OECDpublishing

Please cite this paper as:

Rosenkvist, M. A. (2010), "Using Student Test Results for Accountability and Improvement: A Literature Review", *OECD Education Working Papers*, No. 54, OECD Publishing.
http://dx.doi.org/10.1787/5km4htwzbv30-en

OECD Education Working Papers
No. 54

Using Student Test Results for Accountability and Improvement

A LITERATURE REVIEW

Morten A. Rosenkvist

Unclassified **EDU/WKP(2010)17**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **22-Nov-2010**

English - Or. English

DIRECTORATE FOR EDUCATION

EDU/WKP(2010)17
Unclassified

USING STUDENT TEST RESULTS FOR ACCOUNTABILITY AND IMPROVEMENT: A LITERATURE REVIEW

OECD Education Working Paper No. 54

by Morten Anstorp Rosenkvist

This paper was prepared by Morten Anstorp Rosenkvist during a secondment at the Education and Training Policy Division, Directorate for Education, OECD, from the Norwegian Ministry of Education and Research, for the period February-July 2010. The paper forms part of the work undertaken by the OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes and was discussed at the 2nd meeting of the Group of National Experts on Evaluation and Assessment (9-10 September 2010).

Contact: Mr. Paulo Santiago [Tel: +33(0) 1 45 24 84 19; e-mail: paulo.santiago@oecd.org]

JT03292914

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.

ABSTRACT

This report discusses the most relevant issues concerning using student test results in OECD countries. Initially the report provides an overview of how student test results are reported in OECD countries and how stakeholders in these countries use and perceive of the results. The report then reviews the literature relating to using student test results for accountability and improvement purposes. Two general findings can be drawn from the literature: (1) accountability based on student test results can be a powerful tool for changing teacher and school behaviour, but it often creates unintended strategic behaviour, and (2) no test can be a perfect indicator of student performance. Drawing from these findings the report discusses the advantages and disadvantages of using student test results for accountability and improvement. The discussion touches upon four themes: (1) assessment design, (2) the use of test results, (3) stakeholder involvement, and (4) implementation.[1]

RÉSUMÉ

Ce rapport analyse les questions essentielles sur l'utilisation des résultats de tests standardisés aux élèves dans les pays de l'OCDE. Premièrement, le rapport offre un aperçu sur comment les résultats de tests standardisés sont communiqués dans les pays de l'OCDE et comment les différentes parties prenantes utilisent et perçoivent les résultats. Ensuite le rapport fait une révision de la littérature académique sur l'utilisation des résultats de tests standardisés pour l'amélioration et le rendement de comptes. Deux résultats plus généraux émergent de la littérature : (1) Le rendement de comptes basé sur les résultats de tests standardisés peut avoir un fort effet sur le comportement de l'enseignant et de l'école mais peut aussi créer du comportement stratégique non souhaité, et (2) aucun test ne peut être un indicateur parfait des résultats scolaires d'un élève. En se fondant sur ces résultats, le rapport analyse les avantages et inconvénients de l'utilisation des résultats de tests standardisés pour l'amélioration et le rendement de comptes. L'analyse aborde quatre thèmes : (1) la conception de l'évaluation, (2) l'utilisation des résultats des tests, (3) l'implication des parties prenantes, et (4) l'implémentation.

[1] Morten Anstorp Rosenkvist, a Norwegian national, was part of the team working on the OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes while on secondment at the Education and Training Policy Division, Directorate for Education, OECD, from the Norwegian Ministry of Education and Research, for the period February-July 2010. Morten has a Master's degree in political science from the University of Oslo. He has also studied at the University of Sydney. For the last four years Morten has worked as an analyst in the Norwegian Ministry of Education of Research. He is especially familiar with research relating to teachers and teacher training. Morten is currently working as project manager for "GNIST" – a government initiative to recruit more and better qualified teachers to Norwegian schools.

TABLE OF CONTENTS

1. INTRODUCTION

1. Performance in schools is increasingly judged on the basis of effective student learning outcomes. Information is critical to knowing whether the school system is delivering good performance and providing feedback is the channel through which performance can be improved. Increasingly, countries are developing a range of tools and techniques for evaluation and assessment in school systems as part of their efforts to improve student outcomes.

2. Summative assessments measure what students have learned (and not learned) after completing a specific unit in school – typically at the end of a term. Results from such assessments have a number of potential uses. It can be used to monitor the performance of the education system, inform classroom practice, ensure that students have met required educational standards and reward and/or penalise teacher and schools for their students' performance. Some OECD countries incorporate all or most of the examples above, others only a few. While the motive for, and uses of, summative assessments vary among countries, a common denominator is that it is often used for accountability and improvement purposes.

3. Accountability has become a cornerstone of public sector reform in many countries (Levitt *et al*., 2008). A central assumption in accountability is that substantial improvement necessitates that the producers are held accountable for the outcomes they generate (Hopman, 2008). In an accountability context, teachers and schools – who are trusted with the imperative task of teaching and instructing children – are the "producers", while student test results may be used as a proxy for measuring "outcomes". By measuring student outcomes and holding teachers and schools responsible for results, accountability systems intend to create incentives for improved performance and identify ''underperforming'' schools for remediation (Booher-Jennings, 2007) . It is important to note that student test results can be used for improvement purposes without teachers and schools being held accountable for the results. For example, a teacher can use test results to identify student weaknesses and strengths in order to improve classroom instruction.

4. The focal point of this report is how student test results are used in OECD countries. More specifically, this report aims to answer three broad questions:

- How are summative assessments organised and reported in OECD countries, who are the users of student test results and how do the users use and perceive of the results?

- What is the empirical evidence on the effects of using student test results for accountability and improvement?

- What are the advantages and disadvantages of different approaches to using student test results for accountability and improvement?

5. It is important to note that this report does not provide a general overview of how OECD countries conduct evaluation and assessment. There are related themes such as formative student assessment, accountability for local authorities, student tracking and system level evaluation which is not dealt with in this report.

6. A number of government and government-sponsored websites were surveyed in order to identify country practices and policies. Empirical evidence has been identified through a broad search in the literature using databases and search engines such as *ScienceDirect, Jstore, Google Scholar, ERIC* and *SpringerLink*. The search has been conducted in English, the Scandinavian languages, and to a lesser degree German and Spanish.

7. Chapter 2 defines the terms and concepts used in this report. In Chapter 3, an overview is given on how countries organise and report student test results. Chapter 4 reports on the different users and uses of student test results. Chapter 5 reviews the empirical evidence on the effects of using student test results for accountability and improvement purposes. Chapter 6 offers, based on the findings and data presented in the preceding chapters, a discussion of the advantages and disadvantages of using student test results for accountability and improvement. A final conclusion is offered in Chapter 7.

2. TERMS AND CONCEPTS

8. The terms *assessment* and *evaluation* are often used interchangeably. However, education specialists often make careful distinctions between the two terms in order to clarify different roles. This report follows these distinctions. Assessment is used to refer to the process of deciding, collecting and making judgments about evidence relating to students' achievement of particular goals of learning. Evaluation is used for the process of deciding, collecting and making judgments about systems, programs, materials, procedures and processes (Harlen, 2007). Consequently, assessment encompasses classroom-based assessments as well as large-scale external tests and examinations, while evaluation encompasses school inspections, school self-evaluations and targeted programme evaluation. This report focuses on *student assessment.*

9. Student assessment is a broad term that encompasses several methods and techniques for measuring what students have learned (and not learned). The focal point of this report is the use of student *test* results. A test is an assessment, often administered on paper or on the computer, intended to measure students' knowledge, skills and/or aptitudes. For the purposes of this report, student assessment is understood in the narrow meaning of *student testing*.

10. The literature distinguishes between *formative* and *summative* assessment (EPPI, 2002; OECD, 2005a; Harlen, 2007). Summative assessment is used to measure what students have learnt at the end of a unit, to promote students, to ensure they have met required standards on the way to earning certification for school completion or to enter certain occupations, or as a method for selecting students for entry into further education. Ministries of education may use summative assessment as a way to hold teachers and schools accountable for providing quality education. But assessments may also serve a formative function. In classrooms, formative assessment refers to frequent, interactive assessments of student progress and understanding to identify learning needs and adjust teaching appropriately (OECD, 2005a). This report concentrates on summative student assessment. Formative Assessment is thoroughly dealt with in Looney (forthcoming).

11. Summative assessment can be categorised according to whether or not the assessment is standardised. A standardised assessment is a test designed in such a way that the questions, conditions for administering, scoring procedures, and interpretations are consistent (Popham, 1991). The goal of this consistency is to make the results as objective as possible so that they can be considered valid and meaningful when used to compare the assessed qualities of students (Zucker, 2004). Standardised assessments are usually administered to large groups of students for the purpose of measuring academic achievement and/or comparing members of a cohort. National examinations (*i.e.* assessments that have a civil effect) may or may not be standardised. This report focuses on standardised assessment; assessments whose results are, *comparable among students, regardless of the school they attend.* In order not to exclude too many countries from the descriptive analysis in Chapter 3 and 4, this report employs a broad conception of standardised assessment. The case of Sweden is illustrative. Here standardised assessments are given to all students in certain grades, but the assessments are graded by the students' own teachers. This does not satisfy a stringent definition of standardised assessment. Sweden is nevertheless included in the descriptive analysis in order to provide a broader scope of reference.

12. For matters of simplicity, summative standardised student assessment is in this report referred to as *student assessment*. Assessment can be used for testing the entire student population or a representative

sample of students at one or several stages in the education process. In this report the former is referred to as *national student assessment*, while the latter is referred to as *sample student assessment.*

13. There are different types of assessment. *Norm-referenced* assessments describe what students can do relative to other students, while *criterion-referenced* assessments compare student accomplishment to pre-established achievement standards, rather than to the achievement of other students (Popham, 2003; ECS, 2002). Both norm-referenced and criterion-referenced assessments can be used for measuring value-added student performance. Value-added assessment is a method used to measure the effectiveness of a school and its teachers using data on individual students' academic growth over time (Vaishnav, 2005). It is the act of comparing students' scores with their own past scores that distinguishes value-added assessment. A methodological more advanced version is contextual value-added assessment, which takes into account contextual factors such as students' socioeconomic background.

14. As earlier stated, countries usually employ student assessment as a means for accountability. Accountability is essentially a relationship in which a "principal" holds an "agent" responsible for certain kinds of performance. The agent is expected to provide an "account" to the principal. This account describes the performance for which that agent is held responsible (Jacob and Kirst, 1999). What is meant by accountability in education depends on the agent that is held accountable. For example, Ladd (2007) refers to school accountability as "systems that use measures of student outcomes – primarily student achievement as measured by test scores – to hold schools accountable for improving the performance of their students".

15. Accountability systems depend largely on the culture of social systems in which they operate (Bracci, 2009). The choices of accountability tools (*e.g. which* agents are held accountable and *how* they are held accountable) – and the balance among different forms of accountability – are constantly shifting as problems emerge, as social goals change, and as new circumstances arise (Darling-Hammond, 2004). An agent can be accountable to several principals (*e.g.* teachers being accountable to the local authorities, the school leader and their students' parents). Furthermore, a principal can also be an agent – and vice versa (*e.g.* local authorities can be the principal in its relationship with school leaders, and at the same time the agent in its relationship with central authorities).

16. An aspect concerning accountability is the consequences placed on the outcome. It is common in the literature to distinguish between high and low stake assessment (Klein *et al.,* 2000; Carnoy *et al.,* 2003; Jacob, 2005). High stake implies that substantial advantages and/or disadvantages are coupled with the test results, while low stake implies that none or few such couplings exist. Examples of high stake assessments are examinations with a civil effect for students, pecuniary rewards and sanctions – based on student test results – for teachers, and public disclosure of low performing schools. Examples of low stake assessments are periodic national assessments with no civil effect for students and school leader informal approval or disapproval of teacher performance related to student test results. A single assessment may have different stakes for different stakeholders (*e.g.* high stakes for students and low stakes for teachers). This report concentrates on stakes for schools and teachers. Student stakes are comprehensively dealt with in Nusche (forthcoming).

17. Moreover, a conceptual distinction can be made between (at least) two different accountability models: external accountability (also referred to as bureaucratic or hierarchical accountability) and internal accountability (also referred to as professional accountability) (Adams and Kirst, 1999; Firestone, 2002; O'Day, 2002; Garmannslund *et al.,* 2008; Levitt *et al.,* 2008). The external accountability model is a top-down model where schools are understood as an instrument for education policy on the national, regional and local level. A key feature in this model is to provide information to policy makers and the public about value for money, compliance with standards and regulation and quality of the services provided. Schools and teachers are held accountable for the quality of the education they provide – measured as student test

results and/or other quality indicators. Formal authority alone may be used to enforce compliance in the external accountability model, but that authority can be reinforced with performance incentives such as financial rewards and/or sanctions.

18. The internal accountability model, on the other hand, is rooted in the assumption that teaching is too complex an activity to be governed by top-down defined provisions. Effective teaching rests on professionals acquiring specialised knowledge and skills and being able to apply such knowledge and skills to the specific contexts in which they work. In this model schools and teachers are held accountable for how they conduct their profession – *i.e.* their interaction with colleagues and students – and not their students' test results. A reference is often made to medicine and law, specialised professions where the practitioners first and foremost are accountable to the professional standards of the occupation. Compliance in the internal accountability model may be enforced through holding teachers accountable to high professional standards / codes of conduct (as set by a professional association and/or the government) and peer reviews.

19. Due to lack of evidence relating to internal accountability, this report will deal only with *external accountability*.

3. OVERVIEW OF COUNTRY PRACTICES

3.1. Organisation of student assessments

20. A number of OECD countries conduct student assessments. At first glance, the make-up of these assessments may appear somewhat identical, but in fact there are often substantial differences between countries in design, implementation and use. These differences arise from the fact that assessment is a political phenomenon (as well as a technical one), reflecting the agenda, tensions, institutional norms, and nature of power relations between political actors (Kellaghan *et al.,* 2009).

21. National student assessments are administrated in Australia, Belgium (French Community), Denmark, France, Hungary, Iceland, Ireland, Japan, Luxembourg, Mexico, the Netherlands, Norway, Portugal, the Slovak Republic, Sweden and the United Kingdom (England).

22. In the United Kingdom assessments are held at the end of Key Stage 1 (Years 1-2), 2 (Years 3-6) and 4 (Years 10-11). At the end of Key Stage 1, teachers assess student progress in English and math (measured by tasks and tests that are administered informally). At the end of Key Stages 2, students take national tests in English, math and science. There is no national test at the end of Key Stage 3. At the end of Key Stage 4 students sit exams for the General Certificate of Secondary Education (GCSE) and/or equivalent qualifications. In Australia the National Assessment Program – Literacy and Numeracy (NAPLAN) assesses students in Years 3, 5, 7 and 9. Since 2006, Irish primary schools are required to administer standardised tests in literacy and numeracy to pupils at two points of the primary school cycle (Years 1/2 and 4/5).

23. In the Netherlands, schools must be able to account for their results. The great majority of schools do this through the use of student monitoring systems. The most common is a system developed by the National Institute for Educational Measurement (CITO), which comprises an integrated series of tests with a psychometric basis that allow students' progress to be measured in core subjects. Moreover, the Dutch National Examination Board (CEVO) conducts national student assessments in several subjects. Since 2007, Japan has conducted national student assessments among elementary school sixth-graders and third-year junior high school students. The tests are voluntary for most schools, but a large majority of schools participate.

24. In Belgium (the French Community), the Department of General Affairs, Research on Education and Joint Steering of the Education System is responsible for assessments of students' achievements at the start of Years 3 and 5. As early as 1989, national standardised assessments were introduced in France. National student assessments are administered at the beginning of the school year. Portugal conducts national assessments in Years 4, 6 and 9.

25. The Swedish National Agency for Education yearly conducts national assessments in Years 3, 5 and 9, while the Norwegian Directorate of Education and Training conducts yearly national assessments in Years 5 and 8. Beginning in 2010, the Danish Agency for the Evaluation and Quality of Primary and Lower Secondary Education will yearly conduct national assessments in Years 2, 3, 4, 6 and 8. In Iceland national assessments are conducted in Years 4, 7 and 10. The assessments are organised, composed and marked by The Educational Testing Institute.

26. In federal states national student assessments are often administrated at the state level. This is the case in Canada, Germany and the United States.

27. Under the No Child Left Behind (NCLB) Act in the US, states must measure student progress in reading and mathematics Years 3 through 8 and at least once during Years 10 through 12. Tests of science achievement were added in 2008. Tests must be aligned with state academic content and achievement standards. For some years, the German Länder have been conducting Land-specific as well as Länder-spanning comparative studies, in addition to national and international performance comparisons. This includes for example measurement of language proficiency for different age groups, surveys on learning levels, or comparative studies in different grades or Land-specific performance comparisons.

28. Sample student assessments are administrated in Australia, Austria, Belgium (Flemish Community), Canada, Finland, Germany, Ireland, Italy, Korea, the Netherlands, New Zealand, Spain, the United Kingdom (Scotland) and the United States.

29. In Austria, baseline tests (sample tests based on educational standards) in core subjects were conducted in 2009 and 2010. The Austrian agency BIFIE will conduct nationwide follow-up tests every 3rd year for students in Years 4 and 8. The National Education Monitoring Project (NEMP) in New Zealand aims to obtain a broad picture of the achievement and other educational outcomes of representative samples of students in Years 4 and 8. Each year, over a four-year period, different areas related to the curriculum are assessed. In the 2007/08 school year, the Italian National Institute for the Evaluation of the Education System (INVALSI) introduced national sample assessments in Italian, mathematics and science. The tests take place in Years 2 and 5 of primary school and in Years 1 and 3 of lower secondary school.

30. In Spain, the national Institute of Evaluation (IE) and the corresponding bodies in the Autonomous Communities collaborate in carrying out national assessments of samples of pupils. Other tests covering all students are conducted on the sole responsibility of each Autonomous Community. The Finnish National Board of Education conducts representative sample studies of students learning achievements in different school types and levels. In Scotland, the Scottish Survey of Achievement (SSA) is a sample survey that monitors how well pupils in Scotland are learning. Each year the SSA focuses on a different aspect of the school curriculum. Australia commenced in 2003 a rolling three-yearly cycle of student sample assessments in Year 6 and 10. The sample assessments are designed primarily to monitor national and jurisdictional progress.

31. The National Assessment of Educational Progress (NAEP) is the only nationally representative and continuing assessment of what US students know and can do in various subject areas. The assessment stays essentially the same from year to year. This permits NAEP to provide a clear picture of student academic progress over time. NAEP scores are reported on the state and national level, as well as for selected large urban districts. NAEP results are based on representative samples of students. In Canada the Pan-Canadian Assessment Program (PCAP) is an initiative of the Council of Ministers of Education that complements the other assessments in each province and territory. Reading, mathematics and science tests were administered in the spring of 2007 to randomly selected students (13-year-old) in a random sample of schools with a random assignment of test booklets.

32. In Belgium (the Flemish Community) inter-school tests are organised each year (municipal or inter-diocesan) for certain groups of subjects, such as the mother language and arithmetic in the final year of primary school. The tests are voluntary. The Irish Educational Research Centre conducts national sample surveys of achievement at primary level in English (NAER) and Mathematics (NAMA). It is planned to extend the range of assessments to include other areas of the curriculum in future years. Korea monitors the quality of the education system through the National Assessment of Scholastic Achievement

(NASA). NASA measures student performance against the objectives outlined in the school curriculum. NASA samples 1% of all participants based on school level and region.

3.2. Reporting student test results

33. It is a widely debated question in many countries to what extent and how student test results should be made publicly available. Some contend that there should be an effort towards making public all evidence from the evaluation of public policy (with appropriate analyses) in order to provide evidence to taxpayers and the users of schools on whether the schools are delivering the expected results, to provide a basis for intervening across the systems where results in priority areas are unsatisfactory, to enhance trust in government, or to improve the quality of policy debate. Others consider that the publication of school performance data will be counterproductive as it is subject to erroneous interpretation, particularly when no adjustment for socioeconomic background is made. Also debated is what types of reporting have proven most effective, in terms of raising performance and engaging teachers and schools in school improvement and to what extent the information schools and parents receive goes beyond the performance of their own school (OECD, 2007).

34. Countries consistently publish aggregated assessment (national and sample) results for the education system as a whole, often in a yearly report on the state of the system and/or on an official website. This is the case in Austria, Australia, Belgium (French Community and Flemish Community), Canada, Denmark, Finland, France, Germany, Hungary, Iceland, Ireland, Italy, Japan, Korea, Mexico, the Netherlands, New Zealand, Norway, Portugal, the Slovak Republic, Spain, Sweden, the United Kingdom (England and Scotland) and the United States.

35. Countries less consistently publish aggregated national and/or sample[2] student test results on the local and school level. This is the case in Australia, Canada, France, Iceland, Mexico, the Netherlands, Portugal, Sweden, the United Kingdom (England) and the United States. In Norway results from the national student assessment are published on the local and regional level, while in Germany and Spain results are only published on the regional/state level.

36. In some countries, official documents state clearly that national tests cannot be used to rank schools. This applies to Austria, Belgium (the French Community), Denmark, France (in the case of *évaluations-bilans)* and Ireland. In Finland, there was strong pressure from the media to publish school rankings, but the national consensus in the ensuing debate was against publicizing test results (Eurydice, 2009).

37. Several countries that do not make test results publicly available at the regional, local and school level nevertheless make the results available to selected stakeholders. These normally include local authorities, school leaders and teachers. Furthermore, several countries that publish test results at the local and school level provide selected stakeholders with additional information concerning the test results.

[2] Countries normally don't publish aggregated results from sample student assessments on the school or local level (*n* is usually too small). However, participating teachers, schools and municipalities may receive feedback from sample assessments.

4. USERS AND USES OF STUDENT TEST RESULTS

38. There are a number of uses for student test results. Table 1 provides some examples of uses.

Table 1. Uses of student test results

1. Student monitoring: decide whether students are making sufficient progress in attainment in relation to expectations.
2. Diagnosis: clarify the type and extent of students' learning difficulties in light of well-established criteria. The diagnosis is used as a basis for intervention.
3. School choice: identify the most desirable school for a child to attend.
4. Resource allocation: identify institutional needs and allocate resources.
5. Organisational intervention: identify institutional failure and justify intervention.
6. System monitoring: decide whether the education system is performing in accordance with expectations and, potentially, allocate rewards or sanctions.

(Based on House of Commons, 2007)

39. There are a number of users that use student test results. In this report the users have been categorised into eight broad categories. Table 2 presents an overview of the categories.

Table 2. Users of student test results

Category	Users
Central authorities	Ministries, Directorates, Government agencies (*e.g.* Inspectorates)
Local authorities	Regions, Municipalities, Districts
Schools	School leaders, school administrators
Teachers	Teachers working as educators and Teacher Unions
Students	Students in primary and secondary education
Parents	Parents with children in primary and secondary education
News media	Newspapers, News agencies, TV Channels
Researchers	Universities, Colleges, Research institutions

4.1. Central authorities

40. The state has long been the biggest generator, collector and user of data. Policy makers and bureaucrats have access to records and statistics dealing with almost all aspects of society. As outlined in Chapter 3.2., the great majority of OECD-countries publish aggregated assessment (national and sample) results for the education system as a whole, often in a yearly report on the state of the system and/or on an official website.

4.1.1. Central authorities' perception of student test results

41. Central authorities generally perceive student test results as a tool to support policy-making. Student assessments give a wide range of valuable information about: (a) how well students are learning in the education system; (b) whether there is evidence of particular strengths and weaknesses in students' knowledge and skills; (c) whether particular subgroups in the population perform poorly; (d) which factors are associated with student achievement; (e) whether government objectives and standards are being met; and (f) whether the achievements of students change over time (Kellaghan *et al.,* 2009).

42. Central authorities may also perceive student assessments results as a strategy for making the education system more accountable. This can be done in several ways: (a) informing citizens and parents about how well the education system in general and/or individual schools are meeting the needs of students and society, (b) communicate performance expectations to teachers and schools, and (c) using test results as a basis for rewarding and/or sanction teachers and schools according to student performance.

4.1.2. Examples of uses

43. In Belgium (the Flemish Community) the sample results are published and used as background for large conferences where the central authorities invite stakeholders to reflect on the results. Based on the discussions, a group of experts presents a list of recommendations. The stakeholders will then translate the recommendations into concrete action. Through the publication each year of high school result indicators, the French Education Ministry intends to give an account of the results of the national education public service. Furthermore, communications and conferences on the results of student assessments may be initiated at the request of teachers, researchers, parents or trade unions.

44. In Ireland, the national sample surveys are meant to provide high quality, reliable data for the Department of Education and Science to assist in policy review and formulation, and resource allocation related to English and Mathematics. Furthermore, the surveys are meant to identify factors associated with achievement (school, teacher, home background, and student factors), and identify student performance trends. In Norway the student test results are used to evaluate how successful the school system is in providing all students with basic skills. In Japan the student test results are meant to give the Ministry of Education vital information on academic performance, and to put pressure on teachers and schools to improve. In New Zealand, the stated goal of the National Education Monitoring Project (NEMP) is to provide detailed information about what children know, think and can do, so that patterns of performance can be recognised, successes celebrated, and desirable changes to educational practices and resources identified and implemented.

45. Some OECD countries, mostly European, have inspectorates. In general, Inspectorates of Education exist in order to monitor the quality level of schools and education (Wolf and Janssens, 2007). Since 2007 the Dutch Inspectorate of Education has carried out risk-based inspections of schools, assessing potential problems that could affect the quality of education. By means of full inspections the inspectorate checks whether the school fulfils its social task and/or whether the funds provided are used sensibly. On

this basis, the inspectorate issues advice on whether or not schools should be recognised or subsidised. Inspection reports are available online. Moreover, the Inspectorate publishes a "quality card" for every secondary school containing value-added school performance information. These cards show examination results, performances in individual subjects, the number of drop-outs and a comparison with similar secondary schools in the region.

46. The Department for Education in the United Kingdom (England) publishes the test results by school in achievement and attainment tables which give information on the achievements of students, and how they compare with other schools in the Local Authority area and in England as a whole. Since 2006, "contextual value-added" systems have been used which, in addition to adjusting for a pupil's own prior achievement, also attempt to adjust for factors such as the average prior achievement of a pupil's peers. The purpose of the tables is to provide clear and accessible information to parents on their children's attainment and progress. Furthermore, the tables are meant to assist school improvement and the school inspection process conducted by the Office for Standards in Education, Children's Services and Skills (OFSTEAD). Good and outstanding schools are subject to a lighter touch inspection. If a school's overall effectiveness is judged inadequate, inspectors must decide whether it requires "special measures", or a "notice to improve". A copy of the inspection report is sent to the governing body, the head teacher, the local authority and others. The governing body must send a copy of the report to all parents within five working days of receiving it. The report is subsequently published on OFSTED's website.

47. Established in 2009, the Australian Curriculum, Assessment and Reporting Authority (ACARA) is an independent authority that is responsible for publishing nationally comparable data (NAPLAN) on Australian schools. ACARA has established the *My School* website (www.myschool.edu.au) which provides detailed information about almost 10 000 schools in Australia. Results are published at the school level, and include average scores for statistically similar schools and all Australian schools. ACARA is intended as a key driver for transparency and quality in all Australian schools. Moreover, the establishment of ACARA is meant to reflect the commitment made by the Australian Government and State and Territory governments to provide all young Australians with a world class education (Australian government, 2009).

48. In the United States, the National Assessment of Educational Progress (NAEP) produces the "Nation's Report Card", to inform the public about the academic achievement of elementary and secondary students. NAEP is sponsored by the Department of Education and collects and reports academic achievement at the national level, and for certain assessments, at the state and district levels. The results are widely reported by the national and local media, and are an integral part of government evaluation of the condition and progress of education. Under the No Child Left Behind (NCLB) Act, states must measure student progress in reading, mathematics and science achievement. Schools must make annual progress toward closing the achievement gap between rich and poor, black and white, and bring all students to grade-level proficiency in math and reading by 2014. The make-up of the state-administrated assessments varies between states, especially when it comes to the stakes attached to the assessments. Two out of three states have their own policies for penalising (in addition to the ones mandated by the NCLB) low-performing schools (Chiang, 2009).

4.2. Local authorities

49. Local authorities exercise responsibilities in the field of education in several OECD countries. As outlined in Chapter 3.2., local authorities often obtain aggregated national and/or sample student test results for their own area.

4.2.1. Local authorities' perception of student test results

50. Local authorities generally support student assessment. In Norway, 70% of the municipalities report that the national student assessments have, to a great extent, lead to school improvement in their municipality (Allerup *et al.,* 2009). Support for student assessment is somewhat less apparent among local officials in the United States. A Public Agenda (2006) survey shows that 46% of superintendents consider student test results to be useful.

51. Engeland and colleagues (2008) find that both municipalities and schools have difficulties finding an organisation where the results from national tests and surveys are used for quality improvement. School leaders view the distribution of test results to the municipalities primarily as hierarchical control, and to a much lesser degree as a basis for improvement. At the municipality level there is little discussion concerning how assessment and evaluation results can be used to improve how the administration and political leaders tackle the challenges in their schools. Officials participating in a study of 69 school districts in the United States overwhelmingly agreed that it is not the lack of data that hinders data-driven decision making. Rather, the officials described that they were "being overwhelmed by the sheer number of data collects from the state for compliance reporting" (The American Productivity and Quality Center, 2009 – cited in Kline, 2009).

4.2.2. Examples of uses

52. The United Kingdom (Scotland) has developed systems enabling local authorities to increase the size of the sample student assessment within their territory in order to obtain statistically significant data for their own area. Local authorities that have opted for this system receive a targeted report from the central authorities on their relative performance (Eurydice, 2009).

53. In Finland, local authorities are given much freedom in how they organise and use local school assessment. The evaluation system is predicated on the professionalism and expertise of teachers, and aims for continuous improvement in the quality of education and training. The National Board of Education conducts representative sample studies of students learning achievements in different school types and levels. The sample results are large enough to allow comparisons between regions and municipalities, thus being a useful tool for municipalities.

54. In the United Kingdom (England) the school improvement partner programme, introduced as part of the new relationship between authorities and schools, aims to provide school leaders with challenge and support that is tailored to their needs and delivered to nationally consistent standards. The school improvement partner acts for the local authority and is the main (but not the only) channel for local authority communication about school improvement with the school. Based on a number of inputs, including student test results, the school improvement partner has a limited number of exchanges with the school's leadership about how well the school is serving its pupils and how the school needs to improve.

4.3. School leaders

55. As outlined in Chapter 3.2., school leaders systematically obtain aggregated national and/or sample student test results for their own school and/or individual students in many OECD countries. School leaders are expected to use the results for school development, as well to account for the student performance in their school (*e.g.* to parents, local authorities and/or central authorities) and sometimes also to hold teachers accountable for the results of their students.

4.3.1. School leaders' perception of student assessments results

56. In the United States, about half of the school leaders report that student test results are helpful. A large majority consider knowing how to use results to improve teaching an essential skill for a school leader (Public Agenda, 2006). English secondary schools operate within a performance management system, which includes achievement and attainment tables (formerly performance tables) reporting school performance across a number of indicators. The results from an interview-based study show that school leaders care about their school's place in the achievement and attainment tables, and that they believe this system affects behaviour (Wilson *et al.,* 2006). One the other hand, 50% of school leaders in England report that they find the different accountability measures to be a de-motivating factor in their daily work (PriceWaterhouseCoopers, 2007).

57. Teacher evaluation has historically been the responsibility of school leaders in many countries. In a large scale survey in Colorado, 81% of school leaders answered that teacher evaluation is a primal task of school leadership (Hirsch, 2009). The amount of data available to school leaders is rapidly increasing, thus providing school leaders with a number of tools for evaluation. But the richness in data also generates challenges. A study by the Norwegian Directorate for Education and Training (2008) finds that a majority of school leaders report that they have limited experience with interpreting and using data from the National Quality Assessment System.

58. Likewise, a case study by Earl and Fullan (2003) finds that school leaders in the United Kingdom (England) and Canada (Ontario and Manitoba) have anxieties about using data. Even school leaders that were positively disposed to looking at data as part of their decision-making expressed insecurity about their skill in gathering, interpreting and making sense of the information about their school. Many of them indicated that they had not had training or experience in research, data collection, data management or data interpretation. On the other hand, Allerup and colleagues (2009) find that the more support a school leader receives from the local authorities, the more useful the school leader views the data.

4.3.2. Examples of uses

59. RAISEonline (subsidiary of OFSTED) is an online tool for use by schools, local authorities, inspectors and school improvement partners in the United Kingdom (England). By providing a common set of analyses, it supports school improvement and the school inspection process. External users cannot automatically access this dataset, although schools can choose to allow them access. RAISEonline include functions that allow schools leaders to produce their own "what if" scenarios and set targets based on these, to investigate the performance of pupils in specific curriculum areas, contextual information about schools including comparisons to schools nationally. RAISEonline allows school leaders to focus on areas or groups of pupils where performance is particularly strong as well as areas for improvement.

60. In Hungary, the results of the national student assessment have to be made accessible to educational institutions and their maintainers. This is to ensure that educational institutions can compare their pedagogical work with other institutions and to identify whether an institutional improvement programme is necessary.

61. The province-administered student assessments in Canada (Ontario) are meant to give school leaders more feedback on how well students are meeting the expectations in the provincial curriculum and how effectively teaching strategies and school programmes are meeting students' needs.

4.4. Teachers

62. As outlined in Chapter 3.2., teachers often obtain aggregated national and/or sample student test results for their own students and/or their own school. Student test results can be used to assess whether

students are making sufficient progress in relation to expectations, understand student strengths and weaknesses to target further instruction and inform teacher preparation and training needs (Pathways to College Network, 2006).

4.4.1. Teachers' perception of student test results

63. A number of sample surveys cited in Mons (2009) indicate that teachers generally are positive to student assessment. A Public Agenda (2006) survey finds that only one out of five teachers in the United States considers student assessment to do more harm than good, and that most teachers give their local district rather good marks for being reasonable and putting higher academic standards in place. A Canadian (Ontario) survey reports that around ¾ of all teachers use student and school test results to identify areas of reading, writing and mathematics program strength and areas for improvement (EQAO, 2009). In Sweden, a survey by the National Agency for Education (2004) also finds that a majority of teachers consider that student assessment provides clear guidelines on teaching content, helps to highlight students' strengths and weaknesses and does not limit the scope of their teaching.

64. Statements and declarations from teacher unions are also sometimes positively inclined towards student assessment. The American Federation of Teachers (AFT) has, for many years, been supportive of quality standardised assessment that is fair and timely, and that informs and supports instruction (AFT, 2008). The Norwegian Union of Education (2005) maintains that, used properly, student assessment can be a useful tool for school improvement.

65. Nevertheless, this openness to the *principle* of national student assessment in some countries does not prevent teachers from criticizing specific aspects of national student assessments. In the United States, more than eight in ten teachers say that the schools today place far too much emphasis on standardised test scores (Public Agenda, 2003). A great majority of teachers in Florida responded that the State-administrated student assessments were not taking schools in the right direction. The teachers commented that the test results were used improperly and that the one-time test scores were not an accurate assessment of students' learning and development (Jones and Egley, 2004). Almost 90% of teachers in the United Kingdom (England and Scotland) warn that the "public ranking of schools leads to teaching to the test" and reported that "there was a real danger that public ranking of schools might lead to manipulation of data". They also disagree that competition between schools is needed to drive (school) improvement and disagree that it is necessary to publish school-specific data to enable parents to exercise choice. Scottish teachers hold these views more strongly than do their English counterparts (Croxford *et al.,* 2009).

66. Similar objections have been voiced by teacher unions. In the United Kingdom (England and Wales), the National Union of Teachers (NUT) and the National Association of Head Teachers (NAHT) voted to boycott the 2010 national student assessments. According to NUT, assessments in their current form disrupt the learning process for children in Year 6, and are misused to compile meaningless performance tables which only serve to humiliate and demean children, their teachers and their communities (NUT, 2010). Likewise, the Australian Education Union (AEU) voted to boycott the 2010 national student assessments in protest against the use of results to create school performance tables (AEU, 2010).

67. Randi Weingarten, President of the AFT, summed up this ambiguity to student assessment at the National Convention in 2008: "Tests, if they are fair and accurate, and aligned with a rich curriculum, can play an important role in holding teachers, administrators and schools accountable for much of student achievement. But the narrow numerical measures of [the No Child Left Behind Act] benefit no one, least of all the children they were supposed to help" (AFT, 2008). On a similar note, Mick Brookes, General Secretary of NAHT, has proclaimed that his organisation is in favour of student assessment as long as it is "for the right reasons and with the right instruments"(House of Commons, 2007).

68. In sum, teachers are generally more positive towards using test results for improving instruction, than using them for accountability. It is not so much the principle of being held accountable that teachers object to, but rather that what they view as flawed conclusions from student assessments are used for accountability purposes.

4.4.2. Examples of uses

69. In Canada (Ontario), the province-administered assessments are meant to give teachers more feedback on how well students are meeting the expectations in the provincial curriculum and how effectively teaching strategies and school programs are meeting students' needs. In Australia the national student assessments are meant to help teachers monitor student progress and identify students in need of additional support. The information can be used for diagnostic purposes and can assist them in their planning to cater for the individual needs of each student.

70. In Denmark, Norway and Sweden, teachers use student test results to map the strengths and weaknesses of the students in order to improve classroom instruction. In Sweden, teachers also use test results to ensure fair grading. In Finland, a wide degree of responsibility is given to teachers for the learning and assessment of students, and flexibility in exercising this responsibility.

71. The CITO monitoring and evaluation system aims to help teachers in the Netherlands determine if their students' educational progress is satisfactory and if their own educational programme is working (provides a range of tools and guides for identification, analysis, and the development of action plans). Teachers and schools get access to student data on a secure site at the internet.

4.5. Parents

72. As outlined in Chapter 3.2., a minority of OECD countries publish student test results at the school level. Countries that make such results publically available often do this because they want to bring about the active involvement of different stakeholders, particularly parents, in ensuring education quality at schools. It is assumed that informed parents effectively can challenge schools' weaknesses. Furthermore, such information may enable parents to choose a desirable school for their children (Wolf and Janssens, 2007).

4.5.1. Parents' perception of student test results

73. In general, parents seem to support student assessment schemes. Sample survey studies reviewed in Johnson and Duffett (2003) show that the majority of parents in the United States find student test results useful. Only 1 of 5 parents thinks that their child has to take too many tests (Public Agenda, 2006). In Canada (Ontario), 88% of parents consider the provincial testing program important, and 69% place high importance on having this indication of their child's achievement in relation to the provincial standard (EQAO, 2010). 42% of Norwegian adults say that the national student assessment creates the necessary foundation for improving schools (The Norwegian Directorate of Education and Training, 2009).

74. In the United Kingdom (England), a majority of parents report that they want to be able to compare one school's performance against another, that the performance of each school should be published and publicly available and that test results are one important measure of a school's performance (TNS, 2008). English parents are not so much concerned about 'what's the best school', rather they want to know 'what's the best school for my child' – a question with a highly individual answer' (Counterpoint Research, 2009).

4.5.2. Examples of uses

75. For parents to make informed choices, however, information must be available for all schools on a consistent basis to allow comparisons to be made (Bradley *et al.,* 2000). The most explicit way in which this has been provided is through the annual publication of the school achievement and attainment tables by the Department for Education in the United Kingdom (England), the State-administered School Report Cards in the United States, the province-administered student test results made available at the Education Quality and Accountability Office (EQAO) website in Canada (Ontario) and the My School website administered by the Australian Curriculum, Assessment and Reporting Authority (ACARA).

4.6. Students

76. Although students are the focal point of student assessment, they seldom get to use the results themselves. In some OECD countries students receive feedback on their score and/or their placement relative to other students.

77. Students generally dislike high stakes assessments (EPPI, 2002). Studies cited in Mons (2009) show that testing either reduced the "love of learning", which is one element of intrinsic motivation, or had no significant effect. Likewise, a systematic literature review by Nordenbo *et al.* (2009) emphasise that the test situation leads to increased stress for students, and that test results increases the motivation of high-achieving students, while decreasing the motivation of low-achieving students.

4.7. News media

78. News media have access to student test results at the school level in all OECD countries that publish results for individual schools. The results are often compiled and used as a basis for publishing performance tables and school rankings. In some countries the news media have been instrumental in compelling the government to publish student test results at the school level.

79. In France, newspapers started the process of publishing results from the *baccalauréat* (examination at the age of 18) passing rates in the lycées. Value added indicators were initially (from 1989 onwards) sent to the lycées, and thereafter (from 1991 onwards) published more widely by the Ministry of Education. In the Netherlands the newspaper *Trouw* obtained school performance data from the Ministry of Education. The latter initially refused to provide these data, arguing that it would spoil the relationship between the Dutch Schools Inspectorate and the Ministry of Education, and that publishing raw school scores would have major negative effects on schools (*e.g.* on their reputation and staff motivation). Today the newspapers *and* the Dutch Schools Inspectorate publish test results including value-added performance information (Visscher, 2001).

80. In Australia both newspapers and private enterprises use student test results (available at the government website *My School*) to create school rankings. The Sydney Morning Herald has annually published league tables for New South Wales schools. Furthermore, the entrepreneurial *Australian School Ranking,* which can be downloaded for a fee, ranks every school in Australia on a numerical list, and then provides lists of the "Bottom 100 Suburbs for Primary Education", the "Bottom 100 Suburbs for Secondary Education" and over 90 other ranked lists.

81. Since 1992, The Department for Education in the United Kingdom (England) has published school achievement and attainment tables. Based on the data from the tables, newspapers have annually compiled and published school rankings. The Norwegian Directorate of Education and Training does not facilitate the creation of league tables and school rankings, although newspapers have often managed to get hold of results at the school level. Based on this data they have published school rankings.

4.8. Researchers

82. In several OECD countries data collection and analysis of student assessment is outsourced to independent researchers – be it universities, government sponsored research centres or private research companies. These institutions may also be responsible for making student assessment data available to other researchers.

83. In Canada student assessment data, along with other school and student related information is available through Edudata Canada, an independent organisation that facilitates access to such information for qualified researchers. The National Institute for Educational Measurement (CITO) in the Netherlands is a private company that offers schools a number of products including student assessment and monitoring. In Luxembourg, the University of Luxembourg is responsible for developing, administrating and analysing the national student assessments in the country.

5. EVIDENCE REGARDING THE USE OF STUDENT TEST RESULTS

84. Chapters 3 and 4 revealed that there are numerous uses and users of student test results. Some countries use student test results solely to inform the public about the performance of the education system as a whole, while other countries use student test results to hold schools and teachers accountable to a number of stakeholders. In the latter case high stakes (*e.g.* financial rewards and sanctions) may be attached to the results. Many OECD-countries that conduct student assessments also use the results to inform classroom instruction (*e.g.* mapping student strengths and weaknesses).

85. The lion's share of the evidence regarding the use of student test results originates from countries that employ comprehensive accountability systems, notably North America and the United Kingdom. There is little evidence available from countries with no comprehensive accountability systems. Sections 5.1-5.3 examine evidence relating to the use of student test results for accountability purposes, while section 5.4 looks into evidence regarding the use of student test results for improving classroom instruction. At the end of the Chapter, table 3 sums up the arguments for and against different uses of student test results.

5.1. Publishing student test results in performance tables

86. As outlined in Chapter 3.2, a minority of OECD countries publish student test results for individual schools. Countries that do publish student test results often do this in tables that allow stakeholders easy access to detailed information about a particular school (*e.g.* student test results, student attendance and the socio-economic make-up of the school), as well as allowing stakeholders to do meaningful comparisons among schools. In systems where parents can choose the school they think is best for their children, the tables are meant to support parents in their decision. The information provided in the tables differs somewhat from country to country, although a common denominator is that stakeholders can compare the student test results of different schools. This entails that schools can be ranked according to their students' performance. When not referring to a particular country, this report uses the term "performance tables" to describe the practice of publishing student test results in a way that allows for comparison between schools.

87. The available evidence regarding the effect of publishing student test results in school performance tables is mixed. There is little evidence of a positive relationship between performance tables and increased student performance. There is, however, evidence of performance tables influencing the behaviour of schools, teachers and parents – although not always as originally intended by the authorities. Lastly, there is wide consensus in the literature that reporting student test results in performance tables is coupled with several methodological problems and challenges.

88. Based on data from more than 300 schools in the United Kingdom (England), Levacic (2004) concludes that schools respond positively to competitive pressures to improve on a particularly well publicized and widely used performance indicator. This finding indicates that competitive pressures may stimulate school leaders and teachers to improve on a measure of performance that is accorded a high public profile.

89. Hanushek and Raymond (2005) have tested whether the introduction of the School Report Cards, one of the key elements of the accountability system introduced with the No Child Left Behind act (NCLB), caused the reported increase in student achievement. Their study indicate that just reporting test

results has minimal impact on student performance and that the force of accountability comes from attaching consequences such as monetary awards or takeover threats to school performance. On the other hand, Hoxby (2001) finds that students in states with report card systems improved their reading and math skills faster than students in states that had not yet initiated such systems.

90. Under Florida's accountability program, schools are provided with grades according to their student's performance on state-administrated assessments. The grades "D" and "F" indicate that schools are low performing. Schools assigned an "F" also face the threat of a voucher program. West and Peterson (2006) find that students in Florida schools that were stigmatised as grade "D" performed better the subsequent year than students at similar schools that did not face the same stigma (grade "C"). This finding is interesting. Unlike grade "F" schools, the only stake for schools receiving grade "D" was the stigma of the label. There is also some evidence from Germany of school reputation having a positive effect on student outcome (Jürges *et al.,* 2005).

91. Teachers seem to be responsive to how their school performs on performance tables. A case study from Maryland and Kentucky finds that being labelled as low-performing and the negative publicity that accompanied it came as a shock to some teachers, especially those with more seniority. Although in time, teachers distanced themselves from the verdict of the system, the label nevertheless hurt and many wished to get rid of it (Mintrop, 2004). Likewise, a study of the accountability system in Chicago shows that teachers reported a strong desire to avoid the stigma of their school being labelled as "on probation" (Jacob *et al.,* 2004). Waterreus (2003) shows that in Dutch secondary education, scores on school report cards influence teacher mobility. Schools with "good scores" experience less staff turnover and attract more new teachers than schools with 'bad scores'. However, only minor effects are detected.

92. Wiggins and Tymms (2002) find that English primary schools perceive their accountability system (with School Achievement and Attainment tables) as being significantly more dysfunctional than those of their Scottish counterparts (without tables). English schools were more likely to report concentrating on meeting their targets (at the expense of other important objectives), narrowing the curriculum and concentrating resources on students close to reaching the threshold (who would improve their performance table position). On the other hand, Allerup and colleagues (2009) find that Norwegian teachers, school leaders and local authorities consider that the publication of results from the national student assessments in Norway have changed the way schools operate in a positive direction.

93. Performance tables may be compiled in order to support parents in choosing a desirable school for their children. Karsten and Visscher (2001 – cited in Wolf and Janssens, 2007) conclude, on the basis of a number of international studies, that parents in general pay little attention to public performance indicators. Following the introduction of NCLB in the United States, only a tiny percentage (1-5%) of students have left a "failing" school for a supposedly better school – as measured by student test results (Ravitch, 2010). Based on data from a cohort study of Scottish students entering education in 1984 and interviews with the student's parents, Echols and Willms (1995) find that parents of higher socio-economic status were more likely to value information obtained from teachers and school leaders. Most parents did not feel the need to examine all of the alternatives available; rather, they wanted to find the nearest school with a strong disciplinary climate and a positive social atmosphere.

94. Scholars point to several methodological problems regarding performance tables. A common objection is that performance tables based solely on "raw" student test results essentially measure the quality of the school intake rather than the teaching in the school (Willms, 1997; Hoyle and Robinson, 2003). Ladd and Zelli (2002) argue that a better approach is to use value-added assessment, where gains in achievement of students in specific grades from one year to the next are measured.

95. Regardless of the type of assessment used for measuring student achievement, imprecision will always be a problem. Studies of sampling variation show that the amount of variation due to the idiosyncrasies of the particular sample of students being assessed is often large relative to the total amount of variation in student performance observed. The larger the number of students tested, the more likely is it that erratic scores will cancel each other on average. But most schools are too small to support statistical confidence that student's good and bad days will average out on a single test (Rothstein *et al.,* 2008). Kane and Staiger (2001) estimate that for an average size elementary school in North Carolina, 28% of the variance in student assessment scores are due to sampling variation. Furthermore, imprecision arises from one-time factors that are not sensitive to the size of the sample: a dog barking in the playground on the day of the assessment, the weather or one particularly disruptive student in a class (Kane and Staiger, 2001; Ravitch, 2010).

96. Imprecision can be reduced through publishing student test results with a margin of error. But this margin could be larger than the true differences in average scores that would distinguish effective schools from ineffective ones. Furthermore, human variability (*e.g.* variation of bright and slow students in a single year) necessitates a further margin of error in reporting assessment scores. The margin of error increases even more when test results of subgroups (*e.g.* minorities, disabled students) are reported, because these groups are smaller (Korertz, 2008; Rothstein *et al.,* 2008).

97. A further limitation of performance tables is that differences between schools are generally not statistically significant (Petegem *et al.,* 2005). Performance tables can be a valuable tool for identifying outliers – high and low performing schools – but not for sorting or ranking the majority of schools (Rowe, 2000; Visscher, 2001; Hoyle and Robinson, 2003). In the words of Rowe (1996) – "all rankings are fallible". Moreover, it is hard to use performance tables as a tool for predicting future performance. The most recent published information is based on the current performance of a cohort of students who entered schools several years earlier, whereas for choosing a school it is the future performance of the current cohort that is of interest. Leckie and Goldstein (2009) show that few schools' future performances can be separated from both the overall mean and from one another with an acceptable degree of precision.

98. Lastly, different assessments produce different results. A variety of evidence in the late 1990s (cited in Haney, 2000; Mons, 2009) led a number of observers to conclude that the state of Texas had made near miraculous progress in increasing student achievement – as measured by the Texas Assessment of Academic Skills (TAAS). However, Klein and colleagues (2000) show that student improvements in reading and math in Texas were comparable to the national trend (as measured by NAEP data), except for fourth grade math where Texas student improvements were significantly greater than the national trend. A study of student test results in four states – Texas, North Carolina, Arkansas and Connecticut – also finds that the state-administrated test results have grown almost twice as fast as national student test results (Jacob, 2007). These findings underline that student test results are not a definitive measure of student knowledge or skills. Thus, no single assessment can be a perfect indicator of student performance (Hamilton *et al.,* 2002).

5.2. Using student test results to reward and penalise schools

99. The evidence relating to the effect of using student results to reward and/or penalise schools is mixed. On the one hand, rewards and sanctions for schools seem to have a positive effect on student performance. On the other hand, rewards and sanctions for schools seem to create unintended strategic behaviour among teachers and school leaders.

100. In the United States, 31 states reward high-performing or improving schools, while 32 states sanction low-performing schools (EPE Research Center, 2010). Evidence from Florida's A+ accountability program, which is one of the most comprehensive high stake assessment programs of its kind, is in this

regard interesting. Several scholars find that sanction threats in Florida have raised the assessment scores of students during the time that they are attending the threatened schools (Greene, 2001; Figlio and Rouse, 2005; West and Peterson, 2006; Rouse *et al.,* 2007; Chakrabarti, 2008; Chiang, 2009; Winters *et al.,* 2010). Rouse and colleagues (2007) find that schools that received a grade of "F" in summer 2002 immediately improved the test scores of the next cohort of students, and that these test score improvements were not transitory, but rather remained in the longer term. They also find that "F"-graded schools engaged in systematically different changes in instructional policies and practices as a consequence of school accountability pressure, and that these policy changes may explain a significant share of the test score improvements (in some subject areas) associated with "F"-grade receipt.

101. It is important to note that Florida's increase in student achievement is smaller when NAEP data is used to measure improvement (Figlio and Rouse, 2005), a finding that is reported in many states (see section 5.1.). Furthermore, Chiang (2009) argues that it is important to determine whether persistence of observed improvements in Florida arises from retained knowledge of *subject content* or greater familiarity with the *format* of test questions; the latter type of familiarity is arguably of less value, a point also made by Koretz (2005). Finally, Rouse and colleagues (2007) emphasise that "it is premature to outline a prescription for the improvement of low-performing schools based on these findings, particularly since we do not observe student performance along all relevant dimensions".

102. Other studies find a less clear relationship between high stakes for schools and student improvement. Evidence from a study of 18 states with high-stakes tests show that in all but one analysis, student learning is indeterminate, remains at the same level it was before the policy was implemented, or actually goes down when high-stakes assessment policies are instituted (Amrein and Berliner, 2002). Jacob (2005) finds that Chicago's high-stakes assessment system led to significant learning gains in the low-stakes subjects of science and social studies. However, he finds that these gains in low-stakes subjects due to the policy were smaller than those in the high-stakes subjects. The performance bonus plan in Chile, the *Sistema Nacional de Evaluación de Desempeño de los Establecimientos Educacionales* (SNED), rewards schools whose performance on the national student assessment places them in the top 25% of performance in the region. Mizala and Romaguera (2005 – cited in OECD, 2009) find that the SNED is not an incentive for additional effort for those schools that always score in the top 25%, nor for those that have never scored in that range. However, for those schools that have a chance of success, it has had a positive effect on student achievement. Driscoll and colleagues (2008) report similar findings from California – failure to adjust for initial conditions (*e.g.* social economic background of students) may put awards out of the reach of some schools and thus fail to produce the desired incentives. Furthermore, O'Day (2002) notes that schools with more students from high socio-economic status tend to respond more effectively to the demands of high-stakes assessment systems.

103. Test results are not the output of education, but a proxy for the education taking place every day in classrooms. When stakes are attached to the proxy, rather than the education it is meant to stand for, distortion may occur (House of Commons, 2007). Smith (1995) presents a profound theoretical analysis of the unintended consequences of publishing performance data in the public sector. Most of the effects he describes are of a strategic nature:

- Management emphasis on what is being quantified at the expense of un-quantified performance aspects
- The pursuit of narrow local objectives at the expense of the objectives of the organisation as a whole
- The pursuit of short term targets at the expense of long term objectives

- Emphasis on measures of success rather than the underlying objective
- The deliberate manipulation of data so that reported behaviour differs from actual behaviour
- The misinterpretation of data and deduction of the wrong policy measures due to bounded rationality
- The deliberate manipulation of actual behaviour to secure strategic advantage
- Organisational paralysis brought about by an excessively rigid system of performance evaluation

104. In education, "teaching to the test" is probably the most well-known example of strategic behaviour. There is substantial evidence of teachers and school leaders responding to student assessment schemes through teaching students the specific skills that are assessed, narrowing the curriculum and allocating more resources to subjects that are tested (Klein *et al.,* 2000; Linn, 2000; Stecher and Barron, 2001; Clarke *et al.,* 2003; Jacob, 2005; Center on Education Policy 2007; Hamilton, *et al.,* 2007; Slomp 2008). Hamilton and colleagues (2007) also show that teachers report focusing more on students near the proficient cut score and expressed concerns about negative effects of the accountability requirements on the learning opportunities given to high-achieving students. Likewise, Reback (2008) find that students in Texas perform better than expected when their assessment score is particularly important for their schools' accountability rating. Lastly, a study by Koretz (2005) finds that students generally score lower on an assessment that was unexpected than on an assessment for which teachers had time to prepare.

105. The desirability of teaching to the test is debated in the literature. Advocates of high stakes assessments argue that teaching to the test content is appropriate if tests are properly constructed to measure achievement (Sims, 2008). The fact that school leaders and teachers respond strategically to student assessment implies that assessments can be utilised as a powerful tool for steering classroom instruction in a desirable direction. Lane and colleagues (2002) find that school leaders and teachers tended to support the Maryland School Performance Assessment Program (MSPAP) as a tool for making changes in instruction, teachers were making some positive changes in mathematics instruction because of MSPAP, and the schools for which teachers reported that MSPAP had a greater impact on their mathematics instruction had greater MSPAP performance gains in mathematics over a 5 year period. Ladd and Zelli (2002) show that school leaders responded to North Carolina's ABCs program in ways that are consistent with the state's goal of focusing attention on the basics skills of reading, math, and writing. Koretz and colleagues (1994, 1996) report similar findings for Vermont and Kentucky. In the United Kingdom, the Select Committee report on Testing and Assessment concludes, drawing on a wide range of evidence, that "appropriate testing can help to ensure that teachers focus on achievement and often that has meant excellent teaching, which is very sound" (House of Commons, 2007).

106. Probably of more concern than teaching to the test, is the evidence regarding manipulation and outright cheating. Levitt and Jacob (2002) estimate that serious cases of teacher or administrator cheating on student assessment occur in a minimum of 4-5 percent of elementary school classrooms annually. In Texas, 700 out of the state's 8 000 schools are reported to have unusual test responses (TEA, 2007). Moreover, Figlio and Getzler (2002) estimate that the introduction of the high-stakes FCAT assessment in Florida is associated with a dramatically higher rate of disability classification. The probability that a low-performing student or a student from a low socio-economic background would be reclassified into a disability category exempted from the accountability system increased significantly after the introduction of the high-stakes FCAT assessments. Likewise, Haney (2000) show that a substantial portion of the apparent increases in Texas student test results (TAAS) in the 1990s are due to low performing students being classified as "in special education", and hence not counted in the schools accountability ratings. An investigation in 2007 by the newspaper *Cleveland Plain Dealer* determined that the school districts had

"scrubbed" or tossed out assessment scores from students who were not continuously enrolled during a school year. Most of the scores that were scrubbed were from low performers (Sims, 2008a).

107. Lastly, accountability systems – especially high stakes – are expensive to develop and administer. This is particularly so when the standards that schools are expected to reach are set high. Rothstein and colleagues (2008) have estimated that the costs of a sophisticated accountability system in the United States could be up to 1% of the total spending in primary and secondary school. Estimates of states' costs to implement the No Child Left Behind act in the United States range from $8 billion to $150 billion (Duncombe *et al.,* 2008) – a substantial cost regardless of where one places oneself on the continuum. One should note that any cost calculations is bound to be controversial because scholars do not agree on several key issues (*e.g.* what should be counted).

5.3. Using student test results to evaluate teachers

108. Numerous studies have shown that teachers matter (Hattie, 2003; Rockoff, 2004; OECD, 2005b; Rikvin *et al.,* 2005; Salvanes *et al.,* 2008). Teachers also differ regarding to how effective they are in raising student achievement (Leigh, 2007). What teachers know, do, and care about is important for student outcome. The problem is that assessing teaching quality is a very complex task. The evidence show that student test results – in combination with other measures – may serve as a basis for distinguishing between high and low performing teachers, but it is inadequate as a basis for high stake decisions such as teacher pay and promotion.

109. A fundamental challenge in holding teachers accountable for student achievement is that, as economists put it, education is jointly produced by teachers, schools, families, and communities (Hanushek, 1979; Harris, 2009). Thus, it is hard to single out the effect of a single teacher on the outcome of a single student (McCaffery *et al.,* 2003). The problem may be overcome by using value-added assessment. Value-added assessment is essentially a method for isolating the contribution of individual teachers on growth in student achievement by controlling for other potential influences on student learning, such as prior student achievement and student and family characteristics. Value-added assessment is an approach that has grown in popularity in the later years (Heyburn *et al.,* 2010). In theory value-added offers an opportunity to hold teachers accountable for student results that they can influence. In reality value-added assessment relies on a several problematic assumptions as well as some methodological limitations (Ballou, 2002; Reardon and Raudenbush, 2008; Ravitch, 2010).

110. One of the limitations with value-added assessment is that teachers are not randomly assigned to students, an assumption that most value-added assessments make (Rothstein, 2007). Moreover, a measured increase in student outcome could be due to the hard work of a prior teacher and not the current year teacher (McCaffery *et al.,* 2003). Evidence cited in Harris (2009) suggests that, despite its limitations, value-added assessment can provide useful information about teacher performance. Goldhaber and Hansen (2010) also find statistically significant relationships between North Carolina teachers' value added effectiveness measures and the subsequent achievement of students in their classes. On the other hand, such measures are somewhat unreliable, so that clear distinctions can only be made between the very highest and very lowest level of teacher value-added by traditional statistical standards (Harris, 2009). Value-added assessment, being a rather sophisticated model, often requires an outside research partner to help measure value-added performance (Meyer and Christian, 2008). It is therefore likely that using value-added assessment for teacher evaluation necessitates a rather comprehensive, and possibly costly, support system.

111. Whether value-added or not, student test results are usually available only for certain grades and subjects. In the United States, less than one in four teachers are likely to be in grades and subjects where it would be possible to evaluate teachers based on student test results (Kane *et al.,* 2010). Some groups of

teachers may be evaluated annually (*e.g.* native language teachers teaching in a Year that is assessed), while other groups of teachers may seldom or never be evaluated (*e.g.* social science teachers). This limits the feasibility of using student assessments results for evaluating teachers in general. Moreover, student test results capture only a fraction of the contribution of teachers as well as the overall mission of a school (Podgursky and Springer, 2007). If teachers are not provided with clear signals about legitimate ways in which to improve their practice, there is the danger that teachers will focus instead on teaching test-taking skills at the cost of teaching other, more difficult to measure (but valuable) skills (Kane *et al.,* 2010). This phenomenon commonly referred to as "teaching to the test". Section 5.2 provides a detailed description of "teaching to the test".

112. Student test results are sometimes used as a basis for teacher certification. Certification entails that a teacher has earned professional credentials from an authoritative source, such as a government agency. Certification may be dependent on formal training requirements, skill and ability assessments, classroom observation, student performance and/or other prerequisites. There is evidence of teacher certification having a positive effect on student outcome (Cavalluzzo, 2004; Vandervoort *et al.,* 2004; Darling-Hammond *et al.,* 2005; Smith *et al.,* 2005; Hakel *et al.,* 2008), although some studies find that certified teachers do not perform significantly better than other teachers, in spite of improvements in some grades and areas (McColskey and Stronge, 2005; Harris and Sass, 2007). Furthermore, the evaluation process leading to certification seems to have a positive effect on teacher practice. Studies relating to the National Board for Professional Teaching Standards (NBPTS) in the United States, which represents one of the most complex and comprehensive approaches to teacher certification, show that teachers apply in the classroom what they learnt from the NBPTS evaluation process (Bond *et al.,* 2000; Lustick and Sykes, 2006). Teachers who successfully go through the evaluation process are also likely to contribute to school leadership by adopting new roles including mentoring and coaching of other teachers (Petty, 2002; Freund *et al.,* 2005). It is nonetheless a challenge that high stakes is often attached to teacher certification – *e.g.* higher salary and better job opportunities – thus making it imperative that the data used for the certification decision is reliable and valid.

113. A more controversial issue is whether student test results can be used as a basis for decisions relating to teacher pay (OECD, 2009a). This is commonly referred to as performance incentive pay. The theory behind performance incentive pay is that it will motivate teachers to adapt their professional practice to address performance criteria, whether tied to student achievement measures or other indicators of good practice (Heyburn *et al.,* 2010). The evidence of the overall impact of such schemes is mixed and can be contentious and potentially divisive (OECD, 2005b). In a literature review, Hanushek and Rivkin (2006) conclude that "overall, the studies show that salaries are more likely to be positively related to student achievement than negatively. Nonetheless, only a minority is statistically significant". Furthermore, the OECD (2009b) report *Evaluating and Rewarding the Quality of Teachers* finds several international examples of how performance incentive pay schemes have led to unintended strategic behaviour among teachers.

114. In sum, there is wide consensus in the literature that student test results should not be used as the *sole* measurement of teacher performance. This holds especially true when student test results are used to make high stake decisions such as teacher pay and promotion. A valid and reliable scheme for assessing individual teacher performance requires multiple, independent sources of evidence and multiple, independent trained assessors of that evidence (CAESL, 2004; Ingvarson *et al.,* 2007; Isoré, 2009).

5.4. Using student test results to improve classroom instruction

115. As outlined in Chapter 3.2, OECD-countries that conduct student assessments usually communicate the results to school leaders and teachers. Furthermore, Chapter 4 reported that student test results are often used as a basis for improving class room instruction. The evidence stems mostly from case

studies. These studies show that student test results, when used appropriately, can help teachers understand student strengths and weaknesses to target future teaching – thus improving classroom instruction. Nevertheless, the evidence indicates that schools need capacity to interpret and use student test results.

116. Evidence cited in a literature review by Barneveld (2008) show that teachers report several positive effects of using student test results: greater differentiation of instruction, greater collaboration among faculty, increased sense of teacher efficacy and improved identification of students' learning needs. But it is not an easy task to use test results to inform classroom instruction. Results from a large scale teacher survey in Colorado are illustrative. Although 6 out of 10 teachers agree that student assessments are used to improve student learning, less than half agree that the assessments are useful in efforts to improve student learning and that results are provided in time to impact decision making (Hirsch, 2009).

117. The literature identifies at least two major stumbling blocks for using student test results to improve classroom instruction: (1) teachers generally vary in their conception of what kind of data that is valuable and how data should be used, and (2) most teachers do not have formal training in how to draw meaning from data (Chen *et al.,* 2005; Lachat and Smith, 2005; Wayman and Stringfield, 2006; Barneveld, 2008; Garmannslund *et al.,* 2008). In the latter case many teachers report that they are "flying blind" through the burgeoning amounts of student data (Supovitz and Klein, 2003; Wayman and Stringfield, 2006). Evaluation of school reform in Ontario shows that although teachers agree that data is now being used more than before to help support individual students, many teachers nevertheless indicate that they lack the knowledge and capacity to use the data to drive improvement (Ungerleider, 2007).

118. A common finding in the literature is that schools with capacity to interpret and use data improve student outcomes, while schools that don't have the same capacity are left in a deadlock. According to O'Day (2005) "the most consisting finding in the research on low-performing schools is that they generally lack the capacity to improve on their own". A case study by Isaksen (2008) is illustrative. Isaksen interviewed teachers and school leaders in low-performing schools (as measured by student test results) immediately after the publication of student test results in Norway. The interviews were repeated a year later. The school that was identified with the highest capacity at the beginning of the study had improved the most a year later. The school had essentially used the results as a tool for making necessary adjustments in classroom instruction and school management. On the other hand, the school that was identified with the lowest capacity had improved the least. No real actions had been taken between the first and second interview. The strategy of the school was essentially to talk down the relevance of the test results. Similar findings are reported in studies by Newmann (1997), Debray and colleagues (2003), Curtis and Plut-Pregelj, (2004) and Nemi and colleagues (2007).

119. Informing teachers about student results, without other accompanying polices, may therefore be insufficient – at least in schools with low capacity – to bring about desired changes in classroom instruction. In the words of Hopkins (2001): "simply collecting data, however systematically and routinely, will not of itself improve schools. There needs to be a commitment to scrutinise such data, to make sense of it, and to plan and act differently as a result". Elmore (2008) also makes the point that giving the teachers information about the effects of their practice, other things being equal, does not improve their practice.

120. The importance of school leadership for school improvement is much debated in the literature (Pont *et al.,* 2008). Elmore (2008) and Mulford (2003) propose that an essential function of school leadership is to foster "organisational learning", that is to build the capacity of the school for high performance and continuous improvement through management of the curriculum and teaching programme, development of staff and creating the climate and conditions for collective learning. It is not within the scope of this report to discuss school leadership in detail. This report merely suggests, based on evidence cited in Pont and colleagues (2008), that effective school leadership might overcome some of the

problems relating to using student test results to inform classroom instruction. Moreover, Wößmann, Lüdemann, Schütz, and West (2007) report a positive relationship between school autonomy over hiring decisions and student test results. One might draw from this that school improvement does not only rest on effective leadership, but also necessary autonomy for the school leader to exercise this leadership.

Table 3. Arguments for and against different uses of student test results

	Arguments for	Arguments against
Performance tables	Stimulates improvement Teachers are responsive to their schools' rating and ranking Involves and informs parents Holds schools accountable	Teachers dislike school rankings Parents seldom utilise the full potential of the information Statistical uncertainty and imprecision makes it hard to distinguish in a meaningful way between the majority of schools
Reward and penalise schools	Positive effect on student outcomes Powerful tool for steering classroom instruction in a desirable direction Provides teachers and schools with incentives to improve	Narrowing of the curriculum and allocation of more resources to subjects and skills that are tested Focus on students close to the proficient cut score Manipulation of test scores Teachers and schools are often rewarded and/or penalised for factors that they cannot influence
Evaluate teachers	Holds teachers accountable Value-added assessment makes it possible to isolate teacher effect on student performance Supports teacher certification	Student assessment captures only a fraction of the contribution of teachers and schools Numerous methodological challenges are attached to value-added assessment Only some years and subjects are assessed
Improve instruction	Enhances teacher professionalism Improves identification of students learning needs	Teachers lack training in analysing and using data Schools with low capacity seldom improve

6. DISCUSSION

121. So far this report has reported on how student assessments are organised in different countries, the uses and users of student test results, and the empirical evidence relating to using student test results for accountability and improvement. This chapter attempts to sum up the previous chapters and offer a discussion of the advantages and disadvantages of using student test results for holding teachers and schools accountable, as well as improving classroom instruction.

122. Based on the evidence reported in Chapter 5, four themes are discussed: (6.1) design, (6.2.) the use of test results, (6.3.) stakeholder involvement, and (6.4) implementation.

6.1. Getting the assessment design right

123. OECD countries usually employ student test results as a means for accountability. OECD countries differ quite a lot regarding how student test results are used for holding teachers and schools accountable. Chapter 5 reported evidence from the literature regarding the use of student test results for accountability purposes. A conclusion that can be drawn from the evidence is that accountability can be a powerful tool for steering and informing practice in classrooms and schools. For example, there is some evidence of teachers being responsive to how their school performs on performance tables. At the same time there are some major challenges. Accountability – especially high stakes accountability where rewards and/or sanctions are coupled with student test results – may create unintended incentives for strategic behaviour among teachers and schools (*e.g.* "teaching to the test" and test score manipulation). One can consequently argue that all accountability systems essentially produce both positive and negative effects. The challenge is to enhance the positive effects while at the same time limiting the negative effects.

124. Many OECD countries also use student test results to inform and improve classroom instruction. Evidence reported in Chapter 5 shows that student test results can help teachers understand student strengths and weaknesses to target future teaching. This necessitates that schools and teachers have the capacity to interpret and use student test results. A major challenge is that schools without this capacity have few incentives to change practice (*e.g.* to build capacity).

125. Several scholars (Darling-Hammond and Ascher, 1991; O'Reilly, 1996; Adams and Kirst, 1999; Firestone, 2002; O'Day, 2002; Levin *et al.,* 2008; Fullan, 2009) make the case that accountability needs to be integrated with, and mutually reinforced by, capacity building at the school level. In such a model top-down government performance targets and expectations is combined with bottom-up capacity building. Recent educational reform in Ontario has used such a model with success (Ungerleider, 2007). Levin and colleagues (2008) argue that there are two main lessons to be learned from Ontario: "The first is to recognise that capacity building linked to results must be the main driver. The second is to recognise the fallacy that heavy-handed accountability can create success; instead, getting better results is being more accountable". Barber (2008, 2009) also stresses the need for system leadership along with capacity building. Furthermore, O'Day (2002) argues that accountability systems will only lead to improvement if they "focus attention on information relevant to teaching and learning, motivate individuals and schools to use that information and expend effort to improve practice, build the knowledge necessary for interpreting and applying the new information to improve practice and allocate resources for all the above."

126. Thus, it seems reasonable to conclude that *assessment design – e.g.* how and for what purposes student test results are used – is crucial. Getting the design right necessitates that the different elements are well-adjusted and well-tuned to each other. Failure to do so may increase the possibility of positive effects being offset by negative effects.

6.2. The appropriate use of student test results

127. OECD countries employ a number of different methodological methods for gathering and presenting student test results. Accurate measurement of teacher and school effectiveness is crucial to the legitimacy and desirability of any system that aims to hold teachers and schools accountable. Moreover, student test results need to be valid and reliable if they are to inform and improve classroom instruction.

128. The literature emphasises that student test results is a proxy for the education taking place every day in classrooms – it is not a definitive measure of student knowledge or skills. No single assessment can be a perfect indicator of student performance. There seems to be general agreement in the literature that student assessments results can, *when appropriately used*, provide valuable information to a number of users. The bone of contention is rather how one should understand the *appropriate use* of student test results. In this respect there is a wide divergence in the literature. Some scholars argue that student test results can be sufficiently sophisticated to support high stakes decisions about rewards and sanctions for schools and teachers, while others argue that student test results are at best useful for informing policy at the system level.

129. It is, based on the evidence reported in Chapter 5, possible to make two general observations regarding the appropriate use of student test results. First, no single assessment can meet the information needs of all stakeholders. Preferences regarding what subjects to assess, the level of detail and the cycle of testing may vary among stakeholders. An assessment that is valid for one purpose is not necessarily valid for another. For example, a yearly assessment designed to hold schools accountable may offer little guidance for teachers to improve their daily classroom instruction.

130. Second, if stakes are high for teachers and/or schools, then a variety of sophisticated assessments should be used, as well as other data sources. Teachers and schools should only be accountable for factors within their control. Consider, for example, schools that serve large concentrations of disadvantaged students and that do not have sufficient compensatory resources to offset the educational challenges that such students pose. In that case, schools may be deemed ineffective despite using their insufficient resources more productively and efficiently than other schools (Ladd and Walsh, 2002). In such a case contextual value-added assessment may be a more promising approach.

131. One can conclude from these observations that it is important to have a clear understanding of the decisions that student test results are intended to inform. Decisions about assessment design require trade-offs with respect to reliability, validity, fairness and costs (Hamilton *et al.,* 2002). Comprehensive accountability systems need somewhat different assessments (and consequently results) than systems mainly designed to support classroom instruction. The challenge is employ assessments that support and complement the overall aims of the system, c.f. the discussion in Chapter 6.1.

6.3. Involving stakeholders

132. There are a number of stakeholders in OECD countries that use and/or are affected by student test results. Stakeholder acceptance of, and involvement in, student assessment is vital for the overall quality and legitimacy of the assessment system.

133. As discussed in the previous chapter, it is very difficult to develop a single assessment that satisfies the needs of all stakeholders. Based on the country practices reported in Chapter 4 and the

evidence reported in Chapter 5, one can argue that student assessment generally are more adapted to accommodate accountability needs, than towards informing the day-to-day instruction that occurs in the classroom. This has consequences for stakeholder perceptions of data usefulness. Being perceived primarily as an instrument for top-down control, teachers often find it hard to endorse student test results as a valuable tool for school improvement. Kennedy (2005) argues that highly dedicated teachers' reform rejections do not come from their unwillingness to change or improve, but from "the sad fact that most reforms don't acknowledge the realities of classroom teaching". Similar sentiments, although to a lesser degree, can be found among school leaders. Involving teachers and school leaders in the creation, implementation and review of student assessments, as well as providing them with training in data analysis and use, could help improve their perception and acceptance of student test results as a useful tool.

134. Local authorities occupy the dual position of being both principals (in its relationship with schools) and agents (in its relationship with central authorities, parents and the general public). In order to hold schools accountable and at the same time being accountable to central authorities, local authorities are in need for a wide range of data. It is important to acknowledge that local authorities are not a homogeneous group. While local authorities in cities and large districts may have a well-developed capacity to analyse and use large quantities of data, such a capacity may not be present in small districts. Assisting and supporting local authorities that do not have the sufficient capacity to analyse and use student test results may in time help to build capacity in these districts, thus improving the ability of local authorities to fulfil its dual role as both principal and agent.

135. Parents tend to welcome the publication of student test results in performance tables. Nevertheless, the evidence reported in Chapter 5 shows that parents seldom utilise the full potential of these information sources. Making sense of data is sometimes challenging for parents, especially when student test results are reported with a high degree of accuracy (*e.g.* using several sources of data, employing margins of error). Paying attention to the manner in which student test results are reported to parents, as well as providing parents with the appropriate tools and guidelines for how to use the results, could help improve the ability of parents to hold teachers and schools accountable.

136. Generally, the news media will publish student test results if the data is made available to them. Sometimes the news media is also instrumental in making student test results public. In this respect the news media can play an important role in holding the education sector accountable. In order to execute this role with an adequate degree of accuracy, the news media should be provided with the appropriate tools and guidelines for how to report the results. One should also be conscious that the news media often interpret results from a critical viewpoint and this might impact on the development of assessment policy.

137. In general, one can conclude that student test results predominately are of value to stakeholders when they are perceived as useful by the stakeholders themselves. What is considered useful vary among stakeholders. Thus, the challenge is to tailor student assessments and the information feedback from these assessments so that it suits the needs of different stakeholders – making it a valuable resource for the stakeholders. Properly executed, such an approach may even facilitate a bottom-up demand for data.

6.4. Implementation

138. Drawing on the discussion in sections 6.1-6.3, this chapter offers some general recommendations regarding the implementation of an assessment system where student test results are used to hold teachers and school accountable *and* inform classroom practice.

139. *Find the right balance between accountability and improvement.* Teachers and school leaders generally seem to favour an assessment system that predominately use student test results for informing teacher practice and professionalism. One can therefore presume that it would be easier to gain the support

of these stakeholders in implementing a system that focuses primarily on improving instruction than a system where teacher and teacher and school accountability is the dominant feature. Nevertheless, a system where student test results are employed without accountability runs the risk of not being responsive to organisational and political demands, and/or to the needs of the users (Levitt *et al.,* 2008). Thus, countries that do not have a well-developed tradition for using student test results for improving instruction may need to enforce and support this practice with some measures of accountability.

140. *Clear understanding of what student test results can (and cannot) be used for.* Test results are not the output of education, but a proxy for the education taking place every day in classrooms. A single assessment may be sufficient to measure what students can and cannot do in certain subjects and grades, but it may not be sufficient to inform high stakes decision such as teacher salary and promotion. Teachers often object to being held accountable for student assessments results because they view the assessments as inadequate for measuring the whole registry of teaching. The higher the stakes attached to the test results, the more data is needed to inform the decision-making process. Moreover, teachers and schools should only be accountable for factors that they can influence. This is important for the fairness and legitimacy of the accountability system.

141. *Student test results need to be useful for stakeholders.* The support of key stakeholders may be acquired through providing them with data that they perceive as valuable and useful. This entails that data must to be tailored to the needs of different stakeholders.

142. *Comprehensive use of student test results is costly.* While sample student assessments are cheaper to administer than national student assessment, the usefulness for the former is limited to system level monitoring. The more advanced uses that student test results are intended for, the more sophisticated data and accompanying support systems (including training of users) are needed. This entails considerable expenses for countries.

7. CONCLUSION

143. This report has investigated the use of student test results in OECD countries. The report finds that holding teachers and schools accountable for their students' performance – as measured by student test results – can be a powerful tool for changing teacher and school behaviour. Moreover, student test results can be a useful for identifying students' weaknesses and strengths, thus assisting teachers in improving classroom instruction. The challenge is to get the assessment design right. Too much focus on accountability may create a number of unintended negative effects. On the other hand, a system where student test results are employed without accountability runs the risk of not being responsive to organisational and political demands, and/or to the needs of the users. This report suggests that it is important to find a workable balance between accountability and improvement – a balance where they mutually support and enforce each other.

144. Regardless of how an assessment system is organised, student test results must be reliable, valid and fair. No single assessment can be a perfect indicator of student performance. Thus, several assessments should be used to measure student outcome – especially when the stakes attached to the results are high for teachers and schools. Moreover, teachers and schools should only be accountable to factors within their control. This increases the need for sophisticated assessments, as well as high quality gathering, analysis and use of the test results.

8. APPENDIX

145. In Chapter 3, a brief account was given concerning how OECD countries report student test results. This appendix offers a more thorough outline of how student test results are reported in two countries – Australia and the United Kingdom (England). The countries are chosen because they both represent comprehensive reporting systems.

8.1. Australia

146. Established in 2009, the Australian Curriculum, Assessment and Reporting Authority (ACARA) is an independent authority that is responsible for analysing, evaluating and publishing nationally comparable data on Australian schools. ACARA is intended as a key driver for transparency and quality in all Australian schools. Moreover, the establishment of ACARA is meant to reflect the commitment made by the Australian Government and State and Territory governments to provide all young Australians with a world class education.

147. A main responsibility for ACARA is the National Assessment Program – Literacy and Numeracy (NAPLAN), which assesses students in Years 3, 5, 7 and 9. NASPLAN results are intended to inform a number of users and uses:

- Students and parents may use individual results to discuss achievements and progress with teachers
- Teachers use results to help them better identify students who require greater challenges or additional support
- Schools use results to identify strengths and weaknesses in teaching programs and to set goals in literacy and numeracy
- Authorities use results to review programs and support offered to schools

148. Schools must send parents a report which shows the individual performance of their children on the NAPLAN assessments. Six performance levels are reported for each year level. One of the levels represents the national minimum standard for students. A result at the national minimum standard indicates that the student demonstrated the basic literacy and numeracy skills needed to participate in that year level. The performance of individual students can be compared to the average performance of all students in Australia.

149. In order to increase transparency and accountability in the Australian school system, ACARA has established the *My School* website. The website, which is open to all, provides easy access to detailed information about almost 10 000 schools in Australia. The information reported on the website include: a descriptive statement of the school (done by the school itself), useful facts about the school, the socio-economic background of the school, and the average NASPLAN results of the school. The main features are further outlined in table 4.

150. In addition to the *My School* website, ACARA publishes the National Report on Schooling in Australia. The report is intended to inform the Australian people on progress against the national goals for schooling and agreed national performance measures.

Table 4. Information reported on the *My School* website

School statement: In this section the school can give an account of the school's mission, values, special programs, and other information that gives a broader picture of the school.

School facts:

- School sector: government or non-government school
- School type: primary, secondary, combined (primary *and* secondary) or special purpose (*e.g.* juvenile justice) schools
- Year range offered by the school
- Enrolment: all students (head count) and fulltime equivalent enrolments
- Percentage of Indigenous Australian students: Aboriginal and/or Torres Islander decent
- Location: metropolitan, provincial, remote or very remote
- Student attendance rate: aggregated attendance across levels 1-10
- Number of non-teaching staff: all non-teaching staff (head count) and fulltime equivalent job load
- Number of teaching staff: all teachers (head count) and fulltime equivalent job load

School socio-economic background:

- ICSEA value: The Index of Community Socio-Educational Advantage (ICSEA) is a measure that enables meaningful and fair comparisons to be made across schools. The variables that make up ICSEA include socio-economic characteristics of the area where the students live, the location of the school (regional or remote) and the proportion of Indigenous students enrolled in the school. The average ICSEA value is 1000 – most schools should have a value between 900-1100
- ICSEA quarters: ICSEA quarters for each school are displayed in percentages. This gives contextual information about the socio-educational composition of the student population. If students at a school were drawn proportionally from the broad spectrum of the community, then theoretically there would be 25% in each quarter

NASPLAN results:

- Results are reported as a school average in all tested subjects
- Results are compared to statistically similar schools and all Australian schools
- Participation, absentee and exemption rates are reported: school and national average
- Indicative confidence intervals for the results

Senior secondary outcomes (data is not comparable between jurisdictions):

- Number of seniors that have completed secondary school
- Number of seniors that have completed a specific training program (*e.g.* VET, SBAT)
- The post-school destination of former seniors (vocational training, university studies or in employment)

8.2. United Kingdom (England)

151. In the United Kingdom (England) national student assessments are held at the end of Key Stage 1 (Years 1-2), 2 (Years 3-6) and 4 (Years 10-11). At the end of Key Stage 1, teachers assess student progress in English and math (measured by tasks and tests that are administered informally). At the end of Key Stages 2, students take national tests in English, math and science. There is no national test at the end of Key Stage 3. At the end of Key Stage 4 students sit exams for the General Certificate of Secondary Education (GCSE) and/or equivalent qualifications.

152. The Department for Education in the United Kingdom (England) publishes the test results in School Achievement and Attainment tables which give information on the achievements of students, and how they compare with other schools in the Local Authority area and in England as a whole, as well as a number of background variables. The main features are further outlined in table 5.

153. The purpose of the tables is to provide clear and accessible information to parents on their children's attainment and progress. Furthermore, the tables are meant to assist school improvement and the school inspection process conducted by the Office for Standards in Education, Children's Services and Skills (OFSTEAD). A copy of the inspection report is sent to the governing body, the head teacher, the local authority and others. The governing body must send a copy of the report to all parents within five working days of receiving it. The report is subsequently published on OFSTED's website.

154. The website RAISEonline provides schools, local authorities, inspectors and school improvement partners with an online tool for analysing student test results and other school relevant data. External users cannot automatically access this dataset, although schools can choose to allow them access. RAISEonline include functions that allow schools leaders to produce their own "what if" scenarios and set targets based on these, to investigate the performance of pupils in specific curriculum areas, contextual information about schools including comparisons to schools nationally. RAISEonline allows schools leaders to focus on areas or groups of pupils where performance is particularly strong as well as areas for improvement.

Table 5. Information reported in the *School Achievement and Attainment* tables

School background information:

- Total number of pupils enrolled
- Total number and percentage of pupils with Special Education Needs (SEN)

Key Stage test results:

- Total number and percentage of students with SEN taking the tests
- Percentage of students achieving level 4 or above: reported for each tested subject
- Percentage of students achieving level 5: reported for each tested subject
- Percentage of students absent from the tests: reported for each tested subject
- Average point score for all tested subjects

Contextual value added:

- Contextual Value Added (CVA) score: measures the progress made by students from the end of a Key Stage to the end of another Key Stage using their test results (and some exam results in secondary school – CVA EM). CVA takes into account the varying starting points of each students' test results, and also adjusts for factors which are outside a school's control (such as gender, mobility and levels of deprivation) that have been observed to impact on student results
- Confidence interval: shows the upper and lower limit of the 95% confidence interval
- The average number of qualifications taken by student included in the CVA (EM) calculation

Year on year comparisons:

- Aggregate of test percentages for level 4 and above, and level 5: makes it possible to compare school Key stage results from the years 2006-2009. It is not the same students that are compared, thus results may fluctuate from year to year for reasons to do with the students rather than the school

Progress Measures:

- National targets require schools to ensure that a specified percentage of pupils make at least expected progress in English and, separately, in math between the end of KS1 and the end of KS2. "Expected progress" is two national curriculum levels of progress

Absence record:

- Overall absence: reported as a percentage
- Persistent absence: the percentage of student enrolments equalling or exceeding the threshold number of half-day over the Autumn and Spring terms combined

REFERENCES

Adams, J.E. and M. Kirst (1999), "New Demands for Educational Accountability: Striving for Results in an Era of Excellence", in Murphy, J. and K.S. Louis (eds.), *Handbook of Research in Educational Administration*, San Francisco, CA: Jossey-Bass.

AEU (2010), *www.aeufederal.org.au/Media/MediaReleases/2010/1204(2).pdf*, accessed 27 April 2010.

AFT (2008), *www.aft.org/issues/standards/student-assess*, accessed 26 April 2010.

Allerup, P., V. Kovac, G. Kvåle, G. Langfeldt and P. Skov (2009), *Evaluering av det Nasjonale kvalitetsvurderingssystemet for grunnopplæringen,* FoU rapport/2009, Kristiansand: Agderforskning.

Amrein, A. and D. Berliner (2002), "High-Stakes Testing, Uncertainty, and Student Learning", *Education Policy Analysis Archives*, Vol. 10, No. 18.

Australian Government (2009), *Australian Curriculum, Assessment and Reporting Authority*, Media Release 3 June 2009, the Minister for Education, *www.deewr.gov.au/Ministers/Gillard/Media/Releases/Pages/Article_090603_143742.aspx*, accessed 8 July 2010.

Ballou, D. (2002), "Sizing up Test Scores", *Education Next*, Summer 2002.

Barber, M. (2008), *Instruction to Deliver*, London: Methuen.

Barber, M. (2009), "From System Effectiveness to System Improvement", in A. Hargreaves and M. Fullan (eds.), *Change Wars*, Bloomington, IN: Solution Tree.

Barneveld, C. (2008), *Using Data to Improve Student Achievement*, Research Monograph No. 15, the Literacy and Numeracy Secretariat, Ontario *www.edu.gov.on.ca/eng/literacynumeracy/inspire/research/Using_Data.pdf*, accessed 27 May 2010.

Bond, L., T. Smith, W. Baker and J. Hattie (2000), "The Certification System of the National Board for Professional Teaching Standards: A Construct and Consequential Validity Study", *the National Board for Professional Teaching Standards*.

Booher-Jennings, J. (2007), "Michael Fullan, Turnaround Leadership", Book Review, *Journal of Educational Change,* Vol. 8, No. 3, pp. 291-294.

Bracci, E. (2009), "Autonomy, Responsibility and Accountability in the Italian School System", *Critical Perspectives on Accounting*, Vol. 20, No. 3, pp. 293-312.

CAESL (2004), *Using Student Tests to Measure Teacher Quality*, Assessment Brief No. 9, January 2004 *www.caesl.org/briefs/Brief9.pdf,* accessed 25 May 2010.

Carnoy, M., R. Elmore and L. Siskin (eds., 2003), *The New Accountability: High Schools and High Stakes Testing,* New York, NY: RoutledgeFalmer.

Cavalluzzo, L. (2004), *Is National Board Certification an Effective Signal of Teacher Quality*?, Alexandria, VA: the CNA Corporation.

Center on Education Policy (2007), *Choices, Changes, and Challenges: Curriculum and Instruction in the NCLB Era*.

Chakrabarti, R. (2008), *Impact of Voucher Design on Public School Performance: Evidence From Florida and Milwaukee Voucher Programs,* Staff Report Number 315, Federal Reserve Bank of New York.

Chen, E., M. Heritage and J. Lee (2005), "Identifying and Monitoring Students' Learning Needs with Technology", *Journal of Education for Students Placed at Risk,* Vol. 10, No. 3, pp. 309-32.

Chiang, H. (2009), "How Accountability Pressure on Failing Schools Affects Student Achievement", *Journal of Public Economics,* Vol. 93, No. 9-10, pp. 1045-1057.

Clarke, M., *et al.* (2003), *Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from Interviews with Educators in Low-, Medium-, and High-Stakes States,* Lynch School of Education, Boston College, Mass.: National Board on Educational Testing and Public Policy.

Counterpoint Research (2009), *School Accountability and School Report Card: Qualitative Research. Executive Summary,* DCSF Research Report 106, London: DCSF.

Croxford, L., J. Gray and J. Ozga (2009), "Teacher Attitudes to Quality Assurance and Evaluation (QAE) in Scotland and England", *CES Briefing No. 51*, Edinburgh: Centre for Educational Sociology, University of Edinburgh.

Curtis, K. and L. Plut-Pregelj (2004), "Schools Moving Toward Improvement", in H. Mintrop (ed.), *Schools on Probation: How Accountability Works (doesn't work),* New York, NY: Teachers College Press.

Darling-Hammond, L. (2004), "Standards, Accountability, and School Reform", *Teachers College Record,* Vol. 106, No. 6, pp. 1047-1085.

Darling-Hammond, L. and C. Ascher (1991), "Creating Accountability in Big City School Systems", *Urban Diversity Series*, No. 102.

Darling-Hammond, L., D.J. Holtzman, S.J. Gatlin and J.V. Heilig (2005), "Does Teacher Preparation Matter? Evidence about Teacher Certification, Teach for America, and Teacher Effectiveness", *Education Policy Analysis Archives,* Vol. 13, No. 42.

Debray, E., G. Parson and S. Avila (2003), "Internal Alignment and External Pressure: High School Response in Four States", in M. Carnoy, R. Elmore and L. S. Siskin (Eds.), *The New Accountability: High Schools and High Stakes Testing,* New York, NY: RoutledgeFalmer.

Driscoll, D., D. Halcoussisand and S. Svorny (2008), "Gains in Standardized Test Scores: Evidence of Diminishing Returns to Achievement", *Economics of Education Review*, Vol. 27, No. 2, pp. 211-220.

Duncombe, W., A. Lukemeyer and J. Yinger (2008), "The No Child Left Behind Act: Have Federal Funds Been Left Behind?", *Public Finance Review*, Vol. 36, No. 4, pp. 381-407.

Earl, L. and M. Fullan (2003), "Using Data in Leadership for Learning", *Cambridge Journal of Education,* Vol. 33, No. 3.

Echols, F.H. and J.D. Willms (1995), "Reasons for School Choice in Scotland", *Journal of Education Policy,* Vol. 10, No. 2, pp. 143-156.

ECS (2003), *No Child Left Behind Issue Brief, www.ecs.org/clearinghouse/35/50/3550.pdf,* accessed 20 April 2010.

Elmore, R. (2008), "Leadership as the Practice of Improvement", in B. Pont, D. Nusche and D. Hopkins (eds.), *Improving School Leadership, Volume 2: Case Studies on System Leadership*, Paris: OECD Publishing.

Engeland, Ø., G. Langfeldt and K. Roald (2008), "Kommunalt handlingsrom: Hvordan forholder norske kommuner seg til ansvarsstyring i skolen?" in G. Langfeldt, E. Elstad and S.T. Hopmann (eds.), *Ansvarlighet i skolen – Resultater fra forskningsprosjektet "Achieving School Accountability in Practice"*, Oslo: Cappelen forlag.

EPE Research Centre (2010), *Quality Counts 2010*.

EPPI (2002), *A Systematic Review of the Impact of Summative Assessment and Tests on Students' Motivation for Learning,* London: EPPI-Centre, Institute of Education, University of London.

EQAO (2009), *The Power of Good Information: Assessing Every Student,* Toronto, ON: EQAO.

EQAO (2010), *Parents' Perspectives: The Importance of Provincial Testing and the Information it Provides about Children's Learning,* Toronto, ON: EQAO Research.

Eurydice (2009), *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results*.

Figlio, D.N. and C. Rouse (2005), "Do Accountability and Voucher Threats Improve Low-Performing Schools?", *Journal of Public Economics*, Vol. 90, No. 1-2, pp. 239-255.

Figlio, D.N. and L.S. Getzler (2002), "Accountability, Ability and Disability: Gaming the System"*, NBER Working Paper Series,* Working Paper No. 9307.

Firestone, W. (2002), "Educational Accountability", *Encyclopedia of Education*.

Freund, M., V.K. Russell and C. Kavulic (2005), "A Study of the Role of Mentoring in Achieving Certification by the National Board for Professional Teaching Standards", *the National Board for Professional Teaching Standards.*

Fullan, M. (2009), "Large-Scale Reform Comes of Age", *Journal of Educational Change*, Vol. 10, No. 2-3.

Garmannslund, P., E. Elstad and G. Langfeldt (2008), "Lærernes opplevelse av måling og rangering av kvalitetsaspekter ved undervisning og læringsprosesser", in G. Langfeldt, E. Elstad and S.T. Hopmann (eds.), *Ansvarlighet i skolen: Politiske spørsmål og pedagogiske svar,* Oslo: Cappelen forlag.

Goldhaber, D. and M. Hansen (2010), "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions", *National Centre for Analysis of Longitudinal Data in Education Research*, Working Paper No. 31.

Greene, J. (2001), *An Evaluation of the Florida A-Plus Accountability and School Choice Program*, New York: Manhattan Institute for Policy Research, *www.manhattan-institute.org/html/cr_aplus.htm,* accessed 7 June 2010.

Hakel, M.D., J.A. Koenig and S.W. Elliott (2008), *Assessing Accomplished Teaching: Advanced-Level Certification Program*, Washington D.C.: The National Academic Press.

Hamilton, L.S., B.M. Stecher and S.P. Klein (eds.) (2002), *Making Sense of Test-Based Accountability in Education*, Santa Monica, CA: RAND.

Hamilton, L.S., *et al.* (2007), *Implementing Standards-Based Accountability under No Child Left Behind Responses of Superintendents, Principals, and Teachers in Three States,* Santa Monica, CA: RAND.

Haney, W. (2000), "The Myth of the Texas Miracle in Education", *Education Policy Archives*, Vol. 8, No. 41.

Hanushek, E.A. (1979), "Conceptual and Empirical Issues in Estimating Educational Production Function Issues", *Journal of Human Resources,* Vol. 14, pp. 351-388.

Hanushek, E.A. and M.E. Raymond (2005), "Does School Accountability Lead to Improved Student Performance?", *Journal of Policy Analysis and Management,* Vol. 24, No. 2, pp. 297-328.

Hanushek, E.A. and S.G. Rivkin (2006), "Teacher Quality", *Handbook of the Economics of Education,* Amsterdam: North Holland.

Harlen, W. (2007), "Criteria for Evaluating Systems for Student Assessment", *Studies in Educational Evaluation,* Vol. 33, No. 1, pp. 15-28.

Harris, D. (2009), "Would Accountability Based on Teacher Value Added be Smart Policy? An Examination of the Statistical Properties and Policy Alternatives", *Education Finance and Policy,* Vol. 4, No. 4, pp. 319-350.

Harris, D. and T. Sass (2007), "The Effects of NBPTS-Certified Teachers on Student Achievement", *Center for Analysis of Longitudinal Data in Education Research*, Working Paper No. 4.

Hattie, J. (2003), "Teachers Make a Difference: What is the Research Evidence?", presentation at the Australian Council for Educational Research, October 2003 *www.det.nsw.edu.au/proflearn/docs/pdf/qt_hattie.pdf,* accessed 25 May 2010.

Heyburn, S., J. Lewis and G. Ritter (2010), *Compensation Reform and Design Preferences of Teacher Incentive Fund Grantees,* Policy Paper, The National Center on Performance Incentives.

Hirsch, E., A. Sioberg and A. Germuth (2009), *TELL Colorado: Creating Supportive School Environments to Enhance Teacher Effectiveness*, the New Teacher Center.

Hopkins, D. (2001), *School Improvement for Real,* London: Routledge & Farmer.

Hopmann, S. (2008), *Ansvarliggjøring i praksis. Program Kunnskap, utdanning og læring*, Oslo: Forskningsrådet.

House of Commons (2007), *Testing and Assessment: Third Report of Session 2007-08*, Children, Schools and Families Committee, London.

Hoxby, C.M. (2001), *Testing is About Openness and Openness Works*, *www.hoover.stanford.edu/pubaffairs/we/current/hoxby_0701.html,* accessed 10 April 2010.

Hoyle, R. and J. Robinson (2003), "League Tables and School Effectiveness: A Mathematical Model", *Proceedings of the Royal Society of Biological Science,* Vol. 270, No. 1511, pp. 113-119.

Ingvarson, L., E. Kleinhenz and J. Wilkinson (2007), *Research on Performance Pay for Teachers*, Australian Council for Educational Research.

Isaksen, L.S. (2008), "Skoler i gapestokken", in G. Langfeldt, E. Elstad and S. T. Hopmann (eds.) *Ansvarlighet i skolen: Politiske spørsmål og pedagogiske svar,* Oslo: Cappelen Forlag.

Isoré, M. (2009), "Teacher Evaluation: Current Practices in OECD Countries and a Literature Review", *OECD Education Working Paper* No. 23, Paris: OECD Publishing.

Jacob, A. and M. Kirst (1999), "New Demands and Concepts for Educational Accountability: Striving for Results in an Era of Excellence", in J. Murphy and K. Seashore Lewis (eds.), *Handbook of Research on Educational Administration*, San Francisco, CA: Jossey-Bass, 1999.

Jacob, B.A. (2005), "Accountability, Incentives and Behaviour: The Impact of High-Stakes Testing in the Chicago Public Schools", *Journal of Public Economics*, Vol. 89, No. 5, pp. 761-796.

Jacob, B.A. (2007), "Test-Based Accountability and Student Achievement: An Investigation of Differential Performance on NAEP and State Assessments", *NBER Working Paper Series*, Working Paper No. 12817.

Jacob, B.A. and S.D. Levitt (2003), "Rotten Apples, an Investigation of the Prevalence and Predictors of Teacher Cheating", *The Quarterly Journal of Economics,* Vol. 118, No. 3, pp. 843-877.

Jacob, R., S. Stone and M. Roderick (2004), *Ending Social Promotion: The Effects on Teachers and Students,* Chicago: Consortium on Chicago School Research.

Johnson, J. and A. Duffett (2003), *Where We Are Now. 12 Things You Need to Know about Public Opinion and Public Schools*, New York, NY: Public Agenda.

Jones, B.D. and R.J. Egley (2004), "Voices from the Frontlines: Teachers' Perceptions of High-Stakes Testing", *Education Policy Analysis Archives*, Vol. 12, No. 39.

Jürges, H., W.F. Richter and K. Schneider (2005), "Teacher Quality and Incentives, Theoretical and Empirical Effects of Standards on Teacher Quality", *FinanzArchiv,* Vol. 61, No. 3, pp. 298-326.

Kane, T., E. Taylor, J. Tyler and A. Wooten (2010), "Identifying Effective Classroom Practices Using Student Achievement Data", *NBER Working Paper Series*, Working Paper No. 15803.

Kellaghan, T., V. Greaney and S. Murray (2009), *Using the Results of a National Assessment of Educational Achievement,* Washington D.C.: The World Bank.

Kennedy, M. (2005), *Inside Teaching*, London: Harvard University Press.

Klein, S., L. Hamilton, D. McCaffrey and B. Stencher (2000), *What do Test Scores in Texas Tell us?* Issue paper, Rand Education.

Kline, D. (2009), "Data, Data Everywhere, but Not a Drop to Use", *Education Week*, Vol. 28, Issue 33.

Korertz, D. (2008), *Measuring up: What Educational Testing Really Tells us,* Cambridge, Mass.: Harvard University Press.

Koretz, D. (2005), "Alignment, High Stakes, and the Inflation of Test Scores" in J. Herman and E. Haertel (eds.), *Uses and Misuses of Data in Accountability Testing*, Yearbook of the National Society for the Study of Education, Vol. 104, Part 2, Malden, MA: Blackwell Publishing.

Koretz, D., B. Stecher, S. Klein and D. McCaffrey (1994), "The Vermont Portfolio Assessment Program: Findings and Implications", *Educational Measurement: Issues and Practice,* Vol. 13, No. 3, pp. 5-16.

Koretz, D., S. Barron, K. Mitchell and B. Stecher (1996), *The Perceived Effects of the Kentucky Instructional Results Information System (KIRIS)*, Santa Monica: RAND.

Lachat, M.A. and S. Smith (2005), "Practices that Support Data Use in Urban High Schools", *Journal of Education for Students Placed at Risk,* Vol. 10, No. 3, pp. 333–349.

Ladd, H. (2007), "Holding Schools Accountable Revisited", Spencer Foundation Lecture in Education Policy and Management, presented at the 2007 APPAM Fall Research Conference, Washington DC, November 8.

Ladd, H.F. and A. Zelli (2002), "School-Based Accountability in North Carolina: The Responses of School Principals", *Educational Administration Quarterly,* Vol. 38, No. 4.

Ladd, H.F. and R.P. Walsh (2002), "Implementing Value-Added Measures of School Effectiveness: Getting the Incentives Right", *Economics of Education Review*, Vol. 21, No. 1, pp. 1-17.

Lane, S., C.S. Parke and C.A. Stone (2002), "The Impact of a State Performance-Based Assessment and Accountability Program on Mathematics Instruction and Student Learning: Evidence from Survey Data and School Performance", *Educational Assessment,* Vol. 8, pp. 279-315.

Leckie, G. and H. Goldstein (2009), "The Limitations of Using School League Tables to Inform School Choice", *Centre for Market and Public Organisation*, Working Paper No. 09/208.

Leigh, A. (2007), "Estimating Teacher Effectiveness from two-year Changes in Students' Test Scores", *Economics of Education Review,* Vol. 29, No. 3, pp. 480-488.

Levacic, R. (2004), "Competition and the Performance of English Secondary Schools: Further Evidence", *Education Economics*, Vol. 12, No. 2.

Levin, B., A. Glaze and M. Fullan (2008), "Results without Rancor or Ranking: Ontario's Success Story", *Phi Delta Kappan*, Vol. 90, No. 4, pp. 273-280.

Levitt, R., B. Janta and K. Wegrich (2008), *Accountability for Teachers: A Literature Review,* Santa Monica, CA: Rand Cooperation.

Linn, R.L. (2000), "Assessments and Accountability", *Educational Researcher*, Vol. 29, No. 2, pp. 4-16.

Looney, J. (forthcoming). "Student Formative Assessment within the broader Evaluation and Assessment Framework", OECD Education Working Paper (forthcoming), Paris: OECD Publishing.

Lustick, D. and G. Sykes (2006) "National Board Certification as Professional Development: What are Teachers Learning?", *Education Policy Analysis Archives*, Vol. 14, No. 5.

McCaffery, D., J. Lockwood, D. Koretz and L. Hamilton (2003), *Evaluating Value-Added Models for Teacher Accountability*, Santa Monica, CA: Rand Education.

McColskey, W. and J. Stronge (2005), "A Comparison of National Board Certified Teachers and non-National Board Certified Teachers: Is there a Difference in Teacher Effectiveness and Student Achievement", *the National Board for Professional Teaching Standards*.

Meyer, R. and M. Christian (2008), *Value-Added and other Methods for Measuring School Performance,* NCPI Conference 2008.

Mintrop. H. (ed.) (2004), *Schools on Probation: How Accountability Works (and doesn't work),* New York, NY: Teachers College Press.

Mons, N. (2009), *Theoretical and Real Effects of Standardized Assessment, Background Paper to the Study National Testing of Pupils in Europe: Objectives, Organisation and Use of Results,* EACEA: Eurydice.

Mulford, W. (2003), "School Leaders: Challenging Roles and Impact on Teacher and School Effectiveness", a paper prepared for the OECD *Improving School Leadership* activity.

Nemi, D., E.L. Baker and R.M. Sylvester (2007), "Scaling up, Scaling down: Seven Years of Performance Assessment Development in the Nation's Second Largest School District", *Educational Assessment,* Vol. 12, pp. 195-214.

Newmann, F.M., M.B. King and M. Ringdon (1997), "Accountability and School Performance: Implications from Restructuring Schools", *Harvard Educational Review,* Vol. 67, pp. 41-74.

Nordenbo, S., *et al.* (2009), Pædagogisk bruk av test – en systematisk review, København: Danmarks pædagogiske universitetsforlag og Dansk Clearinghouse for Uddanelsesforskning.

Nusche, D. (forthcoming). "Summative Assessment: What's in it for Students?", OECD Education Working Paper (forthcoming), Paris: OECD Publishing.

NUT (2010), *www.teachers.org.uk/node/9359,* accessed 26 April 2010.

O'Day, J. (2002), "Complexity, Accountability, and School Improvement", *Harvard Educational Review*, Vol. 72, No. 3.

O'Reilly, F.E. (1996), *Educational Accountability: Current Practices and Theories in Use,* Harvard University, Consortium for Policy Research in Education.

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

OECD (2005b), *Teachers Matter*, Paris: OECD Publishing.

OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis,* Paris: OECD Publishing.

OECD (2009a), *Teacher Evaluation: A Conceptual Framework and examples of Country Practices,* Paris: OECD Publishing.

OECD (2009b), *Evaluating and Rewarding the Quality of Teachers: International Practice,* Paris: OECD Publishing.

Petegem, P., J. Vanhoof, F. Daems and P. Mahieu (2005), "Publishing Information on Individual Schools?", *Educational Research and Evaluation*, Vol. 11, No. 1, pp. 45-60.

Petty, T. (2002), "Identifying the Wants and Needs of North Carolina High School Mathematics Teachers for Job Success and Satisfaction", *the National Board for Professional Teaching Standards*.

Podgursky, M. and M. Springer (2007), "Teacher Performance Pay: A review", *Journal of Policy Analysis and Management*, Vol. 26, No. 4, pp. 909-950.

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

Popham, J. (1991), "Why Standardized Tests don't Measure Educational Quality", *Educational Leadership,* Vol. 56, No. 6, pp. 8-15.

Popham, J. (2003), *Test Better, Teach Better: The Instructional Role of Assessment,* Alexandria, VA: Association for Supervision and Curriculums Development.

PriceWaterhouseCoopers (2007), *Independent Study into School Leadership: Main Report*, report prepared for the Department for Education and Skills.

Public Agenda (2003), *Rolling up their Sleeves: Superintendents and Principals Talk about What's Needed to Fix Public Schools*, New York, NY: Public Agenda.

Public Agenda (2006), *Reality Check 2006: Is Support for Standards and Testing Fading?*, New York, NY: Public Agenda.

Ravitch, D. (2010), *The Death and Life of the Great American School System: How Testing and Choice are Undermining Education,* New York, NY: Basic Books.

Reardon, S. and S. Raudenbush (2008), "Assumptions of Value-Added Models for Estimating School Effects", paper prepared for the National Conference on Value-Added Modeling, April 22-24, 2008 University of Wisconsin at Madison.

Reback (2008), "Teaching to the Rating: School Accountability and the Distribution of School Achievement", *Journal of Public Economics,* Vol. 92, pp. 1394-1415.

Rikvin, S., E. Hanushek and J. Kain (2005), "Teachers, School and Academic Achievement", *Econometrica*, Vol. 73, No. 2, pp. 417-458.

Rockoff, J. (2004), "The Impact of Individual Teachers on Students' Achievement: Evidence from Panel Data", *American Economic Review,* Vol. 94, No. 2, pp. 247-52.

Rothstein, J. (2007), "Do Value-Added Models Add Value? Tracking, Fixed Effects, and Causal Inference", *CEPS Working Paper,* No. 159.

Rothstein, R., R. Jacobsen and T. Wilder (2008), *Grading Education: Getting Accountability Right*, Teachers College Press and EPI Book.

Rouse, C.E., J. Hannaway, D. Goldhaber and D. Figlio (2007), *Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure*, Urban Institute.

Rowe, K. (1996), "Assessment, Performance Indicators, League Tables, Value-Added Measures and School Effectiveness: Issues and Implications", *IARTVSeminar Series, 58,* October 1996, Melbourne.

Rowe, K. (2000), "Assessment, League Tables and School Effectiveness: Consider the Issues and 'Let's Get Real'!", *Journal of Educational Enquiry,* Vol. 1, No. 1.

Salvanes, K., *et al.* (2008), *Veien mot kunnskapslandet*, Søf-rapport 01/08.

Sims, D. (2008b), "Strategic Responses to School Accountability Measures: It's all in the Timing", *Economics of Education Review*, Vol. 27, No. 1, pp. 58-68.

Sims, S. (2008a), "Districts 'Scrubbing' Away Thousands of Students' Test Scores", *Cleveland Plain Dealer. http://blog.cleveland.com/metro/2008/09/districts_scrubbing_away_thous.html,* accessed 5 May 2010.

Slomp, D. (2008), "Harming not Helping: The Impact of a Canadian Standardized Writing Assessment on Curriculum and Pedagogy", *Assessing Writing*, Vol. 13, No. 3, pp. 180-200.

Smith, P. (1995), "On the Unintended Consequences of Publishing Performance Data in the Public Sector", *International Journal of Public Administration,* Vol. 18, pp. 277-310.

Smith, T., B. Gordon, S. Colby and J. Wang (2005), "An Examination of the Relationship between Depth of Student Learning and National Board Certification Status", Office for Research on Teaching, Appalachian State University.

Stecher, B. and S. Barron (2001), "Unintended Consequences of Test-Based Accountability When Testing in 'Milepost' Grades", *Educational Assessment,* Vol. 7, No. 4, pp. 259-81.

Supovitz, J. and V. Klein (2003), *Mapping a Course for Improved Student Learning: How Innovative Schools Systematically Use Student Performance Data to Guide Improvement*, Philadelphia, PA: Consortium for Policy Research in Education.

TEA (2007), *Test Security Enhancements Planned*, *http://ritter.tea.state.tx.us/press/07,* accessed 14 April 2010.

The Norwegian Directorate for Education and Training (2008), *Kunnskapsløftet – fra ord til handling. Erfaringer og statistikk etter to søknadsrunder i programmet*.

The Norwegian Directorate of Education and Training (2009), *Informasjon om nasjonale prøver*.

The Norwegian Union of Education (2005), *www.utdanningsforbundet.no/no/Aktuelt/Alle-nyheter/Fra-forbundet/2005/Var-holdning-til-nasjonale-prover/,* accessed 26 April 2010.

The Pathways to College Network, (2006), *Using Data to Improve Educational Outcomes, Institute for Higher Education Policy*, *www.pathwaystocollege.net/pdf/data.pdf,* accessed 26 May2010.

The Swedish National Agency for Education (2004), *Det nationella provsystemet i den målstyrda skolan. Omfattning, användning och dilemma*.

TNS (2008), *School Accountability and School Report Card Omnibus Survey (November 2008) Top Line Findings*, DCSF Research Report 107, London: DCSF.

Ungerleider, C. (2007), *Evaluation of the Ontario Ministry of Education's Student Success Strategy, Phase 1 Report,* Ottawa: Canadian Council on Learning.

Vaishnav, A. (2005), "Adding Value to Student Assessment: Does "Value-Added Assessment" Live up to its Name?", *Harvard Education Letter*, Vol. 21, No. 3, pp. 1-3.

Vandervoort, L., A. Amrein-Beardsley and D. Berliner (2004), "National Board Certified Teachers and Their Students' Achievement", *Education Policy Analysis Archives*, Vol. 12, No. 4.

Visscher, A.J. (2001), "Public School Performance Indicators: Problems and Recommendations", *Studies Educational Evaluation,* Vol. 27, No. 3, pp. 199-214.

Waterreus, I. (2003), *Lessons in Teacher Pay; Studies on Incentives and the Labour Market for Teachers,* Doctoral Thesis, University of Amsterdam.

Wayman, J.C. and S. Stringfield (2006), "Technology-supported Involvement of Entire Faculties in Examination of Student Data for Instructional Improvement", *American Journal of Education,* Vol. 112, No. 4, pp. 549-571.

West, M. and P. Peterson (2006), "The Efficacy of Choice Threats Within School Accountability Systems: Results from Legislatively Induced Experiments", *the Economic Journal*, Vol. 116, No. 510, pp. C46-C62**.**

Wiggins, A. and P. Tymms (2002), "Dysfunctional Effects of Public Performance Indicator Systems: a Comparison between English and Scottish Primary Schools", *Public Money and Management,* Vol. 22, No. 1, pp. 43-48

Willms, J.D. (1997), "Parental Choice and Education Policy", *CES Briefing No. 12,* Centre for Educational Sociology, University of Edinburgh.

Wilson, D., B. Croxson and A. Atkinson (2006), "What gets Measured gets Done: Headteachers' responses to the English Secondary School Performance Management System 1", *Policy Studies*, Vol. 27, No. 2, pp. 153-171.

Winters, M.A., J.R. Trivitt and J.P. Greene (2010), "The Impact of High-Stakes Testing on Student Proficiency in Low-Stakes Subjects: Evidence from Florida's elementary science exam", *Economics of Education Review*, Vol. 29, No. 1, pp. 138-146.

Wolf, I.F. 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.

Wößmann, L., E. Lüdemann, G. Schütz and M.R. West (2007), *School Accountability, Autonomy, Choice and the Level of Student Achievement: International Evidence from PISA 2003*, OECD Education Working Paper No. 13, Paris: OECD.

Zucker, S. (2004), *Administration Practices for Standardized Assessments,* Pearson Assessment Report.

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.