaking community of Belgium, France and Spain administer assessments early in the academic year, targeting students who have just made key transitions in their schooling, *i.e.* from primary to lower secondary school, so that results may be used diagnostically. However, the majority of countries administer national assessments later in the academic year, using the results to monitor school and student performance and to shape provision for future student cohorts. Different countries use different sampling methods (*i.e.* census sampling *vs.* population sampling of students in given year levels).

- Countries and regions take very different approaches to the design and implementation of standards-based assessments. They may rely primarily on multiple-choice formats (as in many American states), or emphasise open formats with performance-based tasks, such as essays, oral presentations, and collaborative-problem solving, or have a combination of the two. Box 4 provides an overview of the advantages and disadvantages of different testing formats.

- A few countries attach high stakes to the results of external assessments, for example the threat of shutdown or reconstitution for underperforming schools (Canada, the United Kingdom and the United States). The majority of OECD countries report that they publish assessment and/or evaluation results, which many teachers perceive as adding to stakes (McDonnell and Choisser, 1997). They include Australia, the Czech Republic, Denmark, England, Hungary, Iceland, Ireland, Italy, Japan (school self evaluations only), Korea, Netherlands, New Zealand, Norway, Portugal, Scotland, Sweden (both school self evaluations and external inspection reports), Turkey, and the United States. The Flemish Community of Belgium, by contrast, legally prohibits publication of results on a comparative basis (UNESCO, 2006).

- Ten countries report to the OECD's *Education at a Glance* survey (2009a) that they promote both external school inspection (usually on a tri-annual basis) and internal school self-evaluation (usually on an annual basis) (the States and Territories of Australia, the Czech Republic, England, Iceland, Korea, New Zealand, Portugal, Scotland, Sweden and Turkey).

14. In their broadest outlines, we can observe a harmonisation of policy approaches to educational accountability and improvement at the international level. At the same time, there are fundamental differences between countries. For example, the balance of power between central authorities and local school districts varies across countries. Local traditions, cultures and values in education also have a strong impact on how policies are implemented.

**Box 2. Defining competences – Selected examples from European countries**

European countries define competences in a variety of ways. These different approaches have implications for how learning is assessed.

- Austria defines "dynamic skills" (*Dynamische Fertigkeiten*), which are transversal, and not tied to specific subjects.
- Finland has introduced the concept of "themes", *i.e.* challenges with social significance.
- France defines the foundation (*socle*) competences as including both subject-based and cross-curricular competences.
- Germany defines subject-independent, general competences essential for learners' personal and working lives. The key competences apply to different subjects and subject areas, are useful for solving complex tasks in real-life contexts, and are transferrable to situations not covered in the curriculum.
- Greece has introduced an interdisciplinary cross-curricular thematic framework (DEPPS), linking all subjects horizontally.
- Hungary defines competences as "capabilities"; values are included in the capabilities (*i.e.* the capability to understand and apply norms and values).
- In Italy, schools help each primary school student to define his or her personal competences in each subject and cycle.
- The Netherlands defines "core objectives" related to specific subjects and "general objectives" (cross curricular).
- Portugal has introduced essential competences – that is, the development of skills and attitudes helpful for using knowledge in different situations.
- Slovenia defines key competences in thematic fields (*e.g.* learning to learn, social skills, ICT, entrepreneurship, environmental responsibility, etc.).
- In Sweden, goals represent a broad range of developmental goals, and cover all aspects of education. Sweden does not use the term "competence".
- Across the United Kingdom and in Ireland, the terms "skills", "core skills" and "key skills" are used. There is a strong emphasis on personal "capabilities" (Northern Ireland) and on the need for young people to become active members of society (Scotland). England emphasises skills for independent thinking, creativity, teamwork and effective participation, and self-management.

*Source*: Gordon *et al.*, 2009.

## SECTION 3. THE COMPONENTS OF STANDARDS-BASED ASSESSMENT AND EVALUATION: TECHNICAL ALIGNMENT

15. The logic underlying standards-based assessment is quite straightforward. Systems set goals for student learning through standards, set out the specific content for learning in curriculum, and measure attainment through external assessments. In several countries, external inspectorates follow the quality of educational provision in schools. Schools may be required to develop self-evaluations of their progress. Data gathered through these processes help teachers to identify gaps between student attainment and standards.

16. At the same time, these straightforward systems involve sophisticated knowledge of student learning and progression and of educational measurement technologies that can capture higher-order learning. There is increasing recognition of the need to adapt standards-based systems to the complexity of educational systems and of learning processes. The following describes technical challenges involved in the design of the different components of standards-based systems, as well as different approaches to achieving better balance and coherence across systems.

### 3.1 Standards

17. Standards typically describe what students should know and be able to do at different stages of their schooling. The process of standard setting inevitably involves political and cultural debates. Those responsible for developing standards in any country face several challenges. Indeed, Cizek (2001) has described standard setting as requiring a blend of artistic, political and cultural ingredients. To this list we might add knowledge of how students learn and progress.

18. While no systematic descriptions of the process of standards or competence development in OECD countries were identified for this report, the focus and intensity of any controversy are necessarily unique to each country. For example, standards writers may have difficulty agreeing on the knowledge and skills that are most important. While the majority of OECD countries now promote skills for "learning-to-learn", including skills for problem-solving, critical analysis, as well as supporting students in developing greater autonomy, and so on, there may still be deep-seated tensions about the goals of education. Such "culture wars" (Finn and Kanstoroom, 2001), may lead to the development of standards that are vague (thereby avoiding controversy), or at the other extreme, standards that are overly detailed, making it difficult to identify priorities for learning, and providing little useful guidance for instruction or the development of assessments (Chudowsky and Pellegrino, 2003).

19. Standards developers must also decide where to set targets for attainment and whether to set the same standards for all students to achieve at the same rate. Standard setting also involves setting out "cut-scores" for broad categories of student proficiency (*e.g.* below basic, basic, proficient, advanced)[3]. These

[3] Standard setting studies establish the validity of the standard setting process (Zieky and Perie, 2006; Cizek and Bunch, 2007). Valid inferences from cut-scores (establishing the different proficiency categories) are based on the assumption that the content and performance standards are aligned with tests specifications. At the same time, there is little agreement as to how to evaluate alignment of standards and assessments (Bhola *et al.*, 2005). Pant and colleagues (2009) find an imbalance between the key role assigned to standard setting and the comparatively weak sources of validity for standard-setting procedures.

"criterion-referenced interpretations" of performance are intended to measure students' progress toward learning goals, rather than in competition with their peers, which is more typical of "norm-referenced interpretations"[4].

20. There are tensions between the idea of setting standards for excellence for all students as well as supporting individual differences and interests. These are fundamental concerns for systems considering how to support both equity and quality (Linn, 1998). Policy makers may choose to set rigorous standards to communicate their efforts to raise school performance to the broader public. There is research supporting the view that students benefit from high expectations (Bransford *et al.*, 1999). But there are also concerns that unreasonably high targets increase incentives for teachers to "teach to the test", thereby raising student scores, while not actually having an impact on student learning (Koretz, 2005) (discussed in more detail below).

21. Some OECD countries have addressed these challenges by setting standards at two levels. For example, Belgium (the Flemish Community) has established minimum objectives for knowledge, skills and attitudes to be attained by the majority of pupils. The Czech Republic notes that students should attain competences "at a level accessible for them" (Eurydice, 2008/09, p. 88). In Victoria, Australia, standards are based on learners' progress on developmental continua (see Box 3). In Scotland, student progression through the curriculum is based on the results of assessments, which are used formatively, to adapt teaching and learning to each students' needs (Eurydice, 2008/09; UNESCO, 2006).

22. Linn (2003) suggests that benchmarks for school and student achievement might be based on the top 10% of schools that have made the most rapid gains over a specified time period (*e.g.* five years). In addition, researchers may explore the effectiveness of systems that set both minimum and higher-level standards. Another approach would be to refine and expand the proficiency classifications, and to tie these classifications to typical student progression within a given subject domain. For example, the United Kingdom's Task Group on Assessment and Testing (TGAT) formed in 1988 recommended the development of a more finely tuned, ten-level criterion-referenced system, with a single set of criteria spanning the age range. These were to be based on evidence that student attainment at any given age may cover a span of several years (Black, 2000).

23. Standards need also to be clear and detailed enough that the knowledge and skills students are expected to attain are readily apparent (Commission on Instructionally Supportive Assessment, 2001). Standards writers may also need to prioritize content and to help teachers to make effective choices within the limited time they have. Indeed, the earliest models of OTL suggested that any student could achieve some level of mastery of a subject if given enough time (Bloom, 1968; Carroll, 1984). Too many attainment targets, on the other hand, may make it difficult for students to learn anything in depth, and may be perceived by teachers as being overly prescriptive. In response to such concerns, the Netherlands significantly streamlined objectives for learning, paring down from an initial set of 464 attainment targets in 1987, to 122 core objectives in 1993. These core objectives were further reduced to 103 in 1998 and then to 58 in 2006 (SLO, 2007).

[4] Assessment results are typically reported as either "norm-referenced" (*i.e.* describing student performance relative to his/her peers), or "criterion-referenced" (*i.e.* describing student performance relative to a performance target). Most standards-based systems prefer criterion-referenced reporting, which is more effectively aligned with goals for all students to meet standards and specific performance goals.

**Box 3. Australia: The Victorian Essential Learning Standards**

The Victorian Essential Learning Standards (the VELS) are based on national and international research about how students' typically progress from novice to expert levels of performance. The progressions take into account students' social, emotional and cognitive development.

As described by the Victorian Curriculum and Assessment Authority, student development involves:

- Pattern recognition;
- Acquisition of relevant content knowledge, reflecting deep understanding of the subject matter;
- Application of knowledge in ways appropriate to context;
- Retrieval of key aspects of knowledge with a degree of automaticity;
- Flexibility in new situations.

The VELS include "progression points" describing:

- Evidence of student progression;
- Guidelines on student assessment in relation to standards;
- Assessment maps;
- Adaptation by schools to reflect curriculum structure and when new content and skills are taught and assessed.

VELS also provides tools to assist with whole school curriculum planning, following the standards.

*Source*: *http://vels.vcaa.vic.edu.au/overview/index.html*.

## 3.2 Curriculum

24. Curriculum plays a vital role in any standards-based system, bridging standards and assessments. In several countries, standards are embedded in curriculum (see list of countries above). The curricula for different subjects concretely set the content and expectations for student learning at different stages of their learning (*e.g.* grade levels, end of primary schooling, etc) and often provide guidance on learning activities, although they do not necessarily define expected performance levels (Cizek, 2001).

25. As with standards, there is the important question of how many topics to address within a limited period of time, and how to ensure coherence in presentation of subject matter. Here, coherence refers to the sequencing of topics in a way that reflects the logical structure of a given discipline, and engages students in reasoning through subject matter. Without attention to sequencing, students may learn topics as loose collections of isolated facts. Content emphasized across standards, curriculum, textbooks and assessments may also be inconsistent (Schmidt and Maier, 2009).

26. Schmidt and Maier (2009) found that curricula in the United States typically cover many more topics than in other countries. They also found that textbooks in the United States are often more than twice as long as those found in other countries. But the problem of overloaded curricula is international.

Teachers frequently feel pressure to move quickly through subjects and unable to take the time to ensure that students have understood a topic before beginning the next (OECD, 2005a). Differences in students' OTL may also occur simply because teachers generally have a certain level of freedom to interpret standards, shape curriculum and decide upon sequencing and content.

### 3.3 Student assessment and school evaluation

27. The effectiveness of standards-based systems ultimately depends upon the validity, reliability and usability of the information gathered in large-scale assessments and school inspections and school self-evaluations. *Validity* refers to the degree to which assessments and evaluations measure what they are intended to measure (*i.e.* how well they are aligned with standards and curriculum). *Reliability* refers to the consistency and stability of results across student populations or across schools. *Usability* refers to how policy makers, school leaders and teachers make sense of and respond to assessment and evaluation results. In regard to *usability* of student assessment results, Abu-Alhija (2007) notes that the tests must be easy to administer and accessible to a wide range of students. Ease of interpretation of the results is also important. Alignment of assessments and evaluations with standards and curriculum is crucial to usability.

28. Usability also relates to the level of detail and timeliness of data. At higher levels of education systems, aggregate data gathered periodically are adequate for decisions related to allocation of resources or adjustment of policies. In classrooms, teachers need more detailed and frequent information on student learning (Black and Wiliam, 1998). Schools may implement more regular assessments aligned with state-level assessments in order to track student progress toward goals.

29. Many of the challenges involved in alignment of standards and large-scale assessments have to do with the difficulty of measuring higher order skills such as problem solving, reasoning and communication. Cognitive scientists have made a great deal of progress in understanding how students learn – including typical learner misconceptions, progression from novice to expert performance, effective learning environments, and so on (Bransford *et al.*, 1999; Pellegrino *et al.*, 1999). However, traditional testing methodologies that treat tasks as discrete items cannot easily capture complex performances and processes. While performance-based assessments are more effective in this regard, there are some concerns regarding the reliability of scores – particularly when scores are awarded by human raters. Caldwell and colleagues (2003) have found, however, that effective training can improve the reliability of scores on performance-based assessments. Box 4 presents an overview of the advantages and disadvantages of several popular assessment formats.

**Box 4. Advantages and disadvantages of different assessment formats**

- *Multiple-choice assessments* provide reliable data on student performance, as assessment are machine-scored, and are therefore less expensive to administer. Well-designed multiple-choice questions may be used to assess higher-order knowledge. They cannot, however, measure skills such as the capacity to develop an argument. Poorly designed multiple-choice assessments are also prone to measurement error (*e.g.* students may misinterpret questions or may make random guesses).

- *Computer Adaptive Tests*, as implied by their name, adapt questions for the test-taker. Students who answer questions correctly are directed to a more difficult question, and those answering incorrectly are directed to an easier question. Since the test is adapted according to each student's responses, no two students take the same test, and it is not possible to compare student performances. Computer-based, adaptive testing (CAT) is generally considered as providing more precise scores of student performance than typical standardised assessments. However, CAT demands a very high number of test questions, which increases development costs. Also, CAT typically draws heavily or solely on multiple-choice formats.

- *Performance-based assessments,* which include tasks such as oral presentations, essays and collaborative-problem solving, are more effective at capturing more complex performance and processes. However, there are concerns regarding the reliability of these assessments – particularly when scores are awarded by human raters. They are also more expensive to administer and score.

- *Computer-based performance assessments* may potentially assess more complex performances through simulation, interactivity, collaboration and constructed response formats. Increasingly sophisticated ICT programmes that score "open-ended performances" may address concerns regarding reliability of human-scored assessments, and validity of multiple-choice assessments that do not effectively measure higher-order skills.

30. Mislevy and colleagues (1998) argue that test developers should first focus on cognitive goals to be measured, and then turn to the content goals – a reversal of the usual process of test design. Test developers first determine the skills they want to measure – for example, students' reasoning processes. They then identify how students use these skills in different subject domains (*e.g.* mathematics, science, social sciences), and decide upon the kinds of tasks that will provide evidence of student capabilities, including performance-oriented tasks. This approach focuses on cognitive demands, rather than specific content, and is therefore more effectively aligned with standards related to higher-order learning skills.

31. It is also important to remember that no single test can measure all the knowledge and skills students are expected to learn in a given domain, as tests take place over a limited period of time and cannot cover all learning priorities. To the extent that teachers focus on tests instead of standards, they will narrow teaching and learning. Teachers may re-allocate time and re-align priorities in order to spend more time on content or performances likely to be covered in the tests, thus narrowing teaching. They may also coach students in test-taking skills (Smith and Rottenberg, 1991; Cizek, 1998; Popham, 2002; Stecher, 2002; Koretz *et al*., 2001, 2005). All of these concerns tend to be magnified in systems attaching high stakes to the results of standards-based assessments

32. Validity, reliability and usability also apply to external school inspection and internal school self-evaluation. Guidelines for inspection should align with standards and include the criteria inspectors will use to judge school performance. Inspectorates should also be able to provide evidence of inter-inspector rating reliability. This is particularly important in situations where schools face high stakes (Fidler *et al.*, 1996). In the context of school self evaluations, staff may need to achieve consensus regarding goals for the evaluation, and the criteria by which they will judge school performance. Staff may also need training in methods of data gathering and analysis.

### 3.4 A comprehensive approach to alignment

33. Given the challenges involved in measurement, experts in educational measurement consistently advise that important decisions should not be based on a single, high-visibility test score. Rather, standards-based systems should incorporate a range of measurements that serve needs of policy makers as well as practitioners. At the policy level, aggregated data highlight trends in student achievement and help to identify outliers in the system (*e.g.* schools performing at both the high and low ends of the spectrum). Classroom-level data provide real-time information on individual student needs, affording teachers the opportunity to adjust instruction (Pellegrino *et al.*, 2001; Wiliam, 2006).

34. Local areas and individual schools may also develop their own standards-based assessments to supplement regional or national measures and to better reflect the local context. The OECD's (2010) PISA found that where schools have greater autonomy over both what is taught and how student learning is assessed, students tend to perform better. Murnane and Nelson (2005) similarly found that schools that use a range of fine-grained measures on different dimensions of performance and develop strategies to address identified problems are consistently more successful[5]. Essentially, teachers discover through trial and error those approaches that are most likely to help particular students.

35. Multiple measures also help to avoid the risk of incorrect decisions based on measurement error, and in high-stakes systems, may lower the risk of score inflation (*i.e.* gains in test scores overstate improvements in actual student learning) (Koretz, 2005). Different assessment methods provide different kinds of information as to how instructional strategies are influencing learning and can help build the knowledge base on "what works" (Baker, 2004; Herman, 2005; Abu-Alhija, 2007).

36. Technical alignment of standards, curriculum and assessment and evaluation is vital if systems are to develop high quality feedback systems. At the same time, the degree to which data are used for improvement depends on social alignment in schools and districts. It is at this intersection of the technical and social where practitioners align efforts where change and improvement can occur.

[5] Murnane and Nelson also refer to progress in research on Cystic Fibrosis, noting that variation in success of health centres in treating the disease had little to do with differences in client mixes, standardised treatment techniques or credentials of the staff. Rather, the most effective centres monitored key indicators of patient health, rapidly identified those with decline in lung function, trained staff to diagnose the source of the problem, and to work with the patient to improve treatment.

## SECTION 4. LEARNING IN COMPLEX SYSTEMS: SOCIAL ALIGNMENT

37. The social aspect of alignment is often overlooked in discussions regarding standards-based systems. Social alignment refers to the social capital in systems, including shared values, motives and efforts (Baker, 2004; Hargreaves, 2003). In socially aligned systems, institutions and actors work together to define challenges and to consider alternative courses of action. This alignment is vital for system learning and improvement.

38. Several recent studies have highlighted exemplary cases where leaders in districts and schools regularly collaborate, referring to data on student learning to adapt and improve instruction. Yet these cases appear to be more the exception than the rule. The OECD's (2009b) Teaching and Learning International Survey (TALIS) points to a lack of deep professional collaboration in schools across a range of countries with large cultural differences. The majority of teachers responding to the TALIS[6] survey reported that they coordinate information on teaching and exchange material more frequently than they engage in joint professional learning activities or other more intensive forms of professional collaboration. Moreover, three-quarters of teachers participating in TALIS reported that there were no incentives to participate actively in school improvement efforts. Nor would individual teachers receive recognition for improvements in the quality or innovativeness of their teaching.

39. Writing on complexity, accountability and improvement in education systems, O'Day (2002) proposes that the structure and norms of many schools, where teachers work in "independent and isolated classrooms", buffers individuals and schools against change and prevents mutual learning (p. 8). Loose coupling across the links and layers of education reinforces this isolation. Where systems limit interaction and interdependence across layers, they also limit opportunities for learning and adaptation.

40. At the same time, schools and teachers face an increasingly complex set of demands. Teachers are expected to help all students to achieve to high levels, to work with increasingly diverse populations of students, to promote social cohesion, to stay in touch with new knowledge and to help students develop higher order skills, to adopt new classroom technologies and to adopt new approaches to student assessment (OECD, 2005b).

41. The question for national and regional policy makers, then, is how to best balance external, bureaucratic controls, which are vital for ensuring quality, equity and accountability across education systems, and support for internal, professional controls, with schools and teachers taking collective responsibility for student learning (O'Day, 2002). Indeed, there is increasing attention as to how systems might strengthen social alignment through tighter coupling across districts and schools, with a strong focus on instructional leadership (Sykes *et al.*, 2009).

42. The studies discussed in the following pages explore school level collaboration and collaboration between schools and across districts. It should be kept in mind that while research on the impact of standards-based approaches is accumulating, there are still significant gaps. The purposes and concerns of

[6] Twenty-three countries participated in the survey: Australia, Austria, Belgium (Flemish Community), Denmark, Hungary, Iceland, Ireland, Italy, Korea, Mexico, Norway, Poland, Portugal, Slovak Republic, Spain and Turkey. Participating partner countries included: Brazil, Bulgaria, Lithuania, Malaysia and Malta. Estonia and Slovenia participated as partner countries but have since become member countries.

different analyses vary significantly. Indeed, there is a need for stronger theory-driven research, hypothesis testing and more systematic study in this area (Berends, 2009; Sykes *et al*., 2009). Nevertheless, the research discussed below highlights current concerns in a range of contexts and points to areas where further research and development are needed.

### 4.1 School level collaboration

43. Several studies from the United States have found that standards-based approaches have been highly effective in focusing attention of schools and teachers on priorities for student learning (Herman and Baker, 2009). Schools are more likely to align curricula with standards and to implement benchmarking assessments to gauge student progress when standards clearly specify goals for learning (Datnow and Park, 2009; Koretz, 2005; Linn, 2005). At the same time, school leaders' and teachers' capacity to interpret data from standards-based assessments and evaluations (both external and internal) varies a great deal (Abelman *et al.*, 1999; Elmore and Fuhrman, 2001; Spillane and Zeuli, 1999; Winkler, 2002). Diagnosing the source of student difficulties and developing appropriate remedies for different students is often challenging even in the highest performing of schools.

44. At the school level, there is some evidence that strong teacher-to-teacher trust, a shared focus on instruction and student learning, and experience are associated with higher levels of student achievement. In general, schools with more socio-economically advantaged student populations have higher capacity along these dimensions and are also more likely to develop coherent strategies to address student needs (Elmore, 2001 cited in O'Day, 2002). The Consortium for Policy Research in Education (CPRE) in Chicago found that among low-performing schools that had been placed on probation, those that had previously developed strong cultures of peer collaboration were able to exit probationary status relatively rapidly (from 1996 to Spring of 1998). The CPRE survey data showed that teachers in the more successful schools had stronger levels of trust. As O'Day (2002) notes, these factors indicate a strong level of internal school control and accountability

45. A large-scale longitudinal study by Seashore Louis and colleagues (2010) based in the United States found that collective leadership at both the school and district levels were associated with stronger impacts on student achievement (collective leadership refers to the extent of influence organisational actors and other stakeholders exert on decisions)[7]. At the school level, collective leadership focused on instructional improvement had a significant impact on teachers' working relationships, and on student achievement. The case studies on teacher collaboration and school leadership in Norway (Box 5) and Ontario, Canada (Box 6) describe different approaches to teacher collaboration and school leadership focused on monitoring performance and improving instruction.

46. At the international level, findings from the OECD's (2009b) TALIS cited above highlight a lack of deep professional collaboration within schools across countries with different policy contexts. However, the report notes evidence of cultural bias in some of the TALIS results, so it is impossible to conduct any comparative analysis. More detailed information on the number of hours teachers spend on different school-level tasks, the role of school leadership in encouraging professional collaboration, and the focus on instructional improvement might facilitate further analysis.

[7] The five-year study (2003 to 2008), gathered data from respondents in nine states, 43 school districts, and 180 schools at primary, lower and upper secondary levels. The researchers also had access to data for student achievement in literacy and mathematics.

**Box 5. Case studies on teacher collaboration and school leadership in Norway**

A common element in schools [visited for the case studies] was that student learning is the focal point of the school's philosophy. This was expressed in the need for a productive, collective learning culture and the interactions between teachers and students. Another important element in relation to teaching practice was the demands in the national curriculum. These demands have been made more visible and forceful among other factors because a new national accountability system has been put in place (Møller *et al.*, 2005).

Within schools it seemed that to a certain extent leadership was distributed to leadership teams and to teacher teams – as one principal put it: "The principal emphasises a shared leadership because of the teachers' feeling of 'co-responsibility', and at the same time he stresses that he is responsible". The distribution of leadership tasks had been in such a way that the principal, it seemed, still retained the final powers of decision making, but at a general level teachers were involved in school development issues and they were expected to take significant responsibility and decisions in their everyday work.

This trend seemed to be elaborated in relation to students, who were encouraged to participate in planning and evaluating teacher and learning. There were strong interconnections in school thinking about learning outcomes and learning processes and relations.

The criteria for selecting "beacon schools" were also very influential for the case schools: pedagogical creativity and development; systematic student assessment and systematic approaches to development strategies–based evaluation; systematic approaches to development, including learning and learning environment; an understanding of successful leadership, including flexibility of instruction, democratic attitudes towards various members of staff and ongoing strategies for development of the organisation.

*Source*: Moos *et al*. (2008).

**Box 6. Case studies on teacher collaboration and school leadership in Canada (Ontario)**

Until the implementation of NPM [New Public Management] principals in schools were seen as administrators only, without educational leadership tasks. However, principals in this case have since then used the provincial and district initiatives as points of departure for setting goals for their schools. They saw their schools in the big picture, but also stressed the need to have a broader view of learning than what is reflected in the provincial achievement tests alone. On that basis they communicated clear goals and high expectations for student achievement to staff. Principals gave individualised support, intellectual stimulation and they acted as role models. They also acted as members of various work teams and give teachers as well as parents significant decision-making roles.

Principals act in a forceful way in planning and supervising instructions that often include monitoring teachers' practice and modifying schools structures, like the school day, to maximise learning. They exercise strong management skills and use systematically collected evidence. In order to maximise learning, principals protect learning time from external, excessive and distracting demands.

Principals from Ontario set clear directions for student learning based on an extended interpretation of the demands and standards set by the educational authorities and they gave teachers support and room for manoeuvre when it came to choosing the means of instruction. Teachers, however, were kept on a short leash as principals monitored student outcomes and teacher practice.

*Source*: Moos *et al.* (2008).

### 4.2 The impact of sanctions on teacher motivation and collaboration

47. In an effort to learn more about the interaction of high stakes[8] incentives and teachers' professional motivations and practices and interactions with colleagues, Finnigan and Gross (2007) surveyed teachers in ten low performing schools in Chicago. All of the schools included in their study had been placed on probation. They found that material incentives (rewards and sanctions) had less influence on teachers than professional motives (their individual purposive motives, or their solidary status as part of a community). However, teachers who felt pressure to increase test scores were less likely to participate in professional development opportunities outside their school (58% as compared to 81%), and were less likely to increase professional collaboration related to reading instruction (80% as compared to 91%). The survey data showed that accountability policies with high stakes might have a counterproductive effect on teacher motivation over time, particularly in schools that continue to struggle. The study also highlights a lack of coherence in school level strategies to improve instruction, and points to a need for capacity building at this level (see also Goertz and Massell, 2005).

48. Several studies have also shown that although teachers in schools on probation often intensify effort and are more focused on student achievement, they do not fundamentally alter instructional approaches in response to the incentives of standards-based systems (see Firestone *et al.*, 1998, 2000; McDonnell and Choisser, 1997; Mehrens, 1998). More typically, strategies involve changing the content emphasised (re-alignment of curriculum), time spent on test preparation (re-allocation of time) and the way in which students are grouped (Jacob *et al.*, 2003). They may also coach students in test-taking skills.

49. To the extent that high-stakes assessments include knowledge and skills deemed to be central to student learning, re-allocation and re-alignment are appropriate. However, no single test can measure proficiencies in any given domain exhaustively; assessments take place over a very limited period of time and cannot cover all learning priorities. They are therefore considered as proxies for wider achievement. But if teachers focus narrowly on content most likely to be on the assessment (*i.e.* they teach to the test), they no longer serve this purpose. If standards are not well designed or if they lack specificity, teachers will have greater incentive to teach to the test, particularly in systems with high stakes (Koretz, 2005; Linn, 2005).

50. Coaching for improved test performance may focus on substantive and/or non-substantive aspects of test performance. For example, if teachers notice certain patterns in tested content and then prepare students to focus on that content, they are engaging in "substantive" coaching. Non-substantive coaching may involving helping students develop tricks for effective test taking (*e.g.* how to recognise distractors in multiple choice tests, while not necessarily learning how to recognise the correct answers)[9] (Koretz, 2005; Popham, 2002).

51. These strategies – re-allocation, re-alignment and coaching – all lead to score inflation. When test scores overstate improvements in actual student learning, and those interpreting results are unable to identify meaningful progress or to develop strategies to meet student needs.

[8] Finnigan and Gross note that teachers viewed school probation as a high stakes strategy, even though there were no threats to school funding or teachers' jobs.

[9] Crocker (2005) proposes that it is appropriate to discuss effective problem-solving strategies for tests and to help students understand how they should behave in the testing milieu. But teachers should also use a variety of formats in their own tests in order to maintain focus on the development of broader cognitive skills– not just those used in the major assessments.

### 4.3 The role of teacher appraisal and feedback in school improvement

52. The OECD's (2005b) *Teachers Matter* report found that many countries lack coherent and well-resourced systems for teacher appraisal, even when such appraisal is required. As a result, teachers do not receive guidance on professional development priorities, do not receive regular feedback on the quality of their work, and may develop a sense of professional isolation. The lack of appraisal thus has a negative impact on social alignment within and beyond schools.

53. At the same time, results of the OECD's (2009b) TALIS indicate that teacher appraisal and feedback can influence teachers' efforts to improve instruction. Respondents to the survey agreed that the stronger the emphasis on a particular aspect of teacher work, the greater its impact. TALIS identified statistically significant relationships between areas emphasised in appraisal, and changes in teachers' subject knowledge and instructional practices. This was the case for all countries participating in the survey. Moreover, the survey found that teachers who participate in professional development opportunities, which may be spurred by appraisal and feedback, are more willing to try new instructional methods.

54. However, most TALIS respondents noted that school leaders do not place a clear focus on any instructional area or school-level involvement, but tend to spread emphasis relatively evenly across the 17 different areas listed in the survey questionnaire. The TALIS (OECD, 2009b) report concludes that countries might encourage increased frequency of teacher appraisal and feedback (through regulations or other policy levers). Evaluative frameworks may also be improved to ensure that appraisals and professional development are linked to policy and school priorities for improvement.

55. Ballard and Bates (2008) suggest that appraisals that are based on a clear definition of high quality teaching will have more impact. This is in line with other research cited above, noting that teachers while maybe willing to "try anything" to help improve students' achievement, are unlikely to succeed when there is no clear instructional strategy (Hargreaves, 2003; Finnigan and Gross, 2007). Black (2000) asserts that professional development is most successful when it is matched to teacher capacity, and allows teachers to evolve or rebuild theories of teaching and learning in a way that supports and gives coherence to their practice.

56. Appraisal aligned with school priorities and needs might strengthen links between individual teachers and their peers while also supporting coherence in school strategies. To the extent that the appraisal process rewards individual improvement and also encourages collaborative work, teachers will be more likely to move beyond their own classrooms and to engage in collective learning.

57. A stronger focus on professionalism also implies the need for significant, sustained and focused investments in professional development. Teachers need strong knowledge of the subject they are teaching and an understanding of student development within that domain. They also need to develop skills to assess learning needs and a broad repertoire of strategies to meet a range of student needs.

58. In view of the research discussed above, policy makers will also need to consider a complex set of factors affecting teacher motivation if they are to create a more effective mix of incentives. These factors include teachers' beliefs in their own and their students' capacity; the role of purposive, solidary, and material motivations for different teachers and in different contexts; and, the importance of instructional leadership at school and policy levels.

### 4.4 School self-evaluation and school improvement

59. School self-evaluations (SSE) require school leaders and teachers to work together to identify both strengths and weaknesses, and to develop long-term strategic plans for improvement. More than half of OECD countries mandate annual, or in a few cases tri-annual, SSE. As might be expected, the effectiveness of SSE varies (Hofman *et al.*, 2009).

60. Most studies on SSE focus on the experiences of staff involved in the process, rather than the impact or uses of data resulting from the process (Blok *et al.*, 2007). Nevo (2002) argues that schools with experience in self evaluation are more likely to adopt a constructive attitude toward external inspection and to make more productive use of the results, he also notes that the research on evaluation utilisation has provided only "meager support" as to their effectiveness, and mainly in relation to conceptual use of evaluation (Cousins and Leithwood, 1986; Shulha and Cousins, 1997; Nevo, 2001).

61. Hofman and colleagues (2009) conducted a large-scale survey of Dutch teachers to learn whether different approaches to SSE are associated with different levels of student achievement. They drew upon data from a questionnaire sent to 939 primary schools (those schools responding to the researchers' query – from among 1 914 randomly selected schools in the initial sample), school level data from the Dutch National Inspectorate, and student data from a large-scale national study, which brought together school and student data. Overlapping data from the two datasets covered 81 primary schools and 2 099 students. Based on this research, Hofman and colleagues found a significant relationship between the quality of teaching and learning processes and SSE. The researchers awarded an "advanced SSE score" to those schools with the highest scores on the teaching and learning scale (*i.e.* curriculum, use of available learning time, pedagogical and didactic performances of teachers, school climate, harmonisation of the educational needs of students, an active and independent role for students, and higher quality of support and guidance for students). The results of the study provide strong support for the importance of collaboration.

62. Hofman and colleagues (2009) note that schools with high SSE scores were also effective learning organisations. Self-evaluation practices within the schools were consistent with Leithwood and Aitken's (1995) definition of the learning organisation as a group of people pursuing common and individual purposes, considering the value of those purposes and modifying them as appropriate, and developing more efficient and effective approaches. These findings are supported by research showing that teachers are better able to adapt teaching to the needs of their students when they share information about instruction methods and student learning (Little, 1990; Newmann and Wehlage, 1995; McLaughlin and Talbert, 2001).

63. It is not clear, however, whether results from these studies are relevant to school self-evaluation in other country contexts. Several of the countries participating in TALIS require schools to develop regular self-evaluation reports (*i.e.* Australia, Hungary, Iceland, Korea, Mexico, Portugal and Turkey). At the same time, as has been noted, three-quarters of the teachers responding to TALIS reported that there were no incentives to participate actively in school improvement efforts.

### 4.5 External inspection and school improvement

64. To some extent, external inspections may serve to strengthen both technical and social alignment across central and/or regional governments and schools. Alignment is technical in the sense that schools align their programmes to the criteria by which they are evaluated. It is social in the sense that inspectors may share observations from their own experiences with a broad set of schools, and provide direct feedback on how schools may improve. Inspectors may also encourage greater collaboration within schools. However, the link between central inspectorates and schools is more distant than with districts, and inspections are relatively infrequent (typically tri-annually).

65. Evidence regarding the impact of school inspection on school performance is also relatively sparse – and, as Grubb (2000) observes, the question as to whether inspection can improve the quality of education is complex. For example, participants' positive and negative perceptions of the inspection process colour perceptions regarding its effectiveness. Moreover, different systems take very different approaches to inspection. The composition of the team will vary (*e.g.* the balance of education *vs.* non-education professionals), the period of time over which it takes place, the balance of institutional views and the views of individual inspectors, the "culture" of inspection, the relative emphasis of regulation versus improvement, and so on, may vary a great deal (Grubb, 2000).

66. De Wolf and Janssens (2007) conducted a review of the empirical evidence on the effects and side effects of school inspection visits and public performance indicators. They note that while inspection visits seem to improve the quality of schools, there is evidence of "window dressing". They note that publication of performance indicators does seem to improve student results, but also conjecture this may be in part to strategic behaviour of schools (*e.g.* reshaping the test pool, "indicator fixation", or fraud).

67. However, a study for the Dutch Central Planning Agency (Luginbuhl *et al.*, 2007) found that in the first two years following an inspection, student performance improved by 2 to 3% of a test score's standard deviation. Gains were strongest for primary school student performance in the area of mathematics, with gains persisting four years following an inspection. They found that more intensive inspections produced larger improvements in school performance. It is not clear, however, whether the more intensive inspection process also had an influence on school level collaboration focused on instructional improvement.

68. Gray and Gardner (1999) surveyed the views of 70 primary and secondary school leaders on school inspections in Northern Ireland. While the approach to inspection in Northern Ireland has certainly evolved over the years since this survey was conducted, a number of findings are of interest. First, a number of school leaders felt that the inspections helped to focus attention and encouraged staff to work cooperatively. At the same time, only a small percentage of school leaders (8%) reported that they would implement recommended changes. Twenty-eight per cent of school leaders reported that they made no changes as a result of the inspections and the remaining 64% planned to make changes related to school planning (30%) or in classroom instruction (22%). However, these were changes initiated by the schools themselves, and not through the inspection process.

69. Several OECD countries support external school inspection and school self-evaluation as complementary approaches to accountability and improvement. External inspectors can help keep school focus on national standards and benchmarks. At the school-level, evaluators bring a local perspective on the unique context that shapes the school and its students. But there are also potential tensions between external and internal evaluation – *e.g.* the degree to which inspectors should also engage in improvement efforts, and whether school self-evaluation can also fulfil accountability functions (Nevo, 2001, 2002; Kyriakides and Campbell, 2004).

70. A number of analyses are aimed at identifying effective strategies for balancing external and internal school evaluation. Nevo (2001) argues that external school inspections and school self-evaluations will be most effective when participants are able to engage in a constructive dialogue. Both external and internal evaluators should be clear about the methods they will use and the data they will gather. Inspectorates will be able to make better use of data gathered through SSE if it corresponds to the framework of school inspection. To this end, they may provide guidelines and exemplars, and may even support training for SSE (Janssens and van Amelsvoort, 2008). At the same time, those engaged in the SSE process may need to increase their evaluation literacy – building technical skills for collecting and analysing information on instruction and school management. With increased skills, they are very likely to

be more open to use of inspection reports and recommendations (Cousins and Leithwood, 1986; Shulha and Cousins, 1997).

### 4.6 Schools and school districts

71. In decentralised education systems, school districts (local education authorities and municipalities) may play a vital role in guiding school improvement. Districts encompass all local stakeholders, including local leaders and administrators, school leaders, teachers, parents and students). Sykes and colleagues (2009) note that districts are nested within larger policy systems, may initiate policies to "coordinate and direct the work of schools within their jurisdiction", and also interpret and implement decisions initiated at the regional and/or national levels.

72. Sykes and colleagues (2009) suggest that policy makers in the United States are beginning to appreciate that districts or school networks may have a comparative advantage over national or regional governments for certain functions, such as instructional leadership, motivating support and capacity-building in schools, and allocating resources. Drawing on Weick's (1976) concept of loose coupling in education systems, they argue that it is possible to change what is tightly and loosely coupled in education systems. School districts and/or teacher networks would then be more tightly linked around tasks focused on improving instruction.

73. Studies from the United States show the impact of districts on school and student performance varies widely. Several studies point to the importance of district leadership in developing strategies for improvement, helping schools to align curriculum to central standards and assessments, and providing support for low-performing schools (see for example, Bitter *et al.*, 2005; Elmore and Burney, 1997; Hill *et al.*, 2000; Newmann *et al.,* 2001). Districts may also draw on school-level assessment and evaluation data to make decisions as to how to allocate discretionary resources in order to address needs. The most successful districts have invested significant resources (human and financial) to develop skills to interpret and act on student performance data (Datnow, 2009).

74. On the other hand, Leithwood and colleagues (1999) reviewed five high-profile standards-based reforms in the United States, and found that with the exception of reforms implemented in Chicago, districts were not able to provide evidence of increases in student achievement. Moreover, improvements in Chicago were apparent only after the seventh year of a ten-year programme. The Leithwood review also found that the reforms had contributed little to the "core technology of schools" (*e.g.* increases in professional development, adaptation to the local context, effective incentives).

75. School districts in different OECD countries have different functions, and may or may not have the capacity to provide instructional leadership. However, international research points to the benefits of inter-school cooperation, even when it is focused primarily on coordination of managerial and administrative tasks. Pont and colleagues (2008a) argue that this type of coordination frees school leaders to concentrate on instructional leadership at the school level. Alternatively, school leaders may also take on leadership roles at the district level (see Boxes 7 and 8).

76. An important caveat for these strategies is that systems promoting school competition as an incentive for improvement and innovation may actually inhibit co-operation at the district level. Given that cross-country analysis of PISA data shows that school competition does not lead to higher levels of student achievement across systems (OECD, 2010), it may be important for national policy makers to reconsider how to re-balance school competition and co-operation.

**Box 7. Distributed leadership in Finland**

In Finland, a municipality proposed a school leadership reform in which it allocated some school leaders to district-wide coordination responsibilities on a part time basis. The overall strategy was to share acting principals at the municipal level: five school principals were working as district principals, with a third of their time devoted to the district and the rest to their individual schools.

- Leadership is redistributed between the municipal authority and the schools. Beyond leading their own schools, they now coordinate various district level functions such as planning, development or evaluation. In this way, the municipality shares some leadership functions with them that move beyond the boundaries of their own school unit.
- The new district heads are part of a municipal leadership team. Instead of managing alone, the head of the municipal education department now works in a group, sharing problems and elaborating solutions cooperatively.
- District heads now distribute their leadership energies, experiences and knowledge between their own schools and others. While coordinating activities like curriculum planning, professional development or special needs provision in their area, they exercise leadership at both the institutional and local district levels.
- Leadership within the largest schools (which are also led by the district heads) has been redistributed internally between the principal and other staff in the school. This releases the principal for the area-based responsibilities and also develops increased leadership experience and capacity within the schools.

In this new web of horizontal and vertical interdependence, new behaviours emerge. Principals start to consider and address broader community needs rather than fiercely and competitively defending the interests of their own organisation. This interaction across schools opens new windows for mutual learning. In addition, as they devote less time and energy to their own school, they are obliged to delegate various management tasks to other staff, which leads to more open lateral leadership within the school, stronger development of distributed leadership capacity and a more constructive approach to leadership succession and sustainability.

*Source*: Hargreaves *et al.* in Pont *et al.*,2008b.

**Box 8. System leadership in England**

In England, various ways for schools to collaborate have developed recently with the view that collaboration can contribute to make "every school a great school". Under the concept of system leadership, system leaders are those principals willing to contribute and care about and work for the success of other schools and communities as well as their own. Different approaches have been promoted to this end:

- Developing and leading a successful educational improvement partnership between several schools, often focused on a set of specific themes that have significant and clear outcomes reaching beyond the capacity of any one single institution. These include partnerships on curriculum design and specialisms, including sharing curricular innovation. While many partnerships remain at a collaboration level, some have moved to "harder" more formalised arrangements in the form of (con)federations (to develop stronger mechanisms for joint governance and accountability) or Education Improvement Partnerships (to formalise the devolution of certain defined delivery responsibilities and resources from their local authority).

- Acting as a community leader to broker and shape partnerships and/or networks of wider relationships across local communities to support children's welfare and potential, often through multi-agency work. Such system leadership responds to, as Osbourne (2000) puts it, "the acceptance [that] some ... issues are so complex and interconnected that they require the energy of a number of organisations to resolve and hence can only be tackled through organisations working together (p.1). ... The concept of [a] full-service school where a range of public and private sector services is located at or near the school is one manifestation" (p.188).

- Working as a change agent or expert leader within the system, identifying best classroom practice and transferring it to support improvement in other schools. This is the widest category and includes: heads working as mentor leaders within networks of schools, combining an aspiration and motivation for other schools to improve with the practical knowledge and guidance for them to do so; heads who are active and effective leaders within more centrally organised system leadership programmes, for instance within the Consultant Leader Programme, School Improvement Partners (SIP) and National Leaders of Education (NLE); and heads who with their staff purposely develop exemplary curricula and teaching programmes either for particular groups of students or to develop specific learning outcomes in a form that is transferable to other schools and settings.

*Source*: Hopkins in Pont *et al.*,2008b.

## SECTION 5. CREATING EFFECTIVE FRAMEWORKS FOR STANDARDS-BASED ASSESSMENT AND EVALUATION: FINDING BALANCE AND COHERENCE

77. The best approach to achieving alignment in standards-based systems may involve abandoning linear definitions altogether (Baker, 2004). A different approach would be to view it as the balancing of components across both the technical and social dimensions. Within the technical dimension, systems designers ensure that overall frameworks for standards, curriculum, assessment and evaluation are comprehensive and coherent. Policy makers and practitioners have appropriate data and in the right time frame to meet their decision-making needs. The definition of measurement constructs is consistent across levels and over time (Herman and Baker, 2009). Within the social dimension, systems provide incentives for district and school level interactions, and support for learning and adaptation. Sophisticated feedback systems provide data on the impact of instructional interventions.

78. As noted at the beginning of this report, there are persistent tensions between external and internal accountability. Systems that emphasise the technical dimensions place a stronger focus on external accountability. Systems with high stakes fall within this category. Systems that emphasise the social dimensions of change are more concerned with internal accountability.

79. It will be important for systems to develop further both the technical and social dimensions of standards-based systems. Sophisticated feedback systems, flexible interactions among institutions and actors, and strong capacities for learning all enhance the potential for systems to improve and innovate (Hargreaves, 2003). But every country will need to address a number of challenges, including:

- Identifying incentives that motivate and strengthen interaction and build professionalism within schools and across districts;
- Developing teachers' skills to evaluate the impact of different instructional approaches with different students;
- Continuing research and development on high quality measurement systems aligned to goals for higher order learning.

80. Stronger teacher professionalism also points to a stronger role for teachers in the development of standards and of assessment and evaluation systems. Based on their review of literature on accountability and classroom instruction, Ballard and Bates (2008) underscore the importance of communication among teachers and those who write standards, develop large-scale assessments, and set out guidelines for external and internal school evaluations. National and regional governments, universities and practitioner networks may also study and identify effective practices, support district–to-district learning and provide guidance and support in schools adopting and adapting new practices (Sykes *et al.*, 2009).

81. System level learning is also vital, and requires strategic investments in research and development. This includes research focused on the development of alternative measurement technologies. It may also include research on how feedback from standards-based systems is mediated in local contexts. This is an area ripe for international exchange and learning.

## SECTION 6. IN CONCLUSION: GENERAL POLICY PRINCIPLES FOR IMPROVING ALIGNMENT IN STANDARDS-BASED SYSTEMS

82. The following seven principles are based on the discussion and evidence presented above, and suggested as ways to improve both technical and social alignment in standards-based systems:

1) Clearly define the purposes of new frameworks for teaching, learning and assessment and evaluation, and the kinds of supports and incentives that will help teachers to create new professional knowledge.

2) Ensure that standards are grounded in evidence of how students learn and progress within and across different subject domains, and represent realistic goals for attainment.

3) Identify and implement incentives that support teachers' individual and collective motivations.

4) Invest in research and development to strengthen the range of measurement technologies available to assess students' higher order skills, such as problem solving, reasoning and communication.

5) Create coherent assessment and evaluation systems, with measurements at each level of the system fit for purpose.

6) Evaluate the impact of standards-based assessment and evaluations – including both intended and unintended impacts – on the quality of teaching and learning, and adapt systems based on findings.

7) Advocate for significant investments in ongoing research and development of standards-based approaches. Also note that systems that are not well aligned and do not provide high quality information waste significant resources.

## REFERENCES

Abelmann, C., R.F. Elmore, J. Even, S. Kenyon and J. Marshall (1999), *When Accountability Knocks, Will Anyone Answer?*, Consortium for Policy Research in Education, University of Pennsylvania, Philadelphia.

Abu-Alhija, F.N. (2007), "Large Scale Testing: Benefits and Pitfalls", *Studies in Educational Evaluation*, Vol. 33, pp. 50-68.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999), *Standards for Educational and Psychological Testing*, AERA, Washington, D.C.

Baker, E.L. (2004), *Aligning Curriculum, Standards, and Assessments: Fulfilling the Promise of School Reform*, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.

Ballard, K. and A. Bates (2008), "Making a Connection between Student Achievement, Teacher Accountability, and Quality Classroom Instruction", *The Qualitative Report*, Vol. 13, No. 4, pp. 560–580.

Berends, M. (2009), "Commentary: Nested Actors and Institutions: The Need for Better Theory, Data, and Methods to Inform Education Policy", in G. Sykes, B. Schneider and D.N. Plank (eds.), *Handbook of Education Policy Research*, AERA, London and New York, pp. 848-854.

Bhola, D.S., J.C. Impara and C.W. Buckendahl (2005), "Aligning Tests with States' Content Standards: Methods and Issues*", Educational Measurement: Issues and Practice*, Vol. 22, No. 3, pp. 21–29.

Bitter, C. *et al.* (2005), *Evaluation Study of the Immediate Intervention/Underperforming Schools Program of the Public Schools Accountability Act of 1999*, American Institutes for Research, Palo Alto, CA.

Black, P. (2000), "Research and Development of Educational Assessment", *Oxford Review of Education*, Vol. 26, No.s 3 and 4, pp. 407-419.

Black, P. and D. William (1998), "Inside the Black Box: Raising Standards through Classroom Assessment", *Phi Delta Kappan*, Vol. 80, No. 2.

Blok, H., P. Sleegers and S. Karsten (2007), "Looking for a Balance between Internal and External Evaluation of School Quality: Evaluation of the SVI Model", *Journal of Education Policy*, Vol. 23, No. 4, pp. 379–395.

Bloom, B.S. (1968), "Learning for Mastery", *Evaluation Comment*, Vol. 1, No. 2, pp. 1–12.

Bransford, J.D., A.L. Brown and R.R. Cocking (1999), *How People Learn: Brain, Mind, Experience, and School*, National Academy Press, Washington, D.C.

Caldwell, C., C.G. Thorton and L.M. Gruys (2003), "Ten Classic Assessment Center Errors: Challenges to Selection Validity", *Public Personnel Management*, Vol. 32, pp. 73–88.

Carroll, J.B. (1984), "The Model of School Learning: Progress of an Idea", in C. Fisher and D.C. Berliner (eds.), *Perspective on Instructional Time*, Longman, New York, pp. 29-58.

Chudowsky, N. and J.W. Pellegrino (2003), "Large-Scale Assessments that Support Learning: What Will it Take?", *Theory into Practice*, Vol. 42, pp. 75-83.

Cizek, G.J. (1998), *Testing in American Schools: Getting the Right Answers*, Lawrence Erlbaum Associates, Mahwah, New Jersey.

Cizek, G.J. (2001), "Conjectures on the Rise and Call of Standard-Setting: An Introduction to Context and Practice", in G.J. Cizek (ed.), *Setting Performance Standards: Concepts, Methods and Perspectives*, Erlbaum, Mahwah, New Jersey, pp. 3-17.

Cizek, G.J. and M.B. Bunch (2007), *Standard-Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*, Sage Publications Inc., California.

Commission on Instructionally Supportive Assessment (2001), *Building Tests to Support Instruction and Accountability*, *www.nea.org/issues/high-stakes/buildingtests.html*.

Cousins, J.B. and K.A. Leithwood (1986), "Current Empirical Research on Evaluation Utilization", *Review of Educational Research*, Vol. 56, No. 3, pp. 331-364.

Crocker, L. (2005), "Teaching FOR the Test: How and Why Test Preparation is Appropriate" in R.P. Phelps (ed.), *Defending Standardized Testing*, Lawrence Erlbaum, Mahwah, NJ, pp. 159-174.

Datnow, A. and V. Park (2009), "Large-Scale Reform in an Era of Complexity" in G. Sykes, B. Schneider and D.N. Plank (eds.), *Handbook of Education Policy Research*, AERA, London and New York, pp. 348-361.

Elmore, R. and D. Burney (1997), *Investing in Teacher Learning: Staff Development and Instructional Improvement in Community School District #2*, National Commission on Teaching & America's Future & the Consortium for Policy Research in Education, New York.

Elmore, R.F. (2001) "Psychiatrists and Light Bulbs: Educational Accountability and the Problem of Capacity", paper presented at the annual meeting of the American Educational Research Association, Seattle, April.

Elmore, R.F. and S.H. Fuhrman (2001), "Holding Schools Accountable: Is It Working?", *Phi Delta Kappan*, Vol. 83, No. 1, pp. 67-70, 72.

Eurydice (2008), Eurybase, *http://eacea.ec.europa.eu/education/eurydice/eurybase_en.php*.

Fidler, B., P. Earley and J. Ouston (eds.) (1996), *Improvement through Inspection? Complementary Approaches to School Development*, David Fulton, London.

Finn, C.E. and M. Kanstoroom (2001), "State Academic Standards", *Brookings Papers on Education Policy*, No. 4, pp. 131-164.

Finnigan, K. and B. Gross (2007), "Do Accountability Sanctions Influence Teacher Motivation? Lessons from Chicago's Low-Performing Schools", *American Educational Research Journal*, Vol. 44, No. 3, pp. 594-629.

Firestone, W.A. and D. Mayrowetz (2000), "Rethinking 'High Stakes', Lessons from the United States and England and Wales", *Teachers College Record,* Vol. 102, pp. 724–749.

Firestone, W.A., D. Mayrowetz and J. Fairman (1998), "Performance-based Assessment and Instructional Change: The Effects of Testing in Maine and Maryland", *Educational Evaluation and Policy Analysis*, Vol. 20, No. 2, pp. 95–113.

Gamoran, A. *et al*. (1997), "Upgrading High School Mathematics Instruction: Improving Learning Opportunities for Low-Achieving, Low-Income Youth", *Educational Evaluation and Policy Analysis*, Vol. 19, pp. 325–338.

Goertz, M.E. and D. Massell (2005), "Summary", in B. Gross and M. E. Goertz (eds.), *Holding High Hopes: How High Schools Respond to State Accountability Policies*, University of Pennsylvania, Consortium for Policy Research in Education, Philadelphia.

Gordon, J., G. Halasz, M. Krawczyk, T. Leney, A. Michel, D. Pepper, E. Putkiewicz and J. Wisniewski (2009), *Key Competences in Europe: Opening Doors for Lifelong Learners across the School Curriculum and Teacher Education* (2009), Study undertaken for the Directorate General Education and Culture of the European Commission, CASE-Center for Social and Economic Research, Warsaw, *http://ec.europa.eu/education/more-information/moreinformation139_ en.htm*.

Gray, C. and J. Gardner (1999), "The Impact of School Inspections", *Oxford Review of Education*, Vol. 25, No. 4, pp. 455–468.

Grubb, N. (2000), "Opening Classrooms and Improving Teaching: Lessons in School Inspections in England", *Teachers College Record*, Vol. 102, No. 4, pp. 696–723.

Hargreaves, A., G. Halász and B. Pont (2008), "The Finnish Approach to System Leadership", a case study report for the OECD Improving School Leadership activity, in B. Pont, D. Nusche and D. Hopkins (eds.), *Improving School Leadership, Volume 2: Case Studies on System Leadership*, OECD, Paris, *www.oecd.org/edu/schoolleadership.*

Hargreaves, D.H. (2003), *Education Epidemic: Transforming Schools through Innovation Networks*, Demos, London.

Herman, J.L. (2005), *Making Accountability Work to Improve Student Learning*, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.

Herman, J.L. and E.L. Baker (2009), "Assessment Policy: Making Sense of the Babel", in G. Sykes, B. Schneider and D.N. Plank (eds.), *Handbook of Education Policy Research*, AERA, London and New York, pp. 348-361.

Hill, P., C. Campbell and J. Harvey (2000), *It Takes a City: Getting Serious about Urban School Reform*, The Brookings Institution, Washington, D.C.

Hofman, R.H., N.J. Kikstra and W.H.A. Hofman (2009), "School Self-Evaluation and Student Achievement", *School Effectiveness and School Improvement*, Vol. 20, No. 1, pp. 47–68.

Hopkins, D. (2008), "Realising the Potential of System Leadership", a case study report for the OECD Improving School Leadership activity, in B. Pont, D. Nusche and D. Hopkins (eds.), *Improving School Leadership, Volume 2: Case Studies on System Leadership*, OECD, Paris, *www.oecd.org/edu/schoolleadership.*

Jacob, B.A., P.N. Courant and J. Ludwig (2003), "Getting Inside Accountability: Lessons from Chicago", *Brookings-Wharton Papers on Urban Affairs*, Brookings Institution Press, pp. 41–81.

Janssens, F.J.G. and G.H.W.C.H van Amelsvoort (2008), "School Self-Evaluations and School Inspections in Europe: An Exploratory Study", *Studies in Educational Evaluation*, Vol. 34, pp. 15–23.

Koretz, D. (2005), "Alignment, High Stakes, and the Inflation of Test Scores", University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.

Koretz, D., D.F. McCaffrey and L.S. Hamilton (2001), *Toward a Framework for Validating Gains under High-Stakes Conditions*, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.

Kyriakides, L. and R.J. Campbell (2004), "School Self-Evaluation and School Improvement: A Critique of Values and Procedures", *Studies in Educational Evaluation*, Vol. 30, pp. 23–36.

Leithwood, K. and R. Aitken (1995), *Making Schools Smarter: A System for Monitoring School and District Progress*, Corwin, Newbury Park, CA.

Leithwood, K., D. Jantzi and B. Mascall (1999), "Large Scale Reform: What Works?", submitted as part of the *External Evaluation of the UK National Literacy and Numeracy Strateg*y, Institute for Studies in Education, Toronto.

Linn, R. (1998), *Assessments and Accountability*, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.

Linn, R.L. (2003), "Accountability: Responsibility and Reasonable Expectations", *Educational Researcher*, Vol. 32, No. 7, pp. 3-13.

Linn, R.L. (2005), *Issues in the Design of Accountability Systems*, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.

Little, J.W. (1990), "The Persistence of Privacy: Autonomy and Initiative in Teachers' Professional Relations", *Teachers College Record,* Vol. 91, pp. 509–536.

Luginbuhl, R., D. Webbink and I. de Wolf (2007), *Measuring the Effect of School Inspections on Primary School Performance: A Study Based on CITO Test Scores*, CPB discussion paper, CPB, Den Haag, The Netherlands.

McDonnell, L.M. and C. Choisser (1997), *Testing and Teaching: Local Implementation of New State Assessments*, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.

McLaughlin, M.W. and J.D. Talbert (2001), *Professional Communities and the Work of High School Teaching,* University of Chicago Press, Chicago.

Mehrens, W.A. (1998), "Consequences of Assessment: What is the Evidence?", *Educational Policy Analysis Archives*, Vol. 6, No. 13.

Mislevy, R.J. *et al.* (1998), *A Cognitive Task Analysis, with Implications for Designing A Simulation-Based Performance Assessment*, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.

Møller, J.A. *et al*. (2005), "Successful School Leadership: The Norwegian Case", *Journal of Educational Administration*, Vol. 43, No 6, pp. 584-584.

Moos, L., J. Kresjsler and K.K. Kofod (2008), "Successful Principals: Telling or Selling? On the Importance of Context for School Leadership", *International Journal of Leadership in Education*, Vol. 11, No. 4, pp. 341–352.

Murnane, R.J and R.R. Nelson (2005), "Improving the Performance of the Education Sector: The Valuable, Challenging, and Limited Role of Random Assignment Evaluations", *Working Paper 11856*, National Bureau of Economic Research, Cambridge, Massachusetts.

Nevo, D. (2001), "School Evaluation: Internal or External?", *Studies in Educational Evaluation*, Vol. 27, pp. 95-106.

Nevo, D. (2002), *School-based Evaluation: An International Perspective*, JAI, Amsterdam.

Newman, F.M., B. Smith, E. Allensworth and A.S. Bryk (2001), "Instructional Program Coherence: What it is and Why it should Guide School Improvement Policy", *Educational Evaluation and Policy Analysis,* Vol. 23, No. 4, pp. 297-321.

Newmann, F.M. and G.G. Wehlage (1995), *Successful School Restructuring: A Report to the Public and Educators by the Center on Organization and Restructuring of Schools*, Center on Organization and Restructuring of Schools, Madison, Wisconsin.

Nusche, D., G. Halász, J. Looney, P. Santiago and C. Shewbridge (2011), *OECD Reviews of Evaluation and Assessment in Education: Sweden*, OECD, Paris.

O'Day, J. (2002), "Complexity, Accountability and School Improvement"*, Harvard Education Review*, Vol. 72, No. 3, *http://gseweb.harvard.edu/~hepg/oday.html*.

OECD (2005a), *Formative Assessment: Improving Learning in Secondary Classrooms*, OECD, Paris.

OECD (2005b), *Teachers Matter: Attracting, Developing and Retaining Effective Teachers,* OECD, Paris.

OECD (2009a), *Education at a Glance: OECD Indicators*, OECD, Paris.

OECD (2009b), *Creating Effective Learning Environments: First Results from TALIS*, OECD, Paris.

OECD (2010), *PISA 2009 Results: What Students Know and Can Do*, OECD, Paris.

Osbourne, S.P. (2000), *Public/Private Partnerships*, Routledge, London.

Pant, H.A., A.A. Rupp, S.P. Tiffin-Richards and O. Köller (2009), "Validity Issues in Standard-Setting Studies", *Studies in Educational Evaluation*, Vol. 35, pp. 95–101.

Pellegrino, J., N. Chudowsky and R. Glaser (eds.) (2001), *Knowing What Students Know: The Science and Design of Educational Assessments,* National Academy Press, Washington, D.C.

Pellegrino, J.W., G.P. Baxter and R. Glaser (1999), "Addressing the 'Two Disciplines' Problem: Linking Theories of Cognition and Learning with Assessment and Instructional Practice", *Review of Research in Education*, Vol. 24, pp. 307–353.

Pont, B., D. Nusche and H. Moorman (2008a), *Improving School Leadership*: *Vol. 1, Policy and Practice*, OECD, Paris.

Pont, B., D. Nusche and D. Hopkins (eds.) (2008b), *Improving School Leadership, Volume 2: Case Studies on System Leadership*, OECD, Paris.

Popham, W.J. (2002), "Right Task, Wrong Tools", *American School Board Journal*, Vol. 189, No. 2, pp. 18-22.

Porter, A.C. and J.L. Smithson (2001), "Defining, Developing and Using Curriculum Indicators", *CPRE Research Report Series RR-048*, Consortium for Policy Research in Education, Philadelphia.

Schmidt, W.H. and A. Maier (2009), "Opportunity to Learn", in G. Sykes, B. Schneider and D.N. Plank (eds.), *Handbook of Education Policy Research*, AERA, London and New York, pp. 541-559.

Seashore Louis, K., K. Leithwood, K.L. Wahlstrom and S.E. Anderson (2010), *Investigating the Links to Improved Student Learning: Final Report of Research Findings"*, University of Minnesota, St. Paul, *www.cehd.umn.edu/CAREI/*.

Shulha, L.M. and J.B. Cousins (1997), "Evaluation Use: Theory, Research and Practice Since 1986", *Evaluation Practice,* Vol. 18, No. 3, pp. 195-208.

SLO (2007), *Case Studies Basis Education in Europe: Core Affairs, the Netherlands*, National Institute for Curriculum Development (SLO), Enschede.

Smith, M.L. and C. Rottenberg (1991), "Unintended Consequences of External Testing in Elementary Schools", *Educational Measurement: Issues and Practice*, Vol. 10, No. 4, pp. 7-11.

Smithson, J.L. and A.C. Collares (2007), "Alignment as a Predictor of Student Achievement Gains", paper presented at the annual meeting of the American Educational Research Association, Chicago, April.

Spillane, J.P. and J.S. Zeuli (1999), "Reform and Teaching: Exploring Patterns of Practice in the Context of National and State Mathematics Reform", *Educational Evaluation and Policy Analysis*, Vol. 21, pp. 1-27.

Stecher, B.M. (2002), "Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practices", in L.S. Hamilton, B.M. Stecher and S.P. Klein (eds.), *Making Sense of Test-Based Accountability in Education*, RAND, Santa Monica, CA, pp. 79-100.

Sykes, G., J. O'Day and T.G. Ford (2009), "The District Role in Instructional Improvement", in G. Sykes, B. Schneider and D.N. Plank (eds.), *Handbook of Education Policy Research*, AERA, London and New York, pp. 767–784.

UNESCO (2006), World Data on Education database *www.ibe.unesco.org/Countries/WDE/2006/index.html*.

Weick, K. (1976), "Educational Organizations as Loosely Coupled Systems", *Administrative Science Quarterly,* Vol. 21, pp. 1–19.

Wiliam, D. (2006), "Formative Assessment: Getting the Focus Right", *Educational Assessment*, Vol. 11, pp. 283–289.

Winkler, A. (2002), "Division in the Ranks: Standardized Testing Draws Lines between New and Veteran Teachers", *Phi Delta Kappan*, Vol. 84, No. 3, pp. 219–225.

Wolf, I.F. de and F.J.G. Janssens (2007), "Effects and Side Effects of Inspections and Accountability in Education: An Overview of Empirical Studies", *Oxford Review of Education,* Vol. 33, No. 3, pp. 379–396.

Ziezy, M. and M. Perie (2006), *A Primer on Setting Cut Scores on Tests of Educational Achievement*, Educational Testing Service, Princeton, NJ.

## THE OECD EDUCATION WORKING PAPERS SERIES ON LINE

The OECD Education Working Papers Series may be found at:

- The OECD Directorate for Education website: www.oecd.org/edu/workingpapers
- The OECD's online library: www.oecd-ilibrary.org/papers
- The Research Papers in Economics (RePEc) website: www.repec.org

If you wish to be informed about the release of new OECD Education working papers, please:

- Go to www.oecd.org
- Click on "My OECD"
- Sign up and create an account with "My OECD"
- Select "Education" as one of your favourite themes
- Choose "OECD Education Working Papers" as one of the newsletters you would like to receive

For further information on the OECD Education Working Papers Series, please write to: edu.contact@oecd.org.