

OECD Education Working Papers No. 65

Student Standardised Testing: Current Practices in OECD Countries and a Literature Review

Allison Morris

https://dx.doi.org/10.1787/5kg3rp9qbnr6-en

OECD

Unclassified **EDU/WKP(2011)10**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **11-Oct-2011**

English - Or. English

DIRECTORATE FOR EDUCATION

EDU/WKP(2011)10
Unclassified

STUDENT STANDARDISED TESTING: CURRENT PRACTICES IN OECD COUNTRIES AND A LITERATURE REVIEW

OECD Education Working Paper No. 65

by Allison Morris

This paper was prepared for the OECD by Allison Morris, a graduate student at the Institut d'Etudes Politiques de Paris (Sciences Po), and is part of the work undertaken by the OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes.

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

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

JT03308839

Document complet disponible sur OLIS dans son format d'origine
Complete document available on OLIS in its original format

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.

The opinions expressed in these papers are the sole responsibility of the author(s) and do not necessarily reflect those of the OECD or of the governments of its member countries.

Applications for permission to reproduce or translate all or part of this material should be sent to OECD Publishing, rights@oecd.org or by fax 33 1 45 24 99 30.

---

www.oecd.org/edu/workingpapers

Copyright OECD 2011

ABSTRACT

This report discusses the most relevant issues concerning student standardised testing in which there are no-stakes for students ("standardised testing") through a literature review and a review of the trends in standardised testing in OECD countries. Unlike standardised tests in which there are high-stakes for students, no-stakes implies that test results have no impact on the student's academic career. The same tests, however, may have high stakes for teachers and schools. The report provides an overview of the standardised testing typology in the no-stakes context, including identifying the driving trends behind the gradual increase in standardised testing in OECD countries and the different purposes of standardised tests. Within this framework the report reviews how standardised tests with no-stakes for students are designed, implemented and used across OECD countries. The report also aims to synthesise the relevant empirical research on the impact of standardised testing on teaching and learning and to draw out lessons from the literature on aspects of standardised tests that are more effective in improving student outcomes. Key debates concerning standardised testing are identified throughout and include (among others): 1) selecting the appropriate test purpose; 2) teacher evaluation based on student test results; 3) the impact of publishing standardised test results; and 4) minimising strategic behaviour by teachers and administrators in standardised testing.[1]

RÉSUMÉ

Ce rapport analyse les questions essentielles sur les tests standardisés des élèves dont les résultats n'ont pas d'implications pour les élèves (« tests standardisés ») à travers une revue de la littérature et une analyse des tendances dans les pays de l'OCDE. Ce type de tests n'a pas d'implications pour le parcours scolaire des élèves. Ces mêmes tests peuvent toutefois avoir des conséquences pour les enseignants et les écoles. Le rapport offre une typologie de l'utilisation des tests standardisés sans conséquences pour les élèves y compris les raisons pour l'augmentation de leur utilisation dans les pays de l'OCDE et ses différents objectifs. Dans ce cadre, le rapport analyse la façon dont les tests standardisés sans conséquences pour les élèves sont conçus, implémentés et utilisés dans les pays de l'OCDE. Le rapport a aussi pour buts de synthétiser la recherche empirique pertinente sur l'impact des tests standardisés sur l'enseignement et l'apprentissage et de retirer des leçons de la littérature sur les aspects des tests standardisés qui sont plus efficaces dans l'amélioration des résultats des élèves. Des débats clés concernant les tests standardisés sont identifiés, notamment (entre autres) : 1) sélectionner l'objectif approprié pour le test ; 2) évaluation des enseignants sur la base des résultats des tests standardisés des élèves ; 3) l'impact de la publication des résultats des tests standardisés des élèves ; et 4) minimisation du comportement stratégique des enseignants et administrateurs dans la mise en place des tests standardisés.

[1] Allison Morris, an American national, was part of the team working on the OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes during the summer 2011 at the Education and Training Policy Division, Directorate for Education, OECD. Allison has a Master's degree in Public Affairs from Sciences Po, Paris with a specialisation in Human Security. She also has research experience in the areas of microfinance, education in emergencies and economic development. Currently, Allison is working as the Director of aidha, Singapore, a social enterprise that delivers financial education to low-income migrant workers. She can be contacted at allison.morris@aidha.org.

TABLE OF CONTENTS

1. INTRODUCTION 5

2. A TYPOLOGY OF STANDARDISED TESTING IN OECD COUNTRIES 7

2.1 Drivers behind the trend of standardised testing 7
2.2 Purposes: What do standardised tests attempt to achieve? 9
2.3 Test design and development 14
2.4 Reference standards 17
2.5 Use of ICT in student standardised testing 20
2.6 Implementation and scoring 20
2.7 Limitations of standardised tests 21
2.8 Suitability criteria 22
2.9 Use of standardised test results 26
2.10 Decisions on reporting results 31

3. COMPETENCIES FOR DEVELOPING AND USING STUDENT STANDARDISED TESTING 33

3.1 Agencies responsible for test design and implementation 33
3.2 Development of capacity and assessment literacy 34

4. EMPIRICAL EVIDENCE ON THE IMPACT OF STANDARDISED TESTS FOR IMPROVING STUDENT OUTCOMES 35

4.1 The impact of standardised tests on student outcomes 35
4.2 The impact of student standardised tests on teaching: Unintended consequences of standardised tests 37
4.3 Standardised tests may not reduce achievement gaps 40

5. ASPECTS OF STANDARDISED TESTING THAT ARE MORE CONDUCIVE TO IMPROVING SCHOOL OUTCOMES 42

5.1 Lesson 1: Clearly establish the purpose of the test and allow this to lead all following test design, implementation and use decisions 42
5.2 Lesson 2: Testing standards should be aligned with the national curriculum to testing standards 42
5.3 Lesson 3: Be cautious in employing large-scale, standardised tests that serve multiple purposes. 43
5.4 Lesson 4: Develop assessment literacy of teachers and administrators 43
5.5 Lesson 5: Reduce distortion and strategic behaviour by increasing teacher involvement and buy-in from an early stage 43
5.6 Lesson 6: Incorporate multiple measures of achievement especially in systems where standardised tests may be perceived as 'high-stakes' for teachers and school administrators 44

6. CONCLUDING REMARKS 45

APPENDIX A: OVERVIEW OF STANDARDISED TESTS IN SOME OECD COUNTRIES 46

REFERENCES 47

1. INTRODUCTION

1. This paper reviews the current practices in large-scale, student standardised testing in OECD countries. It narrows the scope of standardised testing to examine student standardised testing that holds *no stakes or consequences for the student*; but it does not exclude tests in which there may be stakes attached for teachers and/or school administrators. The paper reviews both full-cohort (census based) and sample based standardised testing systems, both of which are often used in OECD countries. It examines the academic and policy literature surrounding the topic and describes the typical framework associated with student standardised testing found in OECD countries today. Further, the paper attempts to bring forth various debates associated with no-stakes standardised tests and it summarises empirical evidence on the effects of testing on teaching and learning outcomes.

2. Standardised testing is often used in OECD countries as a means to assess students, teachers and schools; however across countries substantial differences exist in test purpose, design, implementation and use of test results (Kellaghan *et al.*, 2009). The term standardised test refers to tests that are designed externally and aim to create conditions, questions, scoring procedures and interpretations that are consistent across schools (Popham, 1999; Wang *et al.*, 2006). Such tests are typically given to large groups of students at once for different purposes and the results of the tests are used in various ways, including assessment and evaluation. Assessment refers to systemically collecting evidence relating to student achievement and using this evidence to make a judgement about learning (EPPI, 2002 cited in Nusche, forthcoming; Harlen, 2007). The literature on assessment practices distinguishes that standardised test results (high or low stakes) can be used for different assessment purposes: summative, formative or monitoring and evaluation. Summative assessment refers to "assessment of learning" and involves high stakes consequences for students because the results of the assessment are used to judge the students' performance (Ewell, 2005 cited in Nusche, forthcoming). Formative assessment is often referred to as "assessment for learning" and supports a teacher's pedagogical approach to the student; the results of such an assessment are used to improve teaching strategies and identify learning needs, rather than judge performance (Black and William, 1998 cited in Nusche, forthcoming; Eurydice Network, 2009). Finally, assessment tools such as standardised tests can be used for monitoring and evaluation purposes. Evaluation refers to collecting evidence to judge systems, programs and procedures (Harlen, 2007; Newton, 2007). This paper considers standardised testing for monitoring, evaluation and formative purposes, but because summative assessment inherently involves stakes for students, it is not included in this review. For a comprehensive review of summative assessment trends in OECD countries, see: Nusche (forthcoming).

3. The literature on standardised testing is often dominated by debates over the advantages or disadvantages of standardised tests with *high stakes* for students (Popham, 1999; Wang *et al.*, 2006). This paper looks at another end of the spectrum of standardised tests: those in which there are no-stakes attached for students. The term stakes refers to judgements passed based on test results. Tests with high stakes for students imply that the results feed into decisions about the student's academic or professional career – for example, results from such a test determine whether one passes a grade, enters higher education or obtains a certificate. On the other hand, tests with no stakes for students are those in which results have no effect on the student's career, such as national tests used for monitoring purposes. From this point onward in the paper, the terms "standardised testing", "national tests", or "large-scale, standardised tests" refer to tests with *no stakes for students* (although there may be stakes for teachers and schools). Testing with stakes for students will be specifically denoted as "high-stakes standardised tests."

4. As OECD countries continue to implement accountability measures in public services (Shewbridge, forthcoming), no-stakes standardised tests play an important role in assessing the effectiveness and outcomes of a country's education system. A country may choose to administer low-stakes standardised tests for a number of reasons and in response to different international and national pressures, which in turn link to the test's purpose, design and how test results are used. The trends in OECD countries show a growing reliance on the results of standardised tests for a number of purposes and it is important to keep in mind the resulting effect on teaching and learning. Assessment practices – whether they are focused on system, school, teacher or student results – impact teaching practices and teacher-student relationships and, in certain cases, can restrict learning and teaching (Harlen, 2007; Santiago *et al.*, 2011). This paper addresses this issue by reviewing the current trends with regard to large-scale standardised testing and discussing the impacts of different system choices. Table 1 lists five examples of no-stakes standardised tests in OECD countries, which are often referred to throughout the remainder of this paper.

Table 1: Examples of standardised tests with no-stakes for students

Country	**Standardised test**
Australia	National Assessment Program in Literacy and Numeracy (NAPLAN)
Canada	Pan-Canadian Assessment Programme (PCAP)
Chile	*Sistema de Medición de Calidad de la Educación* (SIMCE) – System to Measure Quality in Education
New Zealand	National Education Monitoring Project
United States	National Assessment of Educational Progress (NAEP)

5. This paper is structured as follows: Section 2 attempts to identify a typology of standardised testing in OECD countries. It highlights key features, including the advantages and disadvantages, of testing practices as they feed into the education system and it reviews the drivers and purposes for standardised tests as well as test design and implementation practices. Section 3 discusses the competencies required for test design and the use of test results. Section 4 introduces the empirical evidence on the effects of different standardised testing practices on teaching and learning. Section 5 reviews certain aspects which are more conducive to an equitable, effective, valid and reliable standardised testing mechanism and Section 6 offers some concluding remarks.

2. A TYPOLOGY OF STANDARDISED TESTING IN OECD COUNTRIES

6. This section describes the key features of how standardised tests are designed and implemented in OECD countries. It reviews the drivers behind the development of a testing scheme, the purposes associated with its implementation and different criteria that feed into the test design.

2.1 Drivers behind the trend of standardised testing

7. Across OECD countries, the literature identifies numerous international trends that act as drivers influencing a country's decision to administer large-scale, standardised tests. While it is difficult to discern exactly which driver may have distinctly played a role in a nation's decision to administer standardised tests, it is nonetheless important to identify the underlying currents and intersecting trends steering assessment systems (Mons, 2009). The primary drivers identified are: 1) New Public Management; 2) Standards-based assessment; 3) International competition; 4) Increasing demand for 21st Century Skills; 5) Test industry pressure.

2.1.1 Driver 1: New Public Management

8. Standardised tests used for national monitoring, evaluation or accountability purposes are often a reflection of government efforts to 'modernise' the education system and incorporate business practices into public service management (Kellaghan, 2003 cited in Greaney and Kellaghan, 2008; Figlio and Kenny, 2009 cited in Figlio and Loeb, 2011). There has also been a growing emphasis on quantitatively measuring outcomes and objectives and reforms towards decentralisation and autonomy which have contributed to the need to develop new means to monitor education systems (Mons, 2009: 5). Shewbridge (forthcoming) also identifies the growing trend across OECD countries to introduce output and/or outcome measures in different government dimensions, such as budgetary procedures.

9. These trends, often referred to as New Public Management (NPM) or Results-Based Management, aim to analyse public sector operations by improving cost-effectiveness, quantifying output and making public bodies with greater autonomy accountable to citizens and system managers (Mons, 2009). NPM shifts measures of efficiency from use of inputs to quantifiable outputs. In terms of the education system, efficiency and effectiveness would be measured by outputs, such as test scores and graduation rates, rather than inputs, like funding, resources and number of school days. The increased attention to quality of public outputs and the insertion of business management techniques into public sector institutions are one of the various reasons a country may adopt and administer a national test (Ball, 1998).

10. New Public Management also addresses increased demands for accountability and efficiency in education systems that are triggered by decentralisation reforms. As stated in the Eurydice Network's comprehensive publication *National Testing of Pupils in Europe, "*In the last two decades, national testing has been increasingly introduced as a natural accompaniment to growing school autonomy, which has resulted in a need to systematise the monitoring of education systems, and in efforts to improve the quality of education" (2009: 21). In Europe, decentralisation reforms of the 1990s gave schools more autonomy, likewise increasing the need for results-based management (Eurydice Network, 2009). For example, in Latvia and Poland national-level measurement tools were implemented following reforms to increase school and teacher autonomy (Eurydice Network, 2009: 19).

11. New Public Management incorporates indicators and benchmarks as a tool to assess a public program's efficiency and throughout OECD country assessment systems, indicators and benchmarks play a significant role in determining whether the education system is meeting national standards or curriculum goals. In fact, government-wide performance measures exist in all OECD countries, except for six European countries (Shewbridge, forthcoming). New Public Management and similar performance measures are used in Europe, Australia, Denmark and Canada. Delvaux and Mangez identify the use of indicators and benchmarks in governing the supra-national European education system (Delvaux and Mangez, 2008 cited in Shewbridge, forthcoming). Australia's National Assessment Program in Literacy and Numeracy (NAPLAN), measures student results against national standards and Denmark's national tests monitor progress against the national Common Objectives (Santiago *et al.*, 2011; Shewbridge *et al.*, 2011). Campbell and Levin cite the Canadian government's broader movement to advance the use evidence-based decision making as a critical driver behind Canada's student assessment practices, including standardised testing (2008). Hence, standardised tests are at times a means to monitor or measure outputs and are often used to determine whether the education system is meeting national education benchmarks.

2.1.2 Driver 2: Standards-based assessment

12. Apart from the larger, government-wide initiatives to incorporate performance measures through New Public Management tools, OECD countries are also developing educational standards as a means to measure national progress and system performance. By creating a set of standards to measure student performance by, governments aim to evaluate students against a desired measurable outcome, rather than against their peers. Educational standards can refer to the level of achievement in a subject at a certain age; for example, the proficiency in English or Mathematics by 13 years of age would be an educational standard. Countries such as the U.S., Australia, the UK and Denmark have implemented educational standards, through which the performance of the education system is measured.

13. This trend is also referred to as Standards-Based Assessment when education systems are evaluated by comparing students to standards, rather than to peers. As stated by Wang *et al.*, "Standards-based assessment is concerned with how well a student's performance is relative to a prescribed set of content standards rather than relative to a norm group of peer students" (2006: 311). This concept is further elaborated by Figlio and Loeb as the identification of clear and measurable standards which students are expected to meet (2011). By setting national or common standards, school outcomes can be more easily controlled for quality and they are more comparable (Wang *et al.*, 2006: 311). Therefore, the movement towards comparing student outcomes to standards also has a role in motivating governments to administer national tests. As countries – like the U.S., Australia, the UK and Denmark have done – incorporating Standards-Based Assessment is often a pre-cursor to measuring standards attainment through standardised tests.

2.1.3 Driver 3: Increased international competition

14. Increased international competition, as a result of globalisation and the dissemination of International Tests, such as the OECD's Programme for International Student Assessment (PISA) and Trends in International mathematics and Sciences Study (TIMSS), is another driver behind the trend of administering standardised tests. This driver has two dimensions: first, globalisation has led to increased competition and higher mobility of labourers; second, as international standardised tests highlight educational differences between countries, governments are motivated to keep up with competitive standards. In Europe, about 1/3 of countries have stated that results of international surveys like PISA and TIMSS have "fuelled" an increased demand for "fuller information about the curriculum and teaching methods" (Eurydice Network, 2009). The U.S.' below average performance on such tests has also been a force behind using standardised tests to improve school outcomes (Wang *et al.*, 2006).

15. In the World Bank publication "Assessing National Achievement Levels in Education", Greaney and Kellaghan (2008) see the role of standardised tests on as a measure of a country's ability to compete in the global market and to drive economic growth. Throughout the literature on education systems, the link between education and economic growth has been explored, as education outputs are increasingly used as a measurement for economic indicators (Barro, 1999 cited in Mons, 2009; OECD, 1989 and Ross, Paviot and Genevois, 2006 cited in Wang *et al.*, 2006).

2.1.4 Driver 4: Increased demand for particular subject areas, such as 21st Century Skills

16. Standardised tests are also driven by the changing definition of education and what it means to have skills suitable for professional and civil life. With the growing use of ICT, students are required to have new skills in technology and communication – referred to as 21st Century Skills – in order to succeed. As OECD countries attempt to develop knowledge economies, students will require higher levels of skills and knowledge in areas like mathematics and science in order to meaningfully contribute to the workforce (Greaney and Kellaghan, 2008).

17. Although there is increasing use of ICT in schooling and in standardised testing, national tests still focus on traditional education outputs of literacy and mathematics. Nonetheless, it can be anticipated that the increasing demand for ICT literacy may, over time, drive countries to incorporate such skills into national tests and national standards. At this time, however, ICT and 21st Century Skills play a role in the design, implementation and scoring of standardised tests; rather, than being an assessment subject outright.

2.1.5 Driver 5: Test industry pressure

18. A final driver behind the trend of administering large-scale, standardised tests comes from the growing and profitable industry of standardised test development. In some countries, private test developers have powerful lobbies and can pressure governments to rely on their tests for national monitoring and evaluation purposes. With the growing number of government-issued standardised tests in countries, the testing market has become more attractive. In the U.S., the No Child Left Behind act requires approximately 45 million standardised tests annually and the costs associated with developing, administering, publishing, scoring and reporting NCLB standardised tests is estimated to be between $5 million and $7 million a year (Toch, 2006). Moreover, the U.S. Testing market is dominated by only a handful of companies, which represent 90% of testing revenue (Toch, 2006). The testing companies that design and analyse standardised tests are likely to have an influence on a government's decision to employ standardised tests; however, additional empirical research should be done in this area to explore the impact the testing industry has on the diffusion of standardised testing.

2.2 Purposes: What do standardised tests attempt to achieve?

19. This section describes the three primary purposes of large-scale, student standardised testing with no stakes involved. As defined by Gipps and Stobart (1993), "The purpose of assessment refers to the intention behind the assessment" (cited in Nusche, forthcoming). Across the literature, it is understood that there can be a multitude of uses of national tests results (described in further detail in Section 2.9); however, each use is typically linked to one or more of the three primary purposes of with standardised tests: 1) monitor and evaluate national education system; 2) hold schools and educators accountable 3) provide information to the public and 4) serve formative purposes (Figlio and Loeb, 2011; Eurydice Network, 2009). For a summary of standardised tests and their purposes in some OECD countries, see Appendix A.

2.2.1 Why it is important to define a purpose

20. There is agreement across the literature that the purpose behind a standardised test should guide the rationale for the assessment and feed into the design and implementation of the test, as well as steer the use of the test results (Greaney and Kellaghan, 2008). The purpose must be clear in order to ensure that the test is appropriately designed and valid evidence is collected (Kellaghan *et al.*, 2009). Since curriculum, instruction and assessment are intertwined, if there are conflicting goals associated with the assessment, the validity of the test is undermined (Alliance for Excellent Education, 2010). Further, if the test is designed for a specific purpose, the results should not be used for a different purpose as it is likely that any inferences made based on test results will not be accurate or valid for other purposes (Alliance for Excellent Education, 2010). Green and Oates comment on the importance of clearly defined and separated testing purposes: "We would like to suggest that separation of purposes and careful alignment of these with adequate and well-matched operational arrangements to deliver on these purposes is vital in respect of responsible and efficient public policy" (2009: 235). Therefore, there is general agreement in the literature that the purpose should be continually aligned with the use of test results in order to increase the validity of the information and that the purpose has an incredible impact on the design of the test.

2.2.2 Purpose 1: Standardised tests to monitor and evaluate the education system

21. Across OECD countries, standardised tests are used to monitor and evaluate a country's education system. In this sense, the test results are used to answer the questions: are students meeting national or minimum standards? How well is the education system functioning? Monitoring and evaluation refers to collecting and analysing data to check performance against goals and to take remedial actions if needed (Eurydice Network, 2009). In this manner, "national test results are used as indicators of the quality of teaching and the performance of teachers, but also to gauge the overall effectiveness of education policies and practices" (Eurydice Network, 2009: 8).

22. Underlying this purpose is the assumption that test results will be used to determine where students stand with regard to standards and that information will be used to improve student outcomes. In this way, national tests that are used to monitor and evaluate the education system are national, complementary efforts to international surveys. In Denmark, for example, the Ministry of Education states that the national test has two purposes: monitor school performance and provide diagnostic information about areas for students' improvement (Shewbridge *et al.*, 2011). In Australia, the NAPLAN test purpose is to compare student results with national minimum standards established for each year level in areas such as literacy and numeracy (Santiago *et al.*, 2011). In a 2009 education survey, more than half of the European countries surveyed indicated having national tests with the objective of monitoring and evaluating schools, or the educations system as a whole (Eurydice Network, 2009: 23). Specifically, national test results are used for national monitoring in Belgium (Flemish Community), Estonia, Ireland, Spain, France, Finland, the United Kingdom (Scotland), Chile, Japan, New Zealand and Canada, among others (Eurydice Network, 2009; Shewbridge, forthcoming). In Korea, the national assessment of educational achievement aims to improve teaching methods, identify student achievement levels nationally, and to collect data to feed into curriculum development (INCA, 2011). In Sweden, national tests are used to measure student progress towards standards 'embedded' in the national curriculum (the issue of aligning standards and tests with national curriculum is discussed in further detail in Section 2.4.4) (Nusche *et al.*, 2011).

23. When standardised tests are used to monitor and evaluate education systems test results are compared with national standards to determine how the education system is performing or can be improved. Proponents of establishing standards for student performance argue that it is a means of removing the ambiguity often associated with traditional testing and instruction goals; therefore, standards

increase comparability across schools while allowing the public to see whether schools are effective, an issue which has previously been shrouded in uncertainty (Wang *et al.*, 2006: 311).

24. If a test is not well-aligned with the national curriculum, it may not be an appropriate tool to monitor the education system. If the test is not fit to be assessing the desired standards, the purpose will not be met. In some cases, countries may be face difficulties effectively measuring educational outcomes if test development is not aligned with the national curriculum or educational standards. Therefore, it is increasingly important for educational entities to coordinate the development of educational standards, the national curriculum and standardised tests (Green and Oates, 2009). A second difficulty presented when using a standardised test to monitor and evaluate the education system is that the results of the test may not be a true account of what students are capable of; that is, since the stakes of the assessment are low for students, students may not try very hard on the assessment (Nusche, forthcoming). A third challenge of this approach to testing is the risk of inaccurately applying "one-size-fits-all" standards and externally imposing standards on all schools and students (Wang *et al.*, 2006: 313). This can have a detrimental effect in cases where "'non-standard kids' simply do not fit into that standard mould" (Wang *et al.*, 2006: 313). Therefore, a government must attempt to strike a balance between effectively monitoring the education system's outputs in a passive way, without undermining the effort of students to perform on standardised assessments.

2.2.3 Purpose 2: Standardised tests hold the education system (and/or its components) accountable

25. Standardised tests that are used to monitor the progress of students in meeting national standards often are also created with the purpose of holding components of the education system accountable to certain targets and outcomes. Thus, Purpose 1 and Purpose 2 are often dual objectives of standardised tests, as seen in the cases of the U.S., Australia and Chile. The primary difference between Purpose 1 and Purpose 2 is that monitoring aims to improve system-wide policies, whereas test-based accountability systems reward or sanction schools and teachers based on their ability to meet desired targets or standards and their ability to improve student performance (Hamilton and Koretz, 2002; Ladd, 2007 cited in Rosenkvist, 2010).

26. Attaching accountability measures to standardised test results is closely related to the growing trend on New Public Management, mentioned in Section 2.1. In accountability systems, actions or decisions are made based on whether schools meet certain performance targets and standards. Rewards can include teacher bonuses or other financial incentives, whereas sanctions can include intervention, school closure, or resource limitations. In terms of who is held accountable, this can vary between school and teacher. In some systems, teachers are held accountable for student test results and results are used to evaluate a teachers' performance. In the U.S., the No Child Left Behind (NCLB) Act of 2001 set forth unprecedented provisions to hold schools accountable for student performance and attached high stakes consequences to assessment outcomes (Wang *et al.*, 2006: 306). Test-based accountability is a highly contested topic, especially as many authors argue that education is produced jointly by teachers, schools, families and communities (for a detailed account of this debate see Rosenkvist, 2010). Therefore, holding solely teachers accountable for student performance in standardised tests ignores the role of parents and other environmental or economic factors in the learning process of an individual.

27. The literature surrounding using both high- and low-stakes test results to hold schools, teachers or students accountable is substantial (see Perie, 2007; Hout and Elliott, 2011; Kane *et al.*, 2002; Hanushek and Raymond, 2004; Chiang, 2009; Wang *et al.*, 2006). Attaching accountability measures to low-stakes standardised tests changes the outlook of teachers and school administrators and the standardised tests are subsequently perceived as "high-stakes". There is consensus in the literature that incentivising standardised tests for teachers or school administrators can lead to distorted practices such as: teaching to the test, narrowing of curriculum, teacher cheating, or student exclusion (Abrams *et al.*, 2003; Braun, 2005;

Guilfoyle, 2006; Figlio and Loeb, 2011; Hout and Elliott, 2011). In general, attaching incentives to standardised tests in accountability systems can lead teachers to perform actions that increase test results, while undermining the value of their work (Hout and Elliott, 2011). The unintended consequences test-based accountability has on teaching and learning must be considered when developing a standardised test for this purpose. These strategic behaviours and distortions decrease the validity of test results and undermine the accountability system. A more detailed description of these implications can be found in Section 2.9, which discusses the impact of different test uses.

28. The advantage of using standardised tests for accountability purposes is that the data is seen as more objective and less ambiguous than classroom tests (Frary, Cross and Weber, 1993 cited in Wang *et al.*, 2006). However, there are a number of risks associated with this approach. First, accountability implies incentives (positive or negative) and consequences, which in turn can work against the goal by motivating actors to distort or manipulate the outcomes. Second, standardised tests are seen as a myopic view of teaching and learning that limits what type of skills and performance are measured and undermines the use of solely standardised test results for accountability purposes. Wang *et al.* (2006) stress this point, stating: "important learning outcomes that do not render themselves easily to an external mechanism for ensuring performance must also be valued and documented" (315). Nonetheless, the risks associated with test-based accountability are often overlooked because "in a political context of tight economy and global competition, educational accountability holds great appeal to taxpayers and funding agencies" (Wang *et al.*, 2006: 316).

2.2.4 Purpose 3: Standardised tests for public information

29. Another form of accountability is providing standardised test results to the general public. This refers to the practice of publishing test results at the school level for use by parents, government officials, the media and other stakeholders. Not only does this serve the purpose of providing information on education system performance to the general public, but the results are often used by stakeholders to take action. For example, parents can use test results to make decisions on their child's schooling: "The information can be used to inform parents and communities, and in some situations, parents can use the information to make choices about school for their children" (Kellaghan *et al.*, 2009). Parents and others can also use test results to increase awareness of education issues, put education on the public agenda and compare or rank schools. For example, in Chile national assessment results are published and have contributed to placing education on the public agenda and highlighting the demand for increased equity in schooling (Benveniste, 2002 cited in Kellaghan *et al.*, 2009). In the US, UK, Australia and Korea results are available online for the general public to access. In Australia, NAPLAN test results are published on an individual school basis on the *My School* website, where the public can access performance and other data on schools across Australia. This provision of information is an important mechanism for the public to be able to hold the education system accountable and to use the information to demand improvement or other changes. Nonetheless, providing school test results to the public can have an impact on a school's ability to recruit and retain teachers and even influence housing prices in high performing school neighbourhoods (Chiodo *et al.*, 2010; Visscher *et al.*, 2010). A more detailed discussion on the impact of publishing test results can be found in Section 2.10.

2.2.5 Purpose 4: Standardised tests for formative purposes

30. Finally, standardised tests serve formative purposes. Formative assessment draws on test results in order to identify learning needs and adjust teaching accordingly (Looney, 2011). In this case, results from tests are used to 1) identify students' strengths and weaknesses or needs and 2) provide teachers with feedback on their instruction. Overall, the aim of the standardised test is to improve instruction and outcomes by gathering information on student performance. Standardised tests for formative purposes assist teachers and schools to target the learning needs to specific students by providing a snapshot of

student performance. According to Volante and Ben Jaafar, "improvement as a purpose lends itself to the idea of using LSA [large-scale assessment] to inform decisions that will yield greater learning for students" (2008: 207). Using test results for formative or diagnostic purposes can be done at the level of the student or the school: "When large-scale assessment serves as diagnostic information at the system level, educational leaders at the school or district level are expected to analyse, interpret, and reflect on the results and make programmatic decision that will yield systematic improvements in teaching and learning (Earl and Torrance, 2000 and Earl and Katz, 2006 cited in Volante and Ben Jaafar, 2008: 207).

31. Standardised tests for formative purposes provide important feedback to teachers on student performance in specific subject areas or on a student's ability to master certain types of tasks. In Sweden, for example, the national tests administered in Years 3 and 5 are for diagnostic and formative purposes to determine students comprehension of Swedish/Swedish as a Second Language, Mathematics, and English (in year 5 only) (Nusche *et al.*, 2011). In Canada, standardised tests create, enhance and apply data in order to support educational improvement (Campbell and Levin, 2008). In France, results from diagnostic tests are used to form groups of students for who personalised assistance programs are offered (Eurydice, 2009). In general, tests for formative purposes are used by teachers to define objectives, adapt or adopt teaching strategies and plan learning activities based on the assessment results (Eurydice, 2009).

32. One drawback of using standardised tests for improvement purposes can also increase the frequency of testing. In some U.S. schools, for example, standardised tests for other purposes (such as judging a student's performance or holding teachers accountable) are being used in *formative* ways. In this instance, a version of the standardised test is issued earlier so teachers can use that as a benchmark and work to improve student outcomes prior to the national standardised testing date (Stiggens, 2005). Using standardised tests for formative purposes can be a useful mechanism for teachers to improve their instruction practices and identify student needs, especially when the results are shared with students so they are aware of their own progress, but policymakers should be wary that this can also imply more frequent testing and a misalignment with test purposes when standardised tests designed for accountability purposes are used for formative purposes as well (Stiggens, 2005). Finally, it is important to note that in order to gain from such test results, the appropriate training must be provided to teachers to enable them to analyse results and test results must be returned to teachers in a timely manner so as to allow sufficient time for teachers to respond to the feedback.

33. Popham (2003) identifies five attributes of an "instructionally useful" test whose goal is improvement: significance, teachability, describability, reportability, and non-intrusiveness (cited in Independent Schools Queensland, 2010). A significant test measures a distinct curricular aim or cognitive skill and teachability refers to the ability for the measurement to be taught, rather than measuring an innate intelligence. Describability refers to the ability for teachers to use the described measures to create appropriate instructions and reportability refers to the specificity of the results and their usefulness in informing teachers. Finally, a non-intrusive test is one that does not take too long to administer, therefore avoiding intrusion on instruction time. For a more detailed analysis of standardised tests for formative purposes please see Looney (2011).

2.2.6 Debate: Single purpose vs. multi-purpose standardised tests

34. Across OECD countries, standardised tests are used to serve numerous purposes (Eurydice Network, 2009; Volante and Ben Jaafar, 2008; House of Commons, 2008). The same test that is used to determine national standards may also be used to offer rewards or sanctions to educators or it may be used to inform teachers of student strengths and weaknesses. Whether a single large-scale, standardised test should serve multiple purposes is a debated issue. The Eurydice Network's 2009 report on education in Europe states: "Assessment experts have warned that the use of a single test for several purposes might be inappropriate where the information ideally required in each case is not the same" (2009: 24). Hamilton

and Stecher also elaborate on this point, stating "it is important to keep in mind that requiring tests to serve multiple purposes sometimes results in a reduction in the utility of the test for any one of these purposes" (2002: 135).

35. The connection between the use of test results and its purpose is referred to as a test's "fitness for purpose" and is developed by Newton (2007) and others (Madaus, 1995 cited in Wang *et al.*, 2006). Newton argues that tests are designed to support certain types of inferences. He elaborates on this point by distinguishing between two types of inferences – design-inference and use-inference. Design-inference is the primary use of the test that is inherently linked to the purpose, while use-inference is where test results are used for a different task than the original purpose (Newton, 2007). When use-inferences result in actions or decisions about an education system the validity and accuracy of the results and resulting actions is questioned (Newton, 2007: 6). Newton warns that "an assessment system which is fit for one purpose may be less fit for another and could, conceivably be entirely unfit for yet another" (2007: 6). This point is echoed by the UK 'Teaching and Assessment Report', which states: "The instrument [national test] will not necessarily be fit for any other purposes for which it may be used and, if it is relied upon for these other purposes, then this should be done in the knowledge that the inferences and conclusions drawn may be less justified than inferences and conclusions drawn from an assessment instrument specifically designed for those purposes" (House of Commons, 2008: 17). The same document highlights the issue in the case of the UK, where national tests are used for a variety of purposes that span across national, local, school and individual levels and formative, summative, evaluative and diagnostic uses (House of Commons, 2008: 15).

36. Another example of this dilemma is the case of Canada, where standardised tests are employed as a policy instrument with multiple purposes. In Canada, standardised tests seek to promote consistency, standards and school improvement across jurisdictions, the standardised test system has been expanded to include quality control and accountability measures (Rogers and Klinger, 2006 cited in Volante and Ben Jaafar, 2008: 206). Volante and Ben Jaafar highlight the discrepancy caused by the multi-purpose assessment in Canada, stating: "The testing structure that is developed for system accountability and those that lend themselves to improvement are not one and the same" (2008: 207). The incompatibility of certain purposes is an important note to take into account, as many national test systems aim to simultaneously hold schools accountable and promote student learning, which are two conflicting goals (House of Commons, 2008).

37. Multi-purpose standardised tests can increase ambiguity in the validity of results, which can undermine the assessment system. In Denmark, teachers have expressed concern over the validity of using test results – that were initially developed for accountability or control mechanisms – for diagnostic purposes (Shewbridge *et al.*, 2011: 57). Since test purpose feeds into the design of the test (including whether the test is given to a cohort or sample of students) it is imperative that purpose and use are clearly indicated so as to avoid inappropriate and invalid uses of results (Eurydice Network, 2009: 24).

2.3 Test design and development

38. This section describes various decisions and components associated with designing a standardised test including the scope of the test, developing test questions and determining the frequency and timing of the standardised test.

2.3.1 Scope

39. The scope of the test refers to which students and what skills are tested. In most OECD countries, standardised tests assess students' attainment of basic skills, such as literacy and mathematics. It is less common for large-scale standardised tests to assess 21st Century skills, such as ICT or science literacy,

although this will likely be an increasing trend in the coming years. It is most common for the subjects to be aligned with the national curriculum, as is done in the UK, Portugal, Belgium (Flemish community), and Austria (Eurydice Network, 2009). Countries that monitor beyond basic competencies in language and mathematics, often do so on a rotating basis, such is the case in Finland, Belgium and the United States.

40. Typically, large-scale tests for monitoring, evaluation, accountability or informational purposes test younger students – between Years 3 and Year 9 of school (see also Appendix A). In Europe, it is most common to administer national tests initially in Year 4, as done in Spain, Austria, Hungary, Portugal, and Iceland (Eurydice Network, 2009). Typically, assessments are determined by grade or year level, rather than student age, as is the case in some international surveys, like PISA. When assessments are done at the primary level, subject areas emphasise language skills like reading and writing as well as mathematics. As students reach higher levels of schooling, standardised tests hold stakes for students, and therefore are outside the scope of this review.

2.3.2 Sample vs. census-based assessments

41. In addition to determining what grade level of students will be assessed, a government must also decide whether the test will be sample or census based. According to Greaney and Kellaghan, "Most national and all regional and international studies use sample-based approaches in determining achievement levels" (2008: 32). Some national tests, such as those in France and Mexico, use both census- and sample-based approaches (Greaney and Kellaghan, 2008).

42. Testing only a sample of the student population is favourable if the primary goal is to gain information for system evaluation or policy purposes (Greaney and Kellaghan, 2008). Sample-based tests are less costly and can have greater accuracy when accompanied by more intense data preparation and analysis (Greaney and Kellaghan, 2008). However, sample-based assessments only allow for assessment at the system-level and it does not identify certain schools that may need attention (Greaney and Kellaghan, 2008). Further, sample assessments may de-motivate students to perform, if the test is perceived as very low stakes for the selected test takers. Sample-based assessments are administered in Belgium (Flemish Community), Hungary, Spain, France, Austria, and Finland (Eurydice, 2009).

43. A census-based assessment lends itself to accountability systems, where sanctions are involved, or to systems that aim to identify schools in need of assistance (Greaney and Kellaghan, 2008). Examples of census-based assessments include the United States' NAEP test, which is required of each student in grades 3-8, and the SIMCE program in Chile, which is required for 4^{th}, 8^{th} and 10^{th} grade students (Toch, 2006; Meckes and Carrasco, 2006). Census-based assessments are also administered in Denmark, Ireland, Italy, Hungary, Portugal, and Sweden (Eurydice, 2009). Such an assessment is advantageous because it allows for direct comparisons of schools, it allows parents to judge the effectiveness of individual schools or teachers and it helps ensure that students reach a certain standard in performance (Greaney and Kellaghan, 2008). Disadvantages associated with the census-based assessment system is that it can lead to unfair ranking of schools, it can lead to cheating and test manipulation by school administrators or teachers, it leads to unfair assessment of effectiveness of the system based on test score performance only (Greaney and Kellaghan, 2008).

2.3.3 Question development and choice

44. Across OECD countries, test questions are developed by test agencies, companies or educational ministry branches which conduct pre-tests and field trials to ensure that the questions gather the intended information and align with curriculum goals and other standards. A technically sound process for developing the test feeds directly into the quality of the test and its results (Toch, 2006). Test developers

are typically psychometricians trained in measurement theory and statistics to enable them to accurately balance the different technical needs of standardised tests (Toch, 2006).

45. Often test questions are taken from a large test-bank of questions that have been pre-validated and refined. In Denmark, for example, large-scale field trials were performed to develop the national test item bank of 7 200 items, 10% of which are renewed each year. Test items were given to 500-700 students in trials to validate their appropriateness against psychometric scales and other comparisons (Shewbridge *et al.*, 2011: 67). In the U.S., test making companies call on curriculum experts to ensure questions are aligned with state standards and field-test questions on thousands of students to ensure that questions accurately measure student's abilities (Toch, 2006). Test development is an extensive process, as highlighted by Shewbridge *et al.*, "Validation [of test questions] is a long-term process of accumulating, interpreting, refining and communicating multiple sources of evidence about appropriate interpretations and use of test information" (2011: p. 67).

46. Below a description of different types of test questions is provided, as well as a brief discussion on the advantages and disadvantages of each. It should be noted that often national tests are a mix of question types (Eurydice Network, 2009).

2.3.3.1 Close-ended questions

47. Many national tests are made up of primarily close-ended questions, such as multiple choice questions, true-false or short fill in the blank tasks. Close-ended questions are advantageous in that they are less costly to develop, administer and score, scoring is more reliable, test results are very comparable, and such questions can be used to test a wide range of outcomes (Hamilton and Koretz, 2002; Anderson and Morgan, 2008; Zucker, 2003). While being highly reliable and comparable, multiple choice questions can be limiting in that they do not test critical thinking or problem solving skills and it is argued such questions encourage surface learning and rote recollection, rather than deep, cognitive processes (Zucker, 2003; Toch, 2006; Anderson and Morgan, 2008; Nusche, forthcoming). Rather than testing thinking skills, multiple choice or other close-ended questions test content only (Nusche *et al.*, 2011).

2.3.3.2 Open-ended constructed response questions

48. Some national tests incorporate a number of open-ended constructed response questions, where the student is instructed to provide a written response, orally respond, solve a problem, or demonstrate a process. Such questions or tasks are hand scored using a rubric, which can be a disadvantage as it is more costly, time consuming and decreases the reliability of the score (Toch, 2006). Such questions are advantageous in that the student is required to recall information by themselves and they can provide more 'sophisticated evaluation' of student performance than multiple choice questions (Anderson and Morgan, 2008). Also, as seen in the case of Sweden, such questions have the opportunity to be more "effectively aligned with curricula that emphasise development of higher order thinking skills and capacity to perform complex tasks (Nusche *et al.*, 2011: 48). In Sweden, national assessments include open-ended performance questions, such as written essays, oral communication skills and collaborative problem solving (Nusche *et al.*, 2011). In the United States a different trend is occurring. Toch (2006) points out that in the U.S. states are moving towards more multiple choice questions due to the cost and time of scoring constructed response questions.

2.3.3.3 Standardisation of Test Questions

49. The agency administering the national test must decide whether to apply the same test questions across the national sample or whether different test questions will be used in a single testing year. The Eurydice Network notes, "The extent to which countries [in Europe] include identical questions in a given

national test varies" (2009: 34). This can vary by country and within countries depending on the national test. Reasons for differentiating tests can include: desire to prevent cheating, attempt to account for learning differences or methodological concerns regarding the evaluation of tests (Eurydice Network, 2009: 35). In Denmark, for example, tests questions are personalised for each student to account for learning differences. By computerising test administration, as a student answers a question, the following question is chosen based on whether their previous response was correct or not (Eurydice Network, 2009).

2.3.4 Frequency and timing of testing

50. The frequency and timing of standardised tests are closely linked to the purpose of the national test. Tests for monitoring purposes can be administered less frequently than tests used for accountability purposes, which are often administered each year. In Chile and England, for example, accountability tests are administered each year (Greaney and Kellaghan, 2008). Finally, if the purpose is to gain information on the system's performance, an assessment of a sample of students in a particular curriculum area every three to five years typically may suffice (Greaney and Kellaghan, 2008). In the US, reading and mathematics are tested every other year, while other subjects are tested less frequently. Korea administers the national test in two subject areas that vary each year (INCA, 2011).

51. The frequency at which national tests are administered varies greatly across OECD countries. Denmark, the UK and France have frequent standardised tests, whereas Belgium, Germany, Spain and the Netherlands administer national tests less frequently (Eurydice Network, 2009). According to the Eurydice Network, the European trend is to administer national tests in 2 or 3 specific school years during compulsory education (2009: 26). Regarding the frequency of tests, governments need to balance two aims: first, there is the desire to have an up to date picture of the education system; second, too frequent tests can increase the burden on teachers and students, reduce teaching time and increase costs (Eurydice Network, 2009). According to Greaney and Kellaghan, over frequent assessment limits the impact of results and is more costly (2008).

52. Like frequency, timing is closely linked with the test purpose. Tests that seek to identify learning needs are often administered at the start of the school year, as is the case in France, Luxembourg and Iceland. In accountability systems, tests are more often held at the end of the year. An important issue with regard to timing is when results are returned to teachers or school administrators. The utility of the test results can be undermined if the results are not provided in a timely manner. For example, the NAPLAN national test in Australia is administered in the autumn and results are provided the following spring. The delay in results makes it difficult for teachers to use results to inform planning in the school year (Santiago *et al.*, 2011). In the case of Denmark, the use of ICT in testing allows for teachers to receive rapid feedback of test results, which in turn encourages the use of results to improve teaching and learning (Shewbridge *et al.*, 2011). As noted in Shewbridge *et al.* (2011), the rapid turnaround of test results in Denmark, "…is in strong contrasts to several national test systems where educators receive student test results several months after the test was administered." Finding the right timing for the test administration and marking can therefore be difficult, given that tests need to be processed and analysed externally before being released to schools.

2.4 Reference standards

53. This section describes the basis by which standardised test results are analysed. In the literature, three methods for examining test results are identified: norm-referenced, criterion-referenced (also referred to as standards based), and growth measures. Reference measures have implications for the design of a test and how a test achieves its purpose.

2.4.1 Norm-referenced testing

54. Norm-referenced assessments – such as those administered in Korea, Mexico and the U.S. – compare students amongst each other and rank a student's proficiency in relative terms (Toch, 2006; Zucker, 2003; INCA, 2011; Ferrer, 2006). Norm-referenced tests compare test results to the results of a reference group that has taken the same test or groups as large as entire school systems are compared. According to Zucker (2003), "Norm-referenced tests are typically designed to cover a broad range of what test-takers are expected to know and be able to do within a subject area" (5). Such tests reveal whether a student is progressing at a slower or faster rate than other children (Hamilton and Koretz, 2002). Norm-referenced reporting is criticised as failing to provide evidence about a student's level of mastery of knowledge and skills [*i.e.* educational outputs], since it focuses on providing an indication of the performance of a student relative to other students (Hamilton and Koretz, 2002). It is for this reason that many standardised tests either report results using both norm- and criterion-referenced methods, the latter of which is described in more detail below.

2.4.2 Criterion-referenced testing or status measures

55. Criterion-referenced tests measure whether a student has met a specific target or performance level (Zucker, 2003; Toch, 2006; Nusche, forthcoming; Hamilton and Koretz, 2002). This same concept is also known in the literature as status measures or standards-referenced, as performance is judged against predetermined standards or target levels (Figlio and Loeb, 2011; Linn, 2005). Systems that look for whether a student attains a certain proficiency level or measures performance based on the average school test score would be examples of status measures assessments. The U.S. NAEP test administered under No Child Left Behind follows a status measure/criterion-referenced approach as does the National System for the Assessment of Educational Quality in Chile and the national tests in Sweden (Ferrer, 2006; Nusche *et al.*, 2011). Criterion-referenced tests are becoming more popular in state administered tests in the U.S., as well, as states can have tests customised to measure whether students are meeting the state-wide standards (Toch, 2006).

56. Criterion-referenced evaluations set a minimum achievement bar, and schools are encouraged to improve student outcomes at least to that level (Krieg, 2008, Neal and Schanzenback, 2010 cited in Figlio and Loeb, 2011). The advantage of this approach, Figlio and Loeb state, is that "…it encourages schools to focus attention on the set of low performing students who in the past may have received little attention" (2011: 392).

57. The challenge for criterion-referenced testing lies in determining the standards and targets. A number of questions need to be answered: What kind of methodology should be employed? How many standards should be created? How is proficiency defined and at which point should the standard be set? Based on the approach to developing standards, establishing a point of reference or a target can have different implications. For example, by establishing a level at which a student is deemed "proficient", an education system provides incentives to schools to focus on bringing students to that level, rather than equitably focusing on over-all student improvement. The problems with consistently applying standards across a nation have recently come to light in the U.S. The National Center for Education Statistics recently mapped the proficiency standards across U.S. States, finding a wide variation in "proficiency" levels across U.S. States and finding that in some cases state "proficiency" levels are even below the national standard of "basic" (NCES, 2011a).

58. Based on the level of the performance threshold, schools and teachers will respond by focusing efforts on different student groups. There may also exist the incentive to set *lower* standards in order to show more progress or performance output; this can especially be the case in systems where test results hold schools and teachers accountable (Toch, 2006). Another disadvantage to the criterion approach is that

the ability of a school to meet standards is a function of factors beyond simply school performance, such as prior achievement levels, and family and community characteristics (Meyers, 2000 cited in Linn, 2005).

2.4.3 Growth or value-added measures

59. A third means of analysis described in assessment literature is growth measures – also called value added or improvement measures. This refers to when results are evaluated on the degree to which student improve over time, for example from fall to spring of a specific year (Figlio and Loeb, 2011). Results may also be adjusted to control for conditions outside teacher control, thereby examining solely the 'value-added' by the teacher (Heyburn *et al.*, 2010 cited in Rosenkvist, 2010). Another application of the growth model is to compare the performance of successive cohorts of students at the school (Linn, 2005).

60. The growth model is advantageous because it encourages schools to focus on improving absolute student performance; hence, some argue this approach is appealing because it is more fair (Figlio and Loeb, 2011). Since the growth model rewards schools for improving student outcomes in any case, there is no desire to focus solely on certain subgroups, like is the case in criterion-referenced systems. Additionally, some researchers claim that by encouraging teachers to improve student outcomes outright, the likelihood of a teacher engaging in distortionary or strategic behaviour is reduced (Heyburn *et al.*, 2010 cited in Rosenkvist, 2010).

61. Figlio and Loeb point out that the growth model can raise political concerns, as it is a less transparent evaluation technique and "some see it as a way of letting schools with low average performance off the hook" (2011: 392). Another potential disadvantage of the growth model is that it fails to recognise changes in student characteristics over the years, for example the number of students transferring into a school (Linn, 2005). Figlio and Loeb summarise the issue appropriately, stating: "Neither the status nor the growth approach to measuring school performance perfectly captures school efficiency – the effectiveness with which schools use their resources to maximise student outcomes, given the students they serve" (2011: 392).

2.4.4 A note on aligning standards to curriculum

62. In order for standards-based systems and tests to be effective, it is critical that the development of standards is in line with student instruction and the national curriculum (Hamilton and Koretz, 2002). In essence, if the goals and expectations of the assessment are not aligned with what happens in the classroom or with teacher's curriculum goals, the system will not effectively measure student performance. Due to the fact that curriculum, instruction and assessment are interdependent, it is important for a government to clearly define education standards or objectives that are aligned with the curriculum (Shewbridge *et al.*, 2011; Alliance for Excellent Education, 2010). Literature and research has shown that if curriculum, instruction and assessment are not aligned student achievement is compromised (Baker and Linn, 2000 cited in Shewbridge *et al.*, 2011).

63. This issue is described as 'system coherence' in the Alliance for Excellent Education's 2010 Policy Brief. The author's argue that the notion of 'alignment' is a key component of criterion-referenced assessment systems and misalignment can have serious consequences on instruction and learning (Alliance for Excellent Education, 2010: 3). One negative externality of disconnected curriculum and standardised tests is termed "teaching to the test", where teachers may emphasise test taking skills and low-level content, rather than "important learning goals expressed by the standards" (Alliance for Excellent Education, 2010: 3). Education standards, when used to guide curriculum development and external standardised assessments, can reinforce instruction that "aims at higher levels of cognitive complexity as well as basic skills and knowledge" (AEE, 2010: 3).

2.5 Use of ICT in student standardised testing

64. Information and communication technology (ICT) has historically been used to mark standardised tests and, more recently, ICT is used to administer or even shape the standardised test. Nonetheless, the use of ICT in testing varies greatly across OECD countries. For example, the Netherlands and Norway administer tests on computers, whereas other countries, such as Belgium (French & Flemish Communities), France, Austria and Luxembourg use ICT primarily to score tests (Eurydice Network, 2009). Implementing ICT for standardised testing can be challenging because it is costly and requires a high level of technical training for teachers and others.

65. ICT can be used to create, administer, and score tests and to analyse and disseminate test results. There are four primary reasons for using ICT in standardised testing: 1) reduce scoring errors and costs; 2) incorporate novel item formats; 3) improve test reports; and 4) adapt the test to the examinee's proficiency. Each of these is described in further detail below.

66. ICT can increase the efficiency of an assessment system by reducing test administration and scoring errors and by reducing costs (AEE, 2010; Eurydice Network, 2009). While a country would incur a substantial initial cost for implementing computerised testing or scoring, computerised scoring and implementation would eventually reduce costs associated with paying teachers or other experts to score tests. ICT can also improve the accuracy and validity of testing by expanding the types of questions that are asked. As stated in the Alliance for Excellent Education's 2010 Policy Brief, "…computerised testing enables the use of simulations, animations, and other techniques that offer opportunities for students to engage in complex tasks that are unlikely in a paper-and-pencil setting" (AEE, 2010: 9). Such item formats are typically not feasible or too costly without the use of ICT (Hamilton and Stecher, 2002).

67. ICT can improve test reporting processes, making test results more accessible and useful to a variety of stakeholders. Results can be downloaded directly by teachers and administrators in easy-to-analyse formats. A further benefit is that results and analysis can be readily available to other stakeholders and results can be linked to instructional tools that provide guidance for teachers and principals about how to improve student outcomes (AEE, 2010: 9).

68. An innovative use of ICT is to use computerised systems to adapt the test to each student's proficiency. This is known as computer-adaptive testing (CAT); this type of national test is currently used in Denmark. Through CAT, each test is "geared to individual levels of ability" (Eurydice Network, 2009: 36). As a student responds to a question, the following question is chosen based on whether the answer was correct or incorrect. A correct answer prompts a more difficult question, whereas an incorrect answer prompts an easier question. Therefore, each item's difficulty corresponds to the students' proficiency. CAT demands technological capacity, as each test must have a large item bank with questions that vary in difficulty; however, the benefits can be great as CAT provides more accurate measures of student performance since it discerns each individual's level of achievement more efficiently (AEE, 2010: 9). In Denmark, adaptive tests are said to provide a "very accurate diagnosis of student performance" that, in addition to accuracy, is also aligned to national objectives (Shewbridge *et al.*, 2011).

2.6 Implementation and scoring

69. In many OECD countries, national tests are administered by teachers, although in some cases external administrators or computers are used. The choice of who should administer the test has implications for the validity and reliability of the assessment. In some cases, separation of teacher and assessor is seen to reduce validity if "teachers' curriculum choices, instruction and guidance do not properly match the expectations of the external test" (Nusche, forthcoming).

70. As discussed in detail above, tests are often scored mechanically, which is very reliable and results can be turned around quickly. In some countries, national tests are scored by teachers, such as is the case in Sweden. The benefit to this approach is that teachers have an incentive to score tests quickly and the scoring is seen as valid, since it has been done from the instructor's perspective. In turn, when scores are seen as valid, teachers are also more likely to use the results (Nusche, forthcoming). The drawbacks to this approach is that it can, at times, be more costly and time consuming than mechanical scoring and scores are less reliable, that is the same performance might not receive the same score (Nusche, forthcoming).

2.7 Limitations of standardised tests

71. While student standardised testing can be a valid and reliable means of monitoring the education system, gathering information on student performance and/or holding schools accountable, the literature also reiterates that there are a number of limitations to standardised tests which weaken the capacity to achieve their purposes. Primarily, standardised tests are limited in scope both in terms of the breadth of their reach and in terms of their depth of assessment.

72. National regulations of standardised tests may not apply to all schools, leading to an uneven assessment of the national student population. In many OECD countries public schools or schools which receive public funding are required to administer standardised tests for a range of purposes. However, this implies that some student bodies, such as those attending private schools, are not monitored. Assessing student outcomes across all student groups would be meaningful if the test is to be used to monitor a nation's educational performance or to compare student outcomes. In Denmark, for example, only public schools are required to administer the national test, which limits the test's value in terms of monitoring national goals; on the other hand, in Australia both government and non-government schools administer the NAPLAN assessment (Nusche *et al.*, 2011; Santiago *et al.*, 2011).

73. Student standardised tests are also limited in the kind of knowledge they assess: the depth of assessment is limited. Test results are only available for certain student populations and for specific subjects. As stated previously, many standardised assessments in OECD countries assess mathematics and reading skills and, due to the nature of a standardised test, the tests often cannot test for critical thinking, analytical or problem solving skills. While narrowing the scope of tests to ensure that basic skills are assessed can have positive effects, it can also negatively impact the students' opportunity to deepen his or her knowledge of other important subject areas, such as science, history and civics. One major drawback of standardised tests is directly related to the limited subject focus – by attaching greater importance to certain subjects, like reading and mathematics, through standardised assessments, school systems are inadvertently corrupting a teacher's motivation to equally concentrate on teaching other subject areas, such as science. This issue of narrowing the curriculum to accommodate standardised test subjects is explored further in Section 4.2.2.

74. According to the literature, one way to counteract the limitations of standardised testing is to implement other monitoring tools to complement the national test (Harlen, 2007; Hamilton and Stecher, 2002; Guilfoyle, 2006). For example, with the introduction of the NAPLAN national assessment in Australia, some states and territories also implemented standardised assessments in other subject areas that were not covered by NAPLAN, such as science, society and environment (Santiago *et al.*, 2011). By complementing the literacy and numeracy focus of NAPLAN, other assessment tools motivate Australian schools and teachers to provide a balanced curriculum.

75. Further, test quality is inherently linked to test design and use of test scores; hence, if a test score is used for a different purpose than was initially envisioned, the quality of the test is questioned (Le and Klein, 2002). The ability to judge the quality of a test rests on a number of suitability criteria, *i.e.* validity,

reliability, fairness, utility, comparability and equity. Most commonly, standardised tests are analysed based on whether the test is seen as a valid tool for gathering data and whether the test is reliable, meaning the score is relatively free from 'chance' effects (Le and Klein, 2002). The two criteria of validity and reliability are discussed in more detail below. The trade-offs and debates associated with designing a test that is valid and reliable is a prevalent discussion in standardised test literature.

2.8 Suitability criteria

76. The following section reviews the different suitability criteria used to evaluate standardised tests. The section begins by defining the two most prominent criteria identified in the literature: validity and reliability. Each criterion is presented, followed by a brief discussion on the factors that influence the validity and reliability of a standardised test. Subsequently, the issue of balancing validity and reliability is discussed, followed by an introduction of other suitability criteria: comparability, utility and equity.

2.8.1 Validity

77. Whether a standardised test is valid or not is an important mechanism to judge whether a country's education evaluation system is functioning and will continue to function. Designing and implementing a standardised test that is valid is a challenging task across OECD countries. According to Eurydice 2009, "A key issue [with regard to standardised tests in Europe] is the need to ensure the validity and fitness-for-purpose of national tests, including their technical accuracy, objectivity and cost-effectiveness" (63). As seen through the discussion below, testing criteria are deeply intertwined and often require trade-offs between validity, reliability, test format and costs.

78. Validity refers to a test's ability to measure what it sets forth to measure and the ability for scores to accurately inform decision making (House of Commons, 2008; Cronbach, 1971 and Messick, 1989 cited in Le and Klein, 2002; Harlen, 2007). According to Harlen (2007), validity refers to "what is assessed and how well this corresponds with the behaviour or constructs that it is intended to assess" (Harlen 2004: 25, cited in Nusche, forthcoming). Validity is closely related with the intentions of a test and whether those intentions are carried out, which further highlights the importance of test development and of clearly establishing the purpose of the test and its results. If a test is not valid, the score will not be a meaningful inference for policy makers to use in decision-making (Santiago *et al.*, 2011).

79. Test validity is compromised when the purpose of the test and the anticipated use of test results is unclear to different stakeholders. In Denmark, for example, OECD reviews have shown that national tests validity is challenged "due to the lack of clarity over the purpose of the tests as communicated by different stakeholders...specifically educators' fears that results will be used to hold them directly accountable" (Shewbridge *et al.*, 2011: 120). In this case, if educators are uncertain of how test results will be used, teachers are less likely to use the test as a monitoring and pedagogical tool; essentially, the under-use of test results or the intense focus on tested subjects undermines the validity of the assessment system (Shewbridge *et al.*, 2011: 120). Similarly, validity is reduced if the test is designed by experts other than teachers, since curriculum and instruction may not properly align with the externally developed test (Nusche, forthcoming).

80. The validity of a standardised test is also undermined if scores fail to capture a true representation of students' performance in a specific subject area, or content domain. As stated by Le and Klein (2002), "If a test fails to capture important elements of the domain, scores can only justify narrow or qualified conclusions about performance" (62). This concept is called 'construct under-representation' (Le and Klein, 2002). Whether a test accurately represents a subject area or domain will feed into its validity as an evaluation mechanism of performance; hence, the choice of test items in test development is critical to the

test's validity and test development must include a clear definition of the subject being assessed (Harlen, 2007).

81. According to the literature, validity of large-scale, standardised tests – specifically those used to assess program effectiveness – is increased through matrix sampling. Matrix sampling refers to administering different sets of questions to different students (Le and Klein, 2002). Administering a standardised test through matrix sampling is said to increase validity because it allows for a more comprehensive evaluation of student performance and knowledge without having to increase testing time (since each student is not responding to every question). NAEP, the national, standardised test in the U.S., applies matrix sampling. It should also be noted that while matrix sampling produces more valid scores because each individual student is answering different questions, such scores are not reliable to make decisions about individual students (Le and Klein, 2002). Essentially, matrix sampling requires individual students to answer only a small proportion of the entire content and subject areas that the complete test is assessing.

82. Finally, the validity of the evaluation system as a whole can be increased by using multiple assessment measures. According to Hamilton and Stecher (2002), multiple measures can refer to administering multiple tests or to including non-test data about students into the decision making process based on test scores. By using multiple measures, validity is increased since a wider range of student outcomes are assessed, which in turn decreases the likelihood of narrowing the curriculum (Hamilton and Stecher, 2002). The concept of relying on multiple assessment measures will be discussed in further detail in Section 5.6.

2.8.2 Reliability

83. A second criterion for evaluating standardised tests is reliability. This concept refers to a test's ability to consistently produce the same outcome for students over repeated occasions. In other words, it refers to the degree to which a test's scores are free from 'chance effects' (House of Commons, 2008; Nusche, forthcoming; Le and Klein, 2002). Reliability can be measured by the extent to which the test, if repeated, would produce the same results (Harlen, 2007). Reliability differs from validity in that reliability looks for consistency of results and the consistency in using results to make judgements, whereas validity looks at the inferences made from test scores and whether they are accurately drawn based on the test's purpose. Often externally developed standardised tests are seen as highly reliable, in that they produce consistent results (Nusche, forthcoming). External tests are criticised in that they reduce the validity of the test, presenting trade-offs for test developers between reliability and validity which is discussed in the following section.

84. Reliability of a test can be influenced by a variety of factors, which can be organised into four classes: item sampling, transitory variables, rater agreement, and test length and format. When a student's score fluctuates as a result of the particular version of the test, such inconsistency in item sampling reduces the test's reliability (Le and Klein, 2002). Transitory variables, such as the student's health on test day, anxiety levels, the quality of the test booklet or teacher encouragement also affect the reliability of a score. Those scoring the tests (the raters) should be able to reproduce scores across the tested population. However, raters can differ from one another in that they can disagree on student scores, undermining the reliability of a test. The detrimental effect of rater disagreement on reliability can be especially challenging when national tests are administered and scored by teachers, as is the case in some OECD countries, such as Sweden (Nusche *et al.*, 2011). Studies have shown that rater disagreement is reduced through extensive training of raters and the use of rubrics, or scoring guides (Shavelson, Baxter and Gao, 1993 cited in Le and Klein, 2002). With regard to test length, the literature finds that longer tests tend to produce more reliable scores than shorter tests (Le and Klein, 2002). The test format – or the types of questions on the

test – also affect reliability. For example, multiple choice questions are highly reliable, yet they are less valid in that they cannot assess certain areas of knowledge.

85. Black *et al.* argue that reliability can be increased by increasing testing time and narrowing the range of question types and topics tested (cited in House of Commons, 2008). However, increasing the testing time simultaneously reduces the amount of instruction time, which can be seen as detrimental to parents and teachers alike. Narrowing the range of topics and question types also is seen as reducing the validity of the test and as limiting the evaluation of student performance (House of Commons, 2008).

86. It is important to evaluate a test's reliability in the test development stage, since the format and length of the test feed into the test's reliability. Further, test developers must seek to create the appropriate balance of reliability and validity which is a function of the purpose of the test. For low-stakes, standardised tests it is important that the test is a valid representation of the student population, yet it is less important that the score is a reliable assessment of the individual students' performance (Le and Klein, 2002). Le and Klein articulate this point, stating: "Scores that are used to make decisions about individual students [*i.e.* high stakes tests] will require higher levels of reliability than scores that are used to make decisions about educational programs" (2002: 59).

2.8.3 Validity & reliability trade offs

87. As seen in the discussion above, often test development involves certain trade offs between the reliability and validity of a test. It should be noted that each decision made in the development process – regarding the test length, content coverage, format, administration and scoring procedures and others – will affect the reliability and validity of the test, along with other criteria and variables such as test time and costs (Le and Klein, 2002). For example, while multiple choice questions are a highly reliable assessment tool, they simultaneously reduce the validity of a test as they limit the scope of content area assessed; on the other hand, whereas essay questions increase test validity, they reduce reliability and increase costs as they are difficult and costly to score. If an assessment aims for maximum reliability it will inherently have limited validity, as it will narrowly focus on testing for factual knowledge and tangible learning outcomes, rather than higher-order reasoning or critical thinking skills (Harlen, 2007 cited in Nusche, forthcoming). This dilemma has been articulated by researchers in the United States, who have found that open-ended, performance-based assessments often do not measure the skills intended, limiting their validity (Baxter and Glaser, 1998; Hamilton *et al.,* 1997; Pellegrino *et al.*, 1999 cited in Nusche, forthcoming). The type of question, therefore, has a direct impact on the reliability and validity of the test.

88. Another trade off is a function of the purpose of the test, as articulated by Hamilton and Stecher (2002) in reference to tests that are used for accountability purposes: "...accountability often requires trade-offs among competing values" (122). For example, policy makers must balance the desire for more reliable test scores (which derive from longer tests) against the concerns of teachers and parents that excessive classroom time is being consumed by testing. There is not a consensus in the literature on how to overcome the reliability-validity trade-off; rather, academics stress the importance of being aware of the factors and decisions that influence the two criteria.

2.8.4 Other suitability criteria: Comparability, utility, equity

89. Although the literature on standardised testing is often dominated by discussions on the validity-reliability trade off, other testing criteria should also be noted to give a complete picture of evaluation aspects that OECD countries aim to achieve.

2.8.4.1 Comparability

90. When standardised tests are used to evaluate the education system as a whole, it is essential that the test and its scores are comparable across the nation's student population. This can imply comparability across sites (schools) and across time, both within and across years. The degree of comparability influences the test validity, as well. According to some scholars, tests can be comparable across various elements, such as purpose, test content, test administration and consequences of testing (Zhang, 2008). As assumed, comparing and linking these elements becomes more difficult when analysing two distinct tests, such as comparing a national test to state tests (Zhang, 2008). Therefore, by ensuring test elements such as purpose and administration are consistent across schools, comparability, and thereby test validity, is enhanced. Some assessment systems even choose incorporate the degree of comparability into the analysis of a system's validity, such is the case in Australia (Santiago *et al.*, 2011).

2.8.4.2 Utility

91. Utility refers to the ability for teachers and other stakeholders to use and interpret results. The utility of a test refers to the tests ability to provide the intended feedback and impact. In some cases, utility is undermined if test results are delayed, which can limit how test scores inform classroom planning or curriculum adjustments. For example, the Australian NAPLAN test is administered in autumn, yet results are not available for teachers until the spring. Similarly, in Sweden tests are administered in the spring, as required by law; results are then available only late in the school year. Such delays in result delivery reduces the utility of the results for teachers, as often results are used for formative purposes and to provide teachers with feedback on instruction and curriculum. While this delay reduces the utility of the test for formative and diagnostic purposes, it must also be noted that this delay does not necessarily reduce the utility of the test as a monitor of the education system as a whole (Santiago *et al.*, 2011).

2.8.4.3 Equity

92. An equitable (or fair) test is accessible, fair and sensitive to a range of student abilities and skills so as to provide an equal opportunity for students to perform well. OECD countries are increasingly diverse and standardised tests often aim to cater towards this diversity; as stated by Le and Klein: "Unrelated characteristics of the test-takers, such as gender, ethnicity, or physical disabilities, and differences in administrative conditions should not affect the scores test-takers receive" (2002: 68). Le and Klein further articulate this point by stating that a fair testing system accounts for three conditions: 1) test items are free of bias; 2) students must have equal opportunities to demonstrate skills; 3) students must have 'sufficient opportunity' to learn the tested material (2002: 68). When these conditions are not met, often students that represent racial minorities or indigenous populations or students whose mother tongue is not the language of instruction are less likely to perform well on such standardised tests. Test developers must consider that certain questions may give an unfair advantage to certain students. For example, The Melbourne Declaration in Australia sets forth the government's goals to promote equity and excellence; however, in developing the national test, NAPLAN, test items may function differently for Indigenous students reducing the equity of the test (Santiago *et al.*, 2011). The concern over the fairness of the test for Indigenous populations should be addressed in design and development stages. In Denmark, the view of 'education of equity and inclusion' forms the foundation of the assessment system and a special test is offered for Danish as a Second Language to accommodate bilingual students (Shewbridge *et al.,* 2011). Other aspects of a standardised test that can reduce the equity include situations where some students are not given the appropriate time allotted to take the test, when test administration conditions differ or if subject matter or questions are biased towards certain ethnic groups or economic levels.

93. The suitability criteria covered in this section – validity, reliability, comparability, utility and equity – are not exclusive, however, they represent a majority of the discussions in the literature as well as

an important debate for test development with regard to balancing the criteria in connection with the standardised tests' purpose.

2.9 Use of standardised test results

94. Whereas the preceding discussion centred on the development and administration phases of standardised testing, the following focuses on the use of the test results. For a more complete synthesis of the literature surrounding using student test results, please refer to OECD Education Working Paper No. 54 by M.A. Rosenkvist (2011). Use of test results, as stated earlier, should be linked to the test's purpose to maximise their validity; however, a review of OECD practices has shown that test purpose is not always clearly stated allowing test results to be used for unrelated purposes.

2.9.1 Test results are used to identify learning needs and instruction

95. One way OECD countries use standardised test results is to inform classroom instruction and to develop curriculum. Results in this case are used by teachers and school administrators and serve as a feedback mechanism to determine which students may require extra attention and whether a teacher's approach to certain subjects is achieving the curriculum goals. In this instance, the use of results has low stakes for both students and for teachers. No repercussions are attached to the results; rather, teachers are encouraged to use the results to guide their instruction practices. Using standardised test results in this manner is often attached to tests which aim to provide information or diagnostics about the education system (see Purpose 3, Section 2.2.4).

96. In Mexico, test result reports are given to schools with the aim of improving teaching (Ferrer, 2006). In Sweden, national tests are administered in Years 3, 5 and 9 and in upper secondary school; only the national assessments in Years 3 and 5 are intended for diagnostic and formative purposes, whereas others are summative and hold stakes for the students (Nusche, forthcoming). In Australia, schools are provided with detailed reports on student results of the NAPLAN test and schools are expected to use the results for formative and diagnostic purposes (Santiago *et al.*, 2011). Specifically, results are reported against the national minimum standards, which assist stakeholders (including policy makers, school administrators, teachers and parents) to monitor student progress. Canada's provincial assessments aim to provide teachers with feedback on how well students are meeting curriculum goals and how effective specific teaching strategies are in meeting student needs (Rosenkvist, 2010).

97. The drawbacks of this approach to using test results are linked to the level of detail provided by the test results and the utility of results delivery. As discussed above, if test results are not provided in a rapid and timely manner, a significant delay can hinder the teacher's ability to utilise the results for formative purposes. Other difficulties are caused by the test design itself and results are too broad and do not provide the level of detail needed to truly respond to individual student needs, as is the case in Sweden (Nusche, forthcoming). Additionally, without adequate training, teachers may not have the assessment literacy and ability to appropriately interpret results and to identify areas where curricular strategies may require adjustment (Nusche, forthcoming; Santiago *et al.*, 2011, Alliance for Excellent Education, 2010). Training strengthens teacher's assessment literacy by improving teachers' awareness for factors that increase test results validity and reliability and by increasing their capacity to analyse and interpret data (Earl and Fullan, 2003; Fullan 2001 cited in Santiago *et al.*, 2011).

98. Using test results for formative and diagnostic purposes is therefore more fruitful when the test is specifically designed to provide detailed profiles of student performance and when results are provided in a timely manner to maximise their utility. Further, the more valid and reliable the test is, the greater the quality of information provided will be.

2.9.2 Test results are used to inform school policy

99. Typically, assessment systems with the purpose of monitoring the education system as a whole use test results to inform school policy and determine whether national standards are met. According to Eurydice (2009), "It is a widespread practice among countries in Europe to provide information enabling schools to measure themselves against the national average results achieved by pupils in national tests and to make improvements on the basis of that comparison". Eurydice finds that most European national tests are designed to monitor schools or the education system as a whole. In Sweden national tests monitor to what extent national goals are being attained (Nusche, forthcoming). By providing schools with data on where they stand in comparison with the national average or national standard, a 'mirror effect' is achieved, where schools can use the information as a basis of action to improve their own performance (Eurydice, 2009). Further, tests can monitor the performance of the education system over time and to assess how educational standards may evolve over time; however, policymakers must take into account the comparability of the test, especially if over time the standardised assessment system is changed.

100. Use of results for system monitoring can take different forms, based on whether the government gathers aggregated or disaggregated data (see Section 2.10.2). In some OECD countries, national reports are prepared which compare national test results over time and analyse other factors which may influence student performance (Eurydice, 2009). Such reports are geared towards informing policy making at a national level, steering the national debate on education, contributing towards action plans to improve the education system and highlighting differences in attainment by student groups (Eurydice, 2009). Examples of countries that actively use test results to monitor the education system include France, Belgium and Denmark. In France, conferences on the results of standardised tests are initiated at the request of various stakeholders and in the different Communities in Belgium the respective Minister of Education initiates a consultation process based on test results geared towards teachers and others (Eurydice, 2009). In Denmark, national tests results are compiled to create a national profile, which serves to monitor how Danish student performance evolves over time (Shewbridge *et al.*, 2011). Finally, it is common for OECD countries to use national standardised tests to compare regions, municipalities and other educational jurisdictions within the same system.

101. Using standardised test results to monitor the education system is beneficial in that national tests can highlight the strengths and weaknesses of an educational system. For instance, "...national tests have been an important means of drawing attention to disparities in the attainment levels of pupils and schools, as well as to factors that may contribute to such differences" (Eurydice, 2009: 59).

2.9.3 Test results are used to reward or sanction schools

102. When no-stakes, standardised tests are used for accountability purposes, results carry consequences for schools in terms of whether schools are able to meet certain performance targets or national standards (for a detailed account of trends in school evaluation in OECD countries see Faubert, 2009). In the UK, US, and Chile student test results are used to hold schools accountable and schools can face rewards or sanctions based on test results. Kellaghan and Greaney write, "In such cases, an assessment becomes a high-stakes operation for schools, with a variety of rewards or punishments attached to student performance" (Kellaghan *et al.*, 2009). Rewards can include monetary bonuses or increased resources, while sanctions include corrective measures, mandatory staff dismissal, school restructuring or even school closure.

103. Examples of education systems where standardised tests have consequences on the school level include the No Child Left Behind Act and the accompanying NAEP national test in the United States and, in Chile, the SIMCE assessment data is used to provide competitive funds for educational projects (Ferrer, 2006).

104. According to Harlen (2007) and echoed by others, such as Popham (2006), disadvantages of holding schools accountable on the basis of student test results stem from the fact that results often do not reflect the "full range of educational outcomes which a school strives for and for which it should be held accountable" (25). Further, Harlen states: "Thus, framing accountability in terms of targets for student achievement, or position in a league table of schools based on test and examination results, distorts the actions of those held accountable in ways that are not intended and are not in the best interests of students" (2007: 25). When national assessments have no stakes for students, but high stakes for teachers and schools, the evaluation system can be threatened by distortions from within. Perceiving tests as a direct measurement of school progress, teachers and administrators are often more likely to teach to the test or narrow the curriculum, leading to inflated student scores. In some more extreme cases, teachers and administrators may be inclined to manipulate student data or exclude low-achieving students from taking the test. Research regarding the causes of such negative consequences and the impact of accountability systems is becoming more prevalent (see Popham, 2006; Wang *et al.*, 2006; Figlio and Loeb, 2011) and is discussed in Section 4.2.

105. Accountability systems based on standardised test results aim to strike a difficult balance between demanding schools to take responsibility for results and haphazardly assigning responsibility to outcomes which are out of the school's control. Although holding schools accountable can – in some cases – motivate teachers and administrators, it is difficult for such systems to clearly delineate between what is the school's capacity and responsibility and what is outside the school's control. To reduce the incentive for actors within the evaluation system to distort or manipulate data, Harlen (2007) recommends: "For a more positive impact, accountability is best based on information about a range of student achievements and learning activities, judged by reference to the context and circumstances of the school and used positively to improve students' opportunities for learning" (25). This concept of adopting a multi-pronged approach to school accountability and evaluation is revisited at the end of this paper (see Section 5.6).

2.9.4 Test results are used to reward, sanction or evaluate teachers

106. Just as test results are used to hold entire schools accountable, in some OECD countries standardised test results are also used to specifically hold teachers accountable. Test results in this case are used to evaluate a teacher's performance, based on certain performance standards or gains students show on the standardised test. This is the case in some U.S. states, such as New York, Delaware and Washington, D.C. In Hungary results from student standardised tests can be used to determine teacher bonuses (cited in Faubert, 2009). It should be noted, however, that in some instances not all teachers can be evaluated based on large-scale, national standardised test results, since typically such no-stakes, standardised tests cover only a few major subjects, such as Reading, Language and Mathematics.

107. When standardised test results are used in teacher evaluation the consequences for the teacher can range from receiving a monetary bonus to losing his or her job. By attaching incentives to the standardised tests, such systems aim to increase teacher's motivation to improve student performance. However, such a system places high stakes on the tests for teachers and, as stated by the United States National Research Council, "Incentives can lead workers to perform actions that increase the performance measures, but not the underlying value of their work" (Hout and Elliott, 2011: 2). Evaluating teachers based on standardised tests which have no stakes for students is a hotly debated topic and a number of authors (such as Popham, 1999; McNeil, 2000; Smith, 1991 cited in Abrams *et al.*, 2003) have detailed the potential negative consequences of such an evaluation system.

108. Authors in favour of attaching incentives for teachers to standardised tests claim "it encourages teachers to internalise norms, values and expectations of stakeholders" and "it supports the operation of market mechanisms in the education system, involving competition, contracting and auditing" (Kellaghan *et al.*, 2009: 9). Essentially, by using test results to evaluate teachers a system of measurement-driven

instruction is created (Kellaghan *et al.*, 2009; Rothman, 1995; Hamilton, Stecher, and Klein, 2002, cited in Wang *et al.*, 2006).

109. Wang *et al.* (2006) offer a synthesis of the arguments for and against using standardised test results to hold teachers and schools accountable. Those in favour of incentivising test results for teacher, cite the links between standardised tests and improved performance, as seen in Phelps' study which concluded that countries that dropped standardised tests saw declining academic standards (Phelps, 2000 cited in Wang *et al.*, 2006). Lauded for their rigorous development standards, standardised tests allow for comparable, objective, and less ambiguous evidence of student performance than teacher-made tests (Frary, Cross and Weber, 1993 cited in Wang *et al.*, 2006). Therefore, using test results to evaluate teachers allows for the most objective assessment of student performance.

110. While incorporating competition into the education system may in some cases improve school performance, there are a number of unintended negative consequences as a result of using test results to evaluate teachers. First, a student's test performance inherently reflects a number of factors which are outside a teacher's control, including a student's home environment and previous academic background, school conditions and resources, and education policies, such as curricula and teacher training (Kellaghan and Greaney, 2001 cited in Kellaghan *et al.*, 2009; and Wang *et al.*, 2006). Since education is a production of a number of factors – teachers being only one of them – it is difficult to single out the effect of a teacher on an individual's student outcome (McCaffery *et al.*, 2003 cited in Rosenkvist, 2010). Recently, some researchers claim that the difficulty in linking teacher effort to student performance can be overcome through value-added assessment, which method attempts to isolate the teacher's contribution by controlling for other influential factors (Heyburn *et al.,* 2010 cited in Rosenkvist, 2010). Yet, value-added assessments are not a panacea for teacher evaluation based on test results; researchers such as Reardon and Ruadenbush (2008) and Ravitch (2010) claim the value-added method rests on various assumptions and has methodological limitations (cited in Rosenkvist, 2010).

111. A second problem with using test results to evaluate teachers is that doing so adds another purpose to standardised tests. Some authors, such as Madaus (1995), Popham (1999) and Shepard (1989), find that the milieu of purposes standardised tests are intended to serve dilute the value of the results, meaning the results are not appropriate to simultaneously foster good teaching, hold schools accountable, and monitor national progress (Wang *et al.*, 2006). Mixing accountability purposes, Shepard claims, is likely to be distorted by incentives and scores will not be a true representation of student outcomes (Wang *et al.*, 2006).

112. Finally, placing a "premium" on student test performance in the form of rewards or sanctions for teachers increases the risk of instruction being reduced to test preparation, which in turn limits the depth of the student experience and reduces the skill needed by teachers (McNeil, 2000; Smith, 1991 cited in Abrams *et al.*, 2003). Additionally, incentives such as bonuses can lead to strategic actions by teachers that distort or manipulate data. These include cases of teacher cheating, exclusion of students in assessments, and teaching to the test, all of which are reviewed in greater detail below.

113. There is consensus in the literature that student test results cannot be the sole measurement of teacher performance, leading to specific rewards and/or sanctions (Rosenkvist, 2010). A multi-pronged approach to teacher evaluation goes hand in hand with a multi-pronged approach to student standardised testing, which is discussed further in the concluding sections of this paper.

2.9.5 Test results are used by parents and stakeholders outside of school system

114. Across OECD countries, student test results are used by parents to make decisions about their children's academic career. If the test results are publicly available, they are often used by stakeholders

outside the school system, as well. The discussion below highlights the debates over using test results to compare and rank schools, to monitor student performance, make demands of the school or public education system and to monitor the education system's competitiveness.

2.9.5.1 Test results are used to compare and rank schools

115. In a few OECD countries, student standardised test results are used to rank schools. This often involves publishing school level results in some manner, be it to the public or within the education system. School rankings can also be done independently – by the media or other outlets – and may not be directly associated with the education evaluation system. For example, in Ontario, Canada school rankings based on assessment results are published by a non-profit policy organisation, rather than the Ministry of Education (Campbell and Levin, 2008).

116. By publishing school results and ranking schools, some claim it provides valid evidence to taxpayers and other stakeholders on the effectiveness of a school and it can serve as a basis to intervene in a school if necessary (Rosenkvist, 2010). Further, ranking schools and providing specific data on school test results offer a means for parents and others to meaningfully compare schools and, in systems which allow for school choice, parents can make decisions about which school their child attends. Publishing student test results with the aim of comparing or ranking schools is referred to as "performance tables". However, the issue of publishing performance tables is highly debated as there is disagreement over what kind of reporting is most effective and to what extent the information is used to improve student experiences (OECD, 2007 cited in Rosenkvist, 2010). Also contested is the need for rankings to take into account contextual factors that are beyond school control. In response to this problem, some rankings use value-added approaches, which take into account additional data other than student test results in preparing rankings (Campbell and Levin, 2008). A second method is a 'relative' ranking system, which ranks and compares schools that have similar characteristics (Campbell and Levin, 2008).

117. According to Eurydice (2009), in a majority of European countries, results of national tests for each school are not publicised; some countries regulate this by clearly stating the results are not to be used for ranking or publication, such is the case in Finland and Belgium. In Denmark, Poland, the UK and Australia results from national assessments are publicised. In some cases, results are published in 'league tables', as used in the UK, which are intended to rank schools by performance and increase competition and, in turn, improve student achievement (Reimers, 2003 cited in Kellaghan *et al.*, 2009). As Vegas and Petrow (2008) note, simply the publication of performance information on schools can pressure schools to improve performance (cited in Kellaghan *et al.*, 2009). Publishing detailed school results can be problematic in that they may not be accurate and school performance should take into account factors over which schools have no control (such as a student's home environment) (Kellaghan *et al.*, 2009; Eurydice, 2009). The issue of accurately stating school performance gains can be resolved by applying value-added models, which take into account such factors in calculating a school's test results ranking (Kellaghan *et al.*, 2009). Further, publishing results is often linked with an increase in corrupt practices or distortionary measures, such as excluding low-achieving students from tests or even teacher cheating. Lastly, a great deal of literature is concerned with the unintended consequences of publishing national assessment data, for more information see: Clotfelter and Ladd, 1996, cited in Kellaghan *et al.*, 2009; Kane *et al.*, 2002; Kellaghan and Greaney, 2001, cited in Kellaghan *et al.*, 2009; and Linn, 2000, cited in Shewbridge *et al.*, 2011.

2.9.5.2 Test results used for monitoring, informational or advocacy purposes

118. Aside from performance tables, which allow stakeholders to compare and rank schools, student test results are used by parents and other stakeholders outside of the education system for informational

purposes like monitoring student performance, informing demands of the school or supporting the public education agenda, and monitoring the education system's competitiveness.

119. In Australia, NAPLAN results of individual students are distributed to parents. The report details the student's results in comparison with other children at the same year level, including the student's performance against the national average, national minimum standard, and school average (Santiago *et al.,* 2011). The results at the school level are publicly available on the *My School* website. Australian parents are therefore well informed of their child's progress and can use the information to make demands of the teacher or the school. Alternatively, in Spain, parents only receive a short summary report of the school system's performance and more detailed and technical reports are only given to the relevant ministries (INCA, 2011).

120. In Chile, publication of national assessment results led to civil society groups putting education on the public agenda (Kellaghan *et al.*, 2009). Also in Chile, national assessment results are used by the independent institution, JUNAEB, which provides free meals and other assistance to poor students. JUNAEB uses Chile student's test results to estimate educational risk factors and focus resources on at-risk students (Meckes and Carrasco, 2006). In Denmark, an OECD review team finds that there are increasing demands for education system-level information from stakeholders outside the education sector (Shewbridge *et al.*, 2011). For instance, annual progress towards educational outcomes is compiled by the Danish Ministry for Economic and Business Affairs as a "Competitiveness Report". It is therefore important to consider the purpose of disseminating national test results and the potential array of uses by external stakeholders.

121. Arregui and McLauchlan (2005) find that in Latin America, informing the public about student achievement levels has come to be regarded as an important outcome of a national assessment (cited in Kellaghan *et al.*, 2009). In the United States, each state administers School Report Cards and in Canada student test results are made available online. Yet, parents can only make informed choices if information is available in a consistent and comparable manner for all schools (Bradley *et al.*, 2000 cited in Rosenkvist, 2010).

2.10 Decisions on reporting results

122. This section reviews different options with regard to reporting results. In some OECD countries, test results are published by the Ministry of Education, whereas in others results are not published. Differences also exist in whether the results are aggregated or disaggregated by groups and whether results are adjusted to account for student characteristics, such as socio-economic status, gender, or language. Often these decisions are a reflection of the over-arching test purpose.

2.10.1 Publishing test results (see also Section 2.9.5)

123. The decision to publish test results – in any aggregate form – should be carefully considered as there can be both positive and negative repercussions (Mons, 2009). Although it is helpful to make scores available so parents and administrators exert pressure on low-scoring schools or to inform school choice, publication can also stigmatise low-performing schools and have ramifications on school enrolment, teacher quality and funding (Hamilton and Koretz, 2002: 44). In deciding whether to publish results, the purpose of publication as it relates to the over-arching test purpose should be considered; publication may not be appropriate if the purpose of the test is to monitor national progress and serve as a pedagogical tool, for example (Shewbridge *et al.*, 2011). Publication continues to be a debated issue and there is no consensus over what types of result reporting is most effective in raising performance and engaging stakeholders in school improvement efforts (OECD 2007 cited in Rosenkvist, 2010).

2.10.2 Aggregated vs. disaggregated data

124. In using and reporting results, a choice must be made as to whether aggregated data or disaggregated data should be used. Student test scores can be aggregated to the classroom level, school, district, state, or national level. The decision about which unit of report is to be used will impact the assessment's design and should be linked with the test's purpose.

125. Policy makers may choose to aggregate scores because this allows for matrix sampling of items and, in turn, increases the validity of the results and reduces the likelihood of curriculum narrowing (Hamilton and Koretz, 2002). The drawback to this approach and to using matrix sampling is that it becomes more difficult to provide individual scores and scores across students are not comparable. Therefore, to gain information on individual student performance, matrix sampling should be avoided. A second reason for aggregating results is that in certain instances aggregation can increase the accuracy of results, as Hamilton and Koretz (2002) note: "school-level scores typically display greater degrees of accuracy than do individual level scores". On the other hand, disaggregated results allow policy makers to compare groups of interests, such as gender, ethnicity, minority groups, and students with disabilities, for example.

2.10.3 Adjusting results

126. In some OECD countries, test results are adjusted to account for factors that are outside school control that may influence a student's performance. These factors can include school characteristics (funding, location, student population) or student characteristics (gender, mobility during the school year, home environment, and levels of deprivation). In the discussion about evaluating teachers based on test scores, it was mentioned that some contest using test scores to evaluate teachers because a student's performance is an outcome of multiple factors. To alleviate this bias, some argue that test scores should be adjusted to account for socio-economic (or other) differences in students (Hamilton and Koretz, 2002). By adjusting scores, schools and test results become more comparable and schools can be compared with others that have similar student populations (Hamilton and Koretz, 2002). Yet, adjusting scores may not be sufficient or can be troublesome if student data is inaccurate or unavailable. A second drawback to score adjustment is the "institutionalising effect" it can have on applying different standards for different students (Hamilton and Koretz, 2002: 37). Due to the strong relationship between student achievement and socioeconomic status – schools that serve poorer populations tend to face lower achievement levels at the time of entry into school in comparison with schools that serve more advantaged families – by allowing for score adjustment, it solidifies or institutionalises the act of applying different standards for different students, which may in turn have repercussions on how these students are treated by teachers and others (Hamilton and Koretz, 2002).

3. COMPETENCIES FOR DEVELOPING AND USING STUDENT STANDARDISED TESTING

127. This section reviews how OECD countries develop competencies for developing and using student standardised tests. This includes allocating responsibility for test development and implementation and developing the capacity of teachers and administrators to use test results effectively. In most OECD countries, the Ministry of Education is responsible for implementing and analysing standardised tests. There is also consensus in the literature that teacher's and administrator's capacity for using standardised test results should be developed in alignment with the standardised test.

3.1 Agencies responsible for test design and implementation

128. Across OECD countries, various agencies and public and private actors are involved in evaluating the education system. With regard to large-scale, standardised tests, responsibilities will likely extend beyond the implementation agency and other agencies will be involved in tracking and analysing data. Typically, a country's Ministry of Education (or similar body) is responsible for developing and implementing standardised tests and recording test results. OECD countries differ, however, in the level of centralisation associated with standards and target setting, curriculum development and responding to standardised test results. In Australia, national curriculum, standards and the national assessment are under the responsibility of ACARA (Australian Curriculum, Assessment and Reporting Authority), yet in the United States states are given more responsibility for defining minimum education standards to apply to the national assessment, NAEP. The delegation of responsibility within the Ministry of Education or other ministries often reflects the objectives of the assessment system and the context of the agency. In France, the education ministry has a central evaluation unit with special offices for student assessment and policy studies (Shewbridge, forthcoming) The purpose of an assessment test will influence the type of agency involved: if the national test aims at monitoring the education system, there may only be one body responsible for tracking data; or, if national tests are used for school evaluation and monitoring, multiple agencies may review and act upon the data. For a more complete discussion on the agencies involved in school evaluation, see Faubert (2009). Among OECD countries there are also differences in the prevalence of evaluation and assessment agencies, research bodies and private test developers. The United States has many independent centres for education evaluation, research and statistics, whereas Japan has few; this reflects the strong role education evaluation plays in the US (Shewbridge, forthcoming).

129. Standardised test design and implementation, including regulating testing standards and managing data collection, is typically the responsibility of an arm of the Ministry/Department of Education (Canada, U.S.), an independent authority (Australia) or a semi-autonomous body (Mexico). It is less common that private institutions develop national tests, yet this is the case in the Netherlands.

130. In Australia, ACARA is an independent authority responsible for developing the national curriculum, the national assessment, and managing data collection and reporting for education. The national test, NAPLAN, is developed by ACARA in consultation with experts in the areas of literacy, numeracy, ICT, sciences, civics and citizenship. Assessment experts, teachers, and education authorities from across Australian schools are involved in test development. Therefore, ACARA aims to collaborate with a wide range of stakeholders, including teachers, principals, and professional education associations, throughout the test development and implementation process (ACARA, 2011).

131. In the Netherlands, although the Dutch Examination Board (CEVO) is responsible for setting education standards, the development of standardised tests are contracted out to a private agency, the National Institute for Educational Measurement (CITO). CITO was originally founded by the Dutch government and privatised in 1999. CITO develops the test by consulting with a CITO subject matter specialist and a group of subject matter teachers. The test is then validated by groups of subject specialists and education representatives from CEVO (Beguin *et al.*, 2008).

132. In the United States, the National Assessment of Educational Progress (NAEP) is based on specifications provided by the National Assessment Governing Board and tests are developed in conjunction with the National Center for Education Statistics. Test development is a rigorous process that aims to design high quality tests by consulting with panels of business representatives, members of the public, local and state policy-makers, curriculum specialists, practitioners and researchers. An extensive review process is followed throughout the test design phase and following the test's implementation a post-review is conducted of test items (NCES, 2011b).

3.2 Development of capacity and assessment literacy

133. Throughout OECD countries there is a trend in involving different stakeholders in the test design process; the insight and expertise of teachers is not only a critical information source, but teacher involvement also increases the perceived usefulness of test results. Following test design and implementation, government agencies across OECD countries differ in their approach to developing teacher and administrator capacity to interpret and use results appropriately. This concept, often referred to as "assessment literacy", encompasses the following actions:

- Capacity to examine student data and make sense of it;
- Ability to make changes in teaching and school derived from those data; and
- Commitment to engaging in external assessment discussions (Rolheiser and Ross, 2001 cited in Campbell and Levin, 2008: 48)

134. The literature stresses that for standardised test results to be used effectively, educators must have the capacity to assess, understand and apply such data (Santiago *et al.*, 2011; Diamond and Spillane, 2004; Earl, 2003; and Mason, 2001 cited in Campbell and Levin, 2008; Ingram *et al.*, 2004, cited in Volante and Ben Jaafar, 2008). Without developing assessment capacity, the result can be "a sorry mixture of confusion, technical naivety and misleading advice" (Goldstein, 1999 cited in Campbell and Levin, 2008: 49).

135. In Ontario, Canada, developing capacity and assessment literacy is the responsibility of the school district. Campbell and Fullan (2006) found that school districts in Ontario that showed improved student outcomes also identified the development of assessment literacy at both the school and district levels as important activities (cited in Campbell and Levin, 2008). Such development activities included: providing professional development on data analysis and assessment literacy for principals and teachers; clearly setting expectations about the use of students assessment information; supporting schools in using and understanding data; encouraging the use of data to inform improvement planning, set goals and provide feedback (Campbell and Levin, 2008). In Denmark, an OECD review has found that recent policies to build teacher capacity in evaluation and assessment have supported teachers and encouraged the incorporation of evaluation and assessment into instruction. Yet, the Danish system can still improve in engaging teachers to effectively use national test results to identify student strengths and weaknesses (Shewbridge *et al.*, 2011).

4. EMPIRICAL EVIDENCE ON THE IMPACT OF STANDARDISED TESTS FOR IMPROVING STUDENT OUTCOMES

136. This section presents the evidence on the impact of standardised tests on improving student outcomes and teaching. Although some authors have warned that "research on the consequences of standardised assessment has been described as yielding 'scarce and equivocal' evidence", the following discussion aims to present the most relevant studies to inform educational policy making (Mehrens, 2002 cited in Wang *et al.*, 2006: 306). The section first synthesises the research studies examining the impact of standardised tests on student outcomes, primarily pulling from studies on the U.S. system. Second, evidence on the impact of standardised tests on teaching is presented, including the unintended consequences and system distortions triggered by testing systems. Finally, the section reviews the evidence for standardised tests to reduce achievement gaps, as this is a common goal of assessment systems in OECD countries.

4.1 The impact of standardised tests on student outcomes

4.1.1 Empirical evidence on the impact on student outcomes is mixed

137. There is not a consensus in the literature as to whether standardised tests improve student outcomes and learning. Moreover, most impact studies are focused on cases in the U.S. and the U.K., as both of these countries have highly developed standardised assessment systems and the empirical evidence on the impact in other OECD countries is somewhat more limited. Synthesising the research on the effect of standardised tests on student outcomes is challenging for two reasons. First, many impact studies look at the effects of tests that have high-stakes for students and there is less research on the impact of no-stakes, standardised tests. Similarly, the research tends to focus on standardised tests for accountability purposes and there is less evidence on the impact of standardised tests for feedback or monitoring purposes. A second challenge relates to the ability of the researcher to identify policy causation; that is, to determine whether impact is a result of standardised testing or other policy choices such as monetary rewards or school reforms. An example of this issue can be seen in the research by Goldhaber and Hannaway (2001) who found that in Florida, some low-performing schools that received resources as a result of poor performance used the money to reduce class sizes, provide new instructional materials and staff development, and offer after-school tutoring programs. Goldhaber and Hannaway were unable to determine whether any resulting student improvements were an effect of standardised testing or the effect of additional spending (cited in Hamilton and Koretz, 2002). These two challenges should be kept in mind when examining the effect and impact of standardised tests.

138. Figlio and Loeb (2011) offer a comprehensive review of test-based accountability systems and their impact. According to their synthesis, "Though no one approach or study is flawless and many inconsistencies remain, taken as a whole, the body of research on implemented programs suggests that school accountability[2] improves average student performance in affected schools" (Figlio and Loeb, 2011: 410). Their review of the research reveals that the No Child Left Behind Act and the accompanying

[2] In their analysis, the authors consider school accountability systems that are "based in large measure on student testing" (Figlio and Loeb, 2011).

NAEP test have led to improved student performance in mathematics, while it is less clear what the effect is on reading.

139. This finding is exemplified by Wong, Cook and Steiner's 2009 research findings (cited in Figlio and Loeb, 2011). These authors used multiple approaches to examine the effects of NAEP and accountability provisions on student achievement in fourth and eighth grades. They evaluated No Child Left Behind using National Assessment of Educational Progress (NAEP) data between 1990 and 2009 for 4th grade reading and 4th and 8th grade mathematics. The authors found that "Across all these analyses, NCLB consistently improved both 4th and 8th grade mathematics, though 4th grade reading effects were limited to states with both high standards and an accountability system that included sanctions only after NCLB" (Wong *et al.*, 2009 cited in Figlio and Loeb, 2011).

140. Cronin *et al.* (2005) compared achievement and student growth in the Northwest United States explicitly looking at growth prior to and immediately following NCLB implementation (cited in Figlio and Loeb, 2011). These authors found that achievement levels in mathematics increased after NCLB implementation; but the same results were not conclusive for reading.

141. Finally, Dee and Jacob (2009) compared NAEP state data from 1990 to 2007 based on achievement growth and whether the school previously had an accountability system before NCLB implementation (cited in Figlio and Loeb, 2011). Dee and Jacob found that there were greater gains in mathematics in the states that did not have a test-based accountability system prior to NCLB, suggesting the positive impact of NCLB.

142. Figlio and Loeb (2011) succinctly summarise the U.S. case stating: "While, in general, the findings of the available studies indicate achievement growth in schools subject to accountability pressure, the estimated positive achievement effects of accountability systems emerge far more clearly and frequently for mathematics than for reading" (410). In her review of the effects of standardised testing, Mons (2009) finds similar differences in the impact of accountability pressure on mathematics and reading scores, stating: "The research revealed inconsistencies in the results for mathematics and reading: in some cases, reforms were linked to performance improvements, and others saw a decline (Amrein and Berliner, 2002 cited in Mons, 2009: 17). Evidence of the relationship between improved student outcomes and standardised testing is "unpredictable" and is a product of a number of complex policy decisions and implementation structures (Mons, 2009: 19).

143. While the above studies point to the positive impact of NCLB on student outcomes in mathematics, other scholars find that NCLB has not improved student outcomes. In her 2010 book *The Death and Life of the Great American School System*, the prominent U.S. education policy analyst Diane Ravitch explains in detail the negative impact NCLB has had on the American education system. Formerly an advocate of NCLB, after years of implementation, Ravitch has reversed her view and claims that the level of education received by students post-NCLB has remained "disastrously low". Ravitch draws on a 2009 Chicago study to emphasise her point. In this study, the improved performance of students in Year 8 in math and reading were a result of changes made to the tests and testing procedures, rather real improvement in student learning (Commercial Club of Chicago, 2009). Further, these gains 'evaporated' by the time students reached secondary school.

144. A 2011 report by the U.S. National Research Council summarises the literature surrounding the effect of NCLB on student outcomes and the report comes to a more sobering conclusion than the Figlio and Loeb review. While Figlio and Loeb highlight studies in which NCLB had a positive impact on mathematics outcomes (see above), the NRC report finds that studies of the impact of NCLB show positive, negative and non-significant results. The review finds that initially, the findings of the impact of NCLB on student achievement appear "substantial", however, upon further analysis the authors find that

statistically significant effects are concentrated in one area: Year 4, mathematics. Results for Year 8 mathematics and for reading in both Year 4 and Year 8 are not significant or in some case are negative. Therefore, the impact of NCLB should not be skewed to be solely positive. The NRC report also explains: "…the evidence related to the effects on achievement of test-based incentives to schools appears to be modest, limited in both size and applicability" (Hout and Elliott, 2011).

145. The NRC report claims that the gain in student outcomes as a result of NCLB is small and, given the ambitious program initiated by the U.S. government to improve significantly student outcomes by 2014, the system is set to fail. NRC finds that current programs that do have positive effect raise achievement of students who are currently in the 50th percentile to the 53rd percentile; whereas to reach the goals set forth by NCLB policy a student in the 50th percentile would need to increase their scores to the current 84th percentile (Hout and Elliott, 2011). NRC summarises this point, stating:

146. "Test-based incentive programs, as designed and implemented in the programs that have been carefully studied, have not increased student achievement enough to bring the United States close to the levels of the highest achieving countries. When evaluated using relevant low-stakes tests, which are less likely to be inflated by the incentives themselves, the overall effects on achievement tend to be small and are effectively zero for a number of programs."

147. The impact of low-stakes, standardised testing is complex as test-based accountability systems, their incentives and the environments within which they act differ across states and nations. In the U.S. case, which tends to dominate the literature due to the controversial effects of the NCLB policy, the impact on student outcomes appears to be positive in terms of elementary mathematics, yet not statistically significant for reading and not great enough to produce the improvement in student learning that the policy aims to achieve. While the academic evidence and ensuing discussion does not necessarily point towards the removal of standardised testing, it does highlight the potential for improvement in the development of assessment systems. Section 5 draws out these lessons from the literature in more detail.

4.2 The impact of student standardised tests on teaching: Unintended consequences of standardised tests

148. Although the evidence is unclear as to whether standardised tests lead to improved student outcomes, there is more certainty that standardised tests lead to increased strategic behaviours on the part of schools and teachers. When student test results are used in accountability systems to reward and/or sanction schools and/or teachers or when student test results are published at the school level to allow for performance tables and other ranking actions, teachers and schools will perceive the no-stakes, standardised tests to be high stakes. Consequently, when such tests are perceived to be high stakes for teachers and schools, the assessment system risks being distorted by the following strategic actions: teaching to the test; narrowing of curriculum; teacher cheating; student exclusion. Evidence shows that such actions can lead to inflated student scores and reduce the validity of student test results. The following section reviews the theoretical and real impact of standardised testing on teaching.

4.2.1 Teaching to the test

149. One response teachers may have when incentives are attached to student test results is to increase instruction on test preparation, also referred to as "teaching to the test". This behaviour is manifested in a teacher increasingly teaching test-taking skills (such as tips for multiple choice tests or focusing on essay writing) or by using test items or similar items in their instruction (Popham, 1999). Popham differentiates between teaching to the test – which he refers to as "item-teaching" – and instruction which aims to focus on the subject matter that will be covered in the test, "curriculum-teaching" (Popham, 1999). When a teacher organises instruction around actual test items, rather than a body of content it is teaching to the test

(Popham, 1999). Teachers may choose to align their teaching to the knowledge and skills assessed in the standardised test, thus neglecting other curriculum areas that are not going to be assessed.

150. The problems with teaching to the test are two-fold. First, by emphasising test-taking skills and concentrating on tested content, scores will become inflated without reflecting an increase in student understanding of concepts (Hamilton and Stecher, 2002; Hout and Elliott, 2011). Shewbridge *et al.* (2011) writes, "Research from the United States has shown that if national tests are considered to be 'high stakes' for teachers and schools, teaching to the test can easily lead to an artificial over-inflation of results and thus render the results useless as a measure of real progress" (*e.g.* Koretz, 2005 cited in Shewbridge *et al.* 2011: 120).

151. When instruction is narrowly focused on specific knowledge, skills and question formats test results become an increasingly misleading measure of student achievement (Hout and Elliott, 2011). This problem is often exacerbated when teachers perceive tests as high stakes as a result of incentive systems; it can also be caused inadvertently if tests are not updated frequently. For example, in Sweden the Year 5 standardised test is used over 2 successive years and teachers may inadvertently "teach to the test" in the second year leading to score inflation and reducing the validity of the results (Nusche, forthcoming). In Denmark, teaching to the test takes the form of increased focus on tested content areas and reduced focus on creative, innovative and oral skills (Wandall, 2010 cited in Shewbridge *et al.*, 2011). Secondly, teaching to the test emphasises rote memorisation and a more passive approach to learning as teachers spend more time developing test-taking strategies rather than cultivating students' problem-solving skills (Kellaghan *et al.*, 2009).

152. By involving teachers in standardised test development and implementation and training teachers on how to effectively use and analyse test results, there can be less risk of teaching to the test. However, incentives attached to test results must be carefully constructed to avoid motivating strategic behaviour such as teaching to the test; the research is mixed as to how to effectively promote positive behaviours as a result of incentives while reducing negative behaviours.

4.2.2 Narrowing curriculum

153. As discussed above, a limitation of standardised tests is the inability to test attainment against the full curriculum. In evaluation systems where incentives are attached to test results and where teachers and schools are pressured to improve test scores, teachers have the tendency to adapt or restrict their content focus accordingly to the aspects of the curriculum which will be tested. Curriculum narrowing differs from teaching to the test in that curriculum narrowing refers to the increasingly unbalanced focus on content areas that will be tested and neglecting non-tested areas whereas teaching to the test refers to using test items (or similar items) to teach. No-stakes, standardised tests often assess language and mathematics and rarely assess other subject areas such as sciences, civics, history and foreign languages. Curriculum narrowing leads to more time spent on tested areas, like mathematics and reading, and less time on non-tested content, such as history. As a consequence of accountability systems, teachers may overemphasise certain subjects that will be tested, even if they make up only a small part of the entire curriculum (Eurydice, 2009). In addition, specific subjects can be omitted from the lesson plans as tested areas are given increasing priority (King and Zucker, 2005).

154. These effects have been documented in education research in the United States and the U.K. In the U.S., a survey by the National Board on Educational Testing and Public Policy (2003) found that 79% of teachers in states with accountability testing reported that instruction in the tested subject areas had either increased a great deal or moderately, and that more time was being devoted to the tested segments of the curriculum than to the non-tested segments (King and Zucker, 2005). According to Eurydice (2009): "An inquiry into the system of national tests undertaken in 2007 (in the UK) by the Children, Schools and

Families Select Committee revealed that many teachers felt obliged to attach undue importance to those aspects of the curriculum that were liable to feature in the tests, and to focus too much attention on pupils who seemed capable of achieving the performance targets set by the government" (Eurydice, 2009: 61).

155. Eurydice recommends that curriculum narrowing can be minimised by rotating subject areas in yearly cycles and by increasing the number of subjects tested annually (2009). Other authors insist upon considering such unintended consequences in the design phases of standardised tests to ensure the purpose, test format and incentives contribute to a functioning evaluation system.

4.2.3 Exclusion of students

156. When standardised tests are linked to accountability measures schools and teachers alike are pressured by incentives and sanctions to improve student test scores. Researchers have found that one response to this pressure is to manipulate the student population and exclude low-performing students from taking standardised tests. Figlio and Loeb write in their synthesis of school accountability research, "The evidence is quite clear that schools have responded to accountability pressures by reclassifying low-performing students as students with disabilities (see: Cullen and Reback, 2006; Deere and Strayer, 2001; Figlio and Getzler, 2007; Jacob, 2005 cited in Figlio and Loeb, 2011: 394). Eurydice (2009) found that in the Netherlands, weaker students were not given standardised tests in anticipation that they would be transferred to remedial classes in the following year; in this way, "schools sought to keep their average marks high and hence protect their image" (Eurydice, 2009). Booher-Jennings (2005) also found that teacher's behaviour changed following the introduction of standardised tests in the United States. Teachers began to classify pupils into three groups – safe cases, suitable for treatment, and hopeless cases – and subsequently focused primarily on the middle group, seeing that this group had the most potential for improvement. In this same case, it was found that weaker pupils received less attention.

157. According to some researchers, incentives to exclude students from testing groups are reduced using growth models rather than status models in testing frameworks. Status and growth models refer to how outcomes are measured and they each create different objectives and incentives for schools. Status models measure the percent of students who achieve certain levels of proficiency and require schools to raise performance to meet the proficient level (Krieg, 2008; Neal and Schanzenbach, 2010 cited in Figlio and Loeb, 2011). Growth models measure improvement by looking at how a school has improved student performance independently of the level of achievement (Figlio and Loeb, 2011). Within each model, teachers are motivated to focus attention on different student groups. In status models, low-performing students receive greater attention to bring them up to proficiency levels; in growth models, teachers are rewarded for improving student outcomes for any level of student. Hence, in growth model systems teachers are less likely to exclude under-performing students, as any improvement in outcome is measured.

4.2.4 Teacher cheating

158. Standardised tests, especially in accountability systems, can lead to cases of teacher cheating. Teachers have been found to engage in a number of activities which manipulate student test data, including: changing student responses, filling in answers that were left blank, allowing additional time for testing, providing correct answers to students or obtaining copies of the exam prior to the test date (Jacob and Levitt, 2002). Research in the United States has shown that "serious cases of teacher or administrator cheating" on standardised tests occurs in at least 4-5% of primary school classrooms each year (Jacob and Levitt, 2002). Jacob and Levitt identify the strong connection between incentive systems and teacher cheating, stating that "cheating appears to respond strongly to relatively minor changes in incentives" (Jacob and Levitt, 2002). In the time of writing this paper, recent U.S. news reports have brought to light teacher cheating scandals in some U.S. States, further highlighting the prevalence of this problem and reigniting the U.S. public debate on the validity and relevance of standardised tests. While there does not

seem to be a consensus as to how to definitively reduce the likelihood of teacher cheating, Figlio and Loeb purport that the growth model approach to accountability reduces such manipulative behaviour since "increases in student achievement in one year would make it more difficult for the school to attain accountability goals the following year" (2011: 400).

4.2.5 Impacts outside of the education sector

159. Certain spill-over effects have also been documented to be linked with standardised testing practices. Bokhari and Schneider (2009) find that in the U.S. the increased use of psychostimulants is linked with school accountability policies, suggesting the health consequences of education policies (cited in Figlio and Loeb, 2011). Figlio (2006) found that some U.S. schools aligned disciplinary actions around testing schedules, in order to improve the average test taker's score (Figlio and Loeb, 2011: 399).

160. Some studies have shown that standardised tests and test-based accountability systems have an impact on the teacher labour market. Figlio and Loeb (2011) summarise this issue, writing: "The research to date suggests that accountability has not dramatically changed the career choices of teachers overall, but that it has likely increased attrition in schools classified as failing relative to other schools" (416). There is evidence that teacher attrition has increased as a result of perceived high-stakes tests being implemented, which can be problematic if this involves effective teachers leaving poor performing schools, which is more often the case (Figlio and Loeb, 2011). Therefore, the effect of standardised tests on the labour market can be detrimental if effective teachers shy away from entering 'failing' or 'low-performing' schools.

161. A further complication introduced by standardised tests is the loss of motivation among teaching staff, particularly in schools with disadvantaged pupils and poor results as documented by (Jones, 2007; Behrens, 2006 in Mons, 2009: 27). Mons (2009) finds that across a number of OECD countries, such as France, the U.S., U.K. and Sweden, teachers are open to the principle of standardised assessment and do not discredit the tool of standardised tests to improve student performance. Yet in these same countries, teachers may criticise specific standardised testing programs because of their negative impact on teaching and learning, because they fail to account for social and economic characteristics of students and because of the link between student performance and teacher rewards (Mons, 2009). Mons cites cases in the U.S., U.K. and France where teachers have expressed reservations towards standardised tests. In the U.K. teachers unions have called for a boycott of national tests and in France teachers unions have spoken out against standardised testing. Standardised testing can not only impact student outcomes, but it also impacts teacher motivation and career choice, which in turn may hinder the implementation and progression of the assessment framework.

162. Overall, the repercussions of standardised tests can extend beyond immediate student outcomes and affect the school system and the teacher labour market in complex and pervasive ways.

4.3 Standardised tests may not reduce achievement gaps

163. In some OECD countries, standardised tests aim to improve student outcomes and reduce racial, social and economic achievement gaps. Authors such as Grissmer *et al.* (2000) and Hong and Youngs (2008) argue for the use of standardised assessment to reduce educational inequalities, claiming that by implementing common standards, assessments level the playing field and motivate teachers to improve student outcomes regardless of the student's social, ethnic or economic characteristics (cited in Mons, 2009). Effectively, by standardising education and educational targets, supporters of standardised testing as a tool to reduce achievement gaps claim this limits inequalities in teaching thereby improving student performance across ethnicities.

164. Figlio and Loeb (2011) find that the empirical evidence to support this claim is mixed, however. In an analysis of NAEP results across U.S. states, Hanushek and Raymond find that state test-based accountability systems may have reduced the achievement gap for some groups of students, but increased it for others (Hanushek and Raymond, 2005). When standardised test results are disaggregated by race, there are differences in gains by racial groups. By examining NAEP test results between 4th and 8th grade, Hanushek and Raymond find that African Americans and Hispanics in the U.S. have lower rates of performance improvement in standardised tests than Caucasians (2005). The authors conclude by stating, "Thus, even though accountability provides a positive gain on average, that dividend is not sufficient to override the prevailing differential in performance when students are broken out by race/ethnicity" (Hanushek and Raymond, 2005). Such findings point to the dual-aim of standardised testing policies and the drawback this may have on the policy's impact. In the U.S., NAEP tests not only aim to improve student outcomes across the U.S., but they also serve as a measure of achievement gaps in the country. Studies like that of Hanushek and Raymond are evidence that these two goals may not be achieved solely through implementing student standardised testing or test-based accountability systems.

165. Mons' (2009) synthesis of the evidence on achievement gaps echoes the findings of Figlio and Loeb (2011): there is no consensus in the literature as to whether standardised assessment reduces educational inequalities among different social and ethnic groups. For example, Carnoy and Loeb (2002) find that standardised testing may benefit ethnic minorities in the U.S., but others such as Lee and Wong (2004) and Nichols *et al.* (2006) did not find that standardised testing had significant benefits for ethnic minorities (cited in Lee, 2008).

166. It would be interesting to expand on this empirical literature. While many studies cite cases in the U.S., a cross-country analysis would perhaps be beneficial as it would highlight the effects of different standardised assessment systems on reducing achievement gaps.

5. ASPECTS OF STANDARDISED TESTING THAT ARE MORE CONDUCIVE TO IMPROVING SCHOOL OUTCOMES

167. In synthesising the relevant literature, this paper has explored the various debates and decisions associated with large-scale, standardised tests. The trends in OECD countries with regard to standardised testing for monitoring, accountability or diagnostic purposes differ widely in the content and form of the test, how their results are used, and the intended and unintended impact on student outcomes. It is evident that there is no blueprint solution for an appropriate use of standardised tests, as results are influenced by each country's context. Yet, a number of lessons can be drawn from the literature to guide countries in building a sustainable and effective assessment system that genuinely improves student outcomes. These lessons are drawn out in the following section.

5.1 Lesson 1: Clearly establish the purpose of the test and allow this to lead all following test design, implementation and use decisions

168. As discussed at the beginning of this paper, standardised tests can serve different purposes. Across OECD countries, standardised tests are tools for monitoring education systems, diagnosing student needs, informing teacher instruction and they often play a part in test-based accountability systems. Clearly establishing a purpose for the test and making this purpose known to the stakeholders is an important initial step in the test design process. Hamilton and Stecher (2002) elaborate this point: "It is important that states clarify the purpose of their testing programs as a basis for making decisions among competing demands and that they monitor the degree to which the tests are serving that purpose" (135). The decisions referred to by Hamilton and Stecher include the following: Firstly, the test's design, including question format, length and frequency, must be informed by the test's purpose. Secondly, the purpose of the test impacts the way test results are used. If a test is created for system monitoring purposes, it is inappropriate to use the results to diagnose student strengths and weaknesses. The purpose of the test also plays a role in determining the suitability criteria and the adequate balance of validity and reliability. Therefore, since the purpose of the test has an impact on subsequent test design and implementation decisions, it should be clearly established at the beginning.

5.2 Lesson 2: Testing standards should be aligned with the national curriculum to testing standards

169. A note on the importance of aligning testing standards to the national curriculum has already been included in Section 2.4.4. There is strong consensus in the literature that assessment measures should be linked with established content standards and curriculum to maximise the test's reliability, validity and utility (Mons, 2009; Meckes and Carrasco, 2006). Essentially, the assessment system should be developed around established learning objectives and educational standards, which feed into the national curriculum and finally into standardised test development. The underlying driver of the educational system is therefore student learning objectives, from which the national curriculum is designed. Aligning curriculum, instruction and assessment around the nation's educational goals and standards supports the validity of standardised tests (Alliance for Excellent Education, 2010). If the testing standards are not aligned with the curriculum, assessment results will provide misleading information about the extent to which students have met such standards (Alliance for Excellent Education, 2010).

5.3 Lesson 3: Be cautious in employing large-scale, standardised tests that serve multiple purposes

170. In order to maximise the utility of the test, researchers argue that tests should not serve a multitude of purposes (Hamilton and Stecher, 2002; Mons, 2009; Eurydice, 2009; Newton, 2007; Alliance for Excellent Education, 2010). When tests serve multiple purposes, the system runs the risk of reducing the utility of the test for any one of those purposes (Hamilton and Stecher, 2002). Mons (2009) echoes this idea, writing: "Establishing a single test to achieve several objectives (*i.e.* assessing system as a whole, supervising schools and monitoring academic progress) results in serious dysfunctions" (33). Eurydice (2009) finds that using one test for several purposes is inappropriate in that the information required for each purpose is likely not the same. One common example is using data from a test designed to measure student attainment to comply with accountability requirements; in this case, both formative and summative objectives are attributed to the same test, which can distort the incentives for teachers and students (Eurydice, 2009). In addition to requiring different information, different purposes may require alternative instruments. When national test data are used in performance tables and for monitoring performance over time, for example, the instruments are not optimised for use in multiple functions (Green and Oates, 2009). A clear, single purpose associated with large-scale, standardised tests aims to maximise test validity, in turn gauging a more accurate measure of student outcomes.

5.4 Lesson 4: Develop assessment literacy of teachers and administrators

171. Developing the capacity of teachers and administrators to use standardised test results appropriately and effectively is a critical pillar of support in an assessment system (Mons, 2009; Green and Oates, 2009; Alliance for Excellent Education, 2010). Training benefits the assessment framework not only by providing teachers and administrators with the specialised skills needed to utilise test results, but also by engaging teachers and administrators in the system thereby increasing stakeholder buy-in. Any large-scale, standardised test should be linked with ongoing training for teachers; in accountability systems, this is sometimes referred to as balancing 'pressure' and 'support', as capacity building is paired with developing accountability requirements (Barber and Fullan, 2005 cited in Campbell and Levin, 2008).

5.5 Lesson 5: Reduce distortion and strategic behaviour by increasing teacher involvement and buy-in from an early stage

172. In order to reduce system distortions, the testing culture and a commitment to the assessment system must be developed in teachers and school administrators. Throughout OECD countries teachers are involved in these stages of the standardised assessment, yet it varies by degrees. For example, in Sweden teachers implement and score the results, whereas in the United States teachers do not have a direct role in the development and scoring of tests. By encouraging teacher involvement and by providing results literacy training and professional development opportunities, teachers and school administrators will likewise be more motivated to appropriately use the test results for their and their students' benefit.

173. By engaging teachers in the design, management and analysis of test results, teachers are more committed in the testing process and are more likely to apply the test results to improve student outcomes (Mons, 2009). In order to promote desired responses on the part of teachers, it is critical that they understand and support the assessment goals (Hamilton and Stecher, 2002; Alliance for Excellent Education, 2010). Establishing clear goals and standards and communicating them to teachers mitigates strategic behaviour such as 'teaching to the test' as teachers have a clearer sense of what they should be aiming for with regards to student outcomes (Hamilton and Stecher, 2002).

174. Demailly (2001) identifies three conditions that should be met in the test design and implementation phases in order for teachers to support standardised tests (cited in Mons, 2009):

1) the evaluation must be developed through participation, with extensive teacher involvement;

2) the assessment objectives must be democratic; and

3) the project's sponsors must be able to persuade teachers and display a firm resolve (*i.e.* clarity in structure and outcomes).

175. In spite of stressing the importance of teacher buy-in, this is not to say that teachers across OECD countries oppose large-scale, standardised tests. In a number of cases, teachers favour standardised tests. In Sweden, for example, a 2004 survey by the National Agency for Education revealed that a majority of teachers favoured the national test because it provided clear guidelines for content instruction, it highlighted student strengths and weaknesses and it did not limit the scope of their teaching (Mons, 2009). Yet, although teachers are theoretically open to the principle of standardised assessment, they often object to certain programs and methods if "the skills tested are too archaic, if the tests fails to take into account the social characteristics of the student population or if there is an inappropriate link between performance and rewards" (Mons, 2009: 25). This opposition can be avoided if teachers are involved in the design process so that appropriate skills are tested and so that the intricacies of the classroom can inform decision making.

5.6 Lesson 6: Incorporate multiple measures of achievement especially in systems where standardised tests may be perceived as 'high-stakes' for teachers and school administrators

176. It is widely agreed that standardised tests are limited in the type of information they gather and that they do not provide a 'full-picture' view of student performance, student abilities or classroom instruction (Hamilton and Stecher, 2002; Harlen, 2007; Earl and Katz, 2006, cited in Volante and Ben Jaafar, 2008; Guilfoyle, 2006). "Only multiple measures of achievement can provide an accurate picture of student learning and school success," writes Guilfoyle (2006: 1). Employing multiple evaluation measures – including incorporating non-test information into decision-making – reduces the risk of making incorrect decisions as a result of the limitations of standardised test scores, improves the validity of the system, and reduces the likelihood of excessive narrowing of curriculum (Hamilton and Stecher, 2002). In addition to pursuing multiple evaluation techniques, educators and others should be made aware that tests provide "only a snapshot of students' achievement levels in select learning targets and subjects" (Wandall, 2010 cited in Shewbridge, forthcoming).

177. When standardised tests are used for accountability purposes, it is especially important to obtain a complete view of student outcomes and teacher instruction, which standardised tests cannot provide. Earl and Katz (2006) recommend gathering data in a wide range of forms, including standardised tests and formative classroom assessments, in order to enhance accountability evaluations (cited in Campbell and Levin, 2008). By implementing a 'toolkit' for understanding student performance and feedback, the concept of accountability becomes a conversation on ideas and challenges and a means to monitor progress, rather than a static approach to data collection and analysis. Such an approach to accountability not only provides more genuine data, but also can increase teacher buy-in and therefore reduce system distortions. For further information on improving test-based accountability systems, see Hamilton and Stecher (2002), Harlen (2007), Linn (2005).

6. CONCLUDING REMARKS

178. This paper aimed to review the relevant literature on large-scale standardised tests with no-stakes for students and to provide an overview of the trends in OECD countries in standardised testing. The debates associated with standardised testing were also introduced, to provide a comprehensive outline of the decisions and difficulties associated with standardised testing. Further, a brief review of the empirical research on the impact of standardised testing was included; yet, it has also been noted that further research on the impact on student outcomes should be done, as the current research lacks geographic diversity and is often found to be inconsistent. Lastly, lessons were drawn out from the literature to guide the development and implementation of standardised testing systems that are sustainable, valid, reliable and useful. The paper concludes by commenting on the delicate balance required of standardised testing systems to avoid over-reliance on national test results.

179. OECD countries readily recognise the potential benefits of implementing large-scale, standardised tests which have no-stakes for students. Often they are used to monitor the education system or they serve as tools to hold schools and teachers accountable. Despite the potential benefits, national testing systems are difficult to implement in a way that will truly improve student outcomes. Implementing and using the results of a standardised test are the outcome of a number of decisions regarding the test's format, content, suitability criteria and purpose. Policy-makers, educators and other stakeholders should be aware of distortionary practices such as teaching to the test, narrowing curriculum, student exclusion and teacher cheating and aim to create systems which are the result of collaboration between educators and policy-makers. The lessons above aim to outline the primary challenges in developing a standardised test that will reduce strategic behaviour and gather valid information on student performance.

180. While the appropriate use of standardised test results can feed into a school system that improves student outcomes, more empirical research should be done to identify the link between standardised test and student outcomes, as the current research is lacking in geographic diversity and often inconsistent (Mons, 2009; Wang *et al.*, 2006).

181. As a final note, in praising the ideal assessment system that supports improving student performance, it should also be stated that there is fair potential for educators and others to be over-reliant on national tests as an indicator of progress. In a system which places too much emphasis on the national test, students can lose out on learning skills that are not tested and teachers are motivated to 'teach to the test' (Santiago *et al.*, 2011). For example, heavy focus on a national test can drown out attention to other classroom-based assessment techniques and inherently increase the stakes of the national test (Santiago *et al.,* 2011). Over-reliance on national test results can be mitigated by increasing the range of indicators reported to the public (Santiago *et al.,* 2011).

APPENDIX A: OVERVIEW OF STANDARDISED TESTS IN SOME OECD COUNTRIES

Country	Test Name	Grades Tested	Subjects Tested	Sample/ Census	Intended Purposes
Australia[3]	NAPLAN	Years 3, 5, 7, 9	Reading, grammar, numeracy	Census	System monitoring School accountability Public Information
Canada[4]	PCAP	Students aged 13 (Years 6, 7, or 8)	Maths, reading and sciences	Sample	System monitoring Formative
Chile[5]	SIMCE	Years 4 and 8	Varies (language, mathematics, natural and social sciences)	Census	System monitoring School accountability Public Information
United States[6]	NAEP	Years 4 and 8	Reading, maths	Census	System monitoring School accountability
Denmark[7]	National Test	Years 2 through 8	Danish, Maths, English, Sciences (varies)	Census	Formative
Sweden[8]	National Test	Years 3, 5	*Year 3*: Swedish, Swedish as a second language, maths *Year 5*: Same as Year 3 + English	Census	System monitoring Formative
Korea[9]	National Assessment of Scholastic Achievement	Years 6, 9, 10	Various	Sample	System monitoring Formative
U.K[10]	National Curriculum Assessment	Years 3, 4, 5, 6	English, Maths, Sciences	Census	System monitoring Formative Public Information
Austria[11]	Educational Standards Tests	Years 4 and 8	*Year 4*: German, reading, writing and maths *Year 8*: German, maths, English	Sample	System monitoring
Norway[12]	National tests	Years 5, 8	Norwegian, English & maths	Census	System monitoring Formative

[3] Santiago *et al.*, 2011
[4] *www.cmec.ca/Programs/assessment/pancan/Pages/default.aspx*
[5] Ferrer, 2006
[6] Hout and Elliott, 2011
[7] Eurydice, 2009
[8] Nusche *et al.*, 2011 and Eurydice, 2009
[9] INCA, 2011
[10] Eurydice, 2009
[11] Eurydice, 2009
[12] Eurydice, 2009

REFERENCES

Abrams, L., J. Pedulla and G. Madaus. (2003), "Views from the Classroom: Teachers' Opinions of Statewide Testing Programs", *Theory into Practice*, Vol. 42, No. 1, pp. 18-29.

Alliance for Excellent Education (AEE) (2010), *Principles for a Comprehensive Assessment System*, Policy Brief, *www.all4ed.org/files/ComprehensiveAssessmentSystem.pdf*, accessed 12 July 2011.

Anderson, P. and G. Morgan (2008), *Developing Tests and Questionnaires for a National Assessment of Educational Achievement,* The World Bank, Washington, D.C.

Australian Curriculum, Assessment and Reporting Authority (ACARA) (2011), *About Us*, *www.acara.edu.au/about_us/about_us.html*, accessed 24 August 2011.

Ball, S. (1998), "Big Policies/Small World: An Introduction to International Perspectives in Education Policy", *Comparative Education,* Vol. 34, No. 2, pp. 119-130.

Beguin, A., E. Kremers and R. Alberts (2008), "National Examinations in the Netherlands: Standard-Setting Procedures and the Effects of Innovations," paper presented at the IAEA Conference in Cambridge, *www.iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/180395_Beguin.pdf*, accessed 24 August 2011.

Black, P. and D. William (1998), "Inside the Black Box: Raising Standards through Classroom Assessment", *Phi Delta Kappan,* Vol. 80, No. 2, pp. 139-148.

Booher-Jennings, J. (2005), "Below the Bubble: 'Educational Triage' and the Texas Accountability System", *American Educational Research Journal*, Vol. 42, No. 2, pp.231-268.

Braun, H. (2005), *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models*, Educational Testing Services, *www.ets.org/research/policy_research_reports/pic-vam*, accessed 12 July 2011.

Campbell, C. and B. Levin (2008), "Using Data to Support Educational Improvement", *Educational Assessment, Evaluation and Accountability*, Vol. 21, pp. 47-65.

Carnoy, M. and S. Loeb (2002), "Does External Accountability Affect Student Outcomes? A Cross-State Analysis", *Educational Evaluation and Policy Analysis,* Vol. 24, No. 4, pp. 305-331.

Chiang, H. (2009), "How Accountability Pressure on Failing Schools Affects Student Achievement", *Journal of Public Economics*, Vol. 93, pp. 1045-1057.

Chiodo, A. *et al.* (2010), "Nonlinear Effects of School Quality on House Prices", *The Federal Reserve Bank of St. Louis Review*, Vol. 92, No. 3, pp. 185-204.

Commercial Club of Chicago (2009), *Still Left Behind: Student Learning in Chicago's Public Schools*, *www.civiccommittee.org/Still%20Left%20Behind%20v2.pdf*, accessed September 2011.

Eurydice Network (2009), *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results, http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/109EN.pdf*, accessed 12 July 2011.

Faubert, V. (2009), "School Evaluation: Current Practices in OECD Countries and Literature Review", *OECD Education Working Paper No. 42*, OECD, Paris.

Ferrer, G. (2006), *Educational Assessment Systems in Latin America: Current Practice and Future Challenges*, PREAL, Washington, D.C.

Figlio, D. and S. Loeb (2011), "School Accountability", in E. Hanushek, S. Machin and L. Woessman (eds.), *Handbooks in Economics,* Vol. 3, North-Holland, The Netherlands, pp. 383-421.

Greaney, V. and T. Kellaghan (2008), *Assessing National Achievement Levels in Education*, The World Bank, Washington, D.C.

Green, S. and T. Oates (2009), "Considering Alternatives to National Assessment Arrangements in England: Possibilities and Opportunities", *Educational Research*, Vol. 51, No. 2, pp. 229-245.

Guilfoyle, C. (2006), "NCLB: Is There Life Beyond Testing?" *Educational Leadership*, Vol. 64, No. 3, pp. 8-13.

Hamilton, L. and B. Stecher (2002), "Improving Test-Based Accountability", in L. Hamilton, B. Stecher and S. Klein (eds.), *Making Sense of Test-Based Accountability in Education,* RAND Publishing, Santa Monica, California.

Hamilton, L. and D. Koretz (2002), "Tests and their Use in Test-Based Accountability Systems", in L. Hamilton, B. Stecher and S. Klein (eds.), *Making Sense of Test-Based Accountability in Education,* RAND Publishing, Santa Monica, California.

Hanushek, E. and M. Raymond (2005), "Does School Accountability Lead to Improved Student Performance?", *Journal of Policy Analysis and Management*, Vol. 24, No. 2, pp. 297-327.

Harlen, W. (2007), "Criteria for Evaluating Systems for Student Assessment", *Studies in Educational Evaluation*, Vol. 33, pp. 15-28.

House of Commons (2008), *Testing and Assessment, Volume 1*, The Stationery Office Limited, London.

Hout, M. and S. Elliott (eds.) (2011), *Incentives and Test-Based Accountability in Education*, National Research Council, The National Academies Press, Washington, D.C.

INCA – International Review of Curriculum and Assessment Frameworks Internet Archive (2011), "Korea: Assessment Arrangements" and "Spain: Assessment Arrangements", *www.inca.org.uk/1401.html*, accessed 13 July 2011.

Independent Schools Queensland (2010), "Standardised Testing: Getting it Right for Educational Improvement", *Briefings*, Vol. 14, No. 6, *www.aisq.qld.edu.au/files/files/Communications/briefings/JulyBriefings_10_A4.pdf*, accessed September 2011.

Jacob, B. and S. Levitt (2002), "Rotten Apples: An Investigation of the Prevalence and Predicators of Teacher Cheating", *NBER Working Paper No. 9413*, *www.nber.org/papers/w9413.pdf?new_window=1*, accessed 18 August 2011.

Kane, T. *et al.* (2002), "Volatility in School Test Scores: Implications for Test-based Accountability Systems", *Brookings Papers on Education Policy,* No. 5, pp. 235-283.

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

King, K. and S. Zucker (2005), *Curriculum Narrowing*, Pearson Education, *www.pearsonassessments.com/NR/rdonlyres/D3362EDE-7F34-447E-ADE4-D4CB2518C2B2/0/CurriculumNarrowing.pdf*, accessed 18 August 2011.

Le, V. and S. Klein (2002), "Technical Criteria for Evaluating Tests", in L. Hamilton, B. Stecher and S. Klein (eds.), *Making Sense of Test-Based Accountability in Education,* RAND Publishing, Santa Monica, California.

Lee, J. (2008), "Is Test-based Accountability Effective? Synthesizing the Evidence from Cross-State Causal-Comparative and Correlational Studies", *Review of Educational Research*, Vol. 78, p. 608.

Linn, R. (2005), "Issues in the Design of Accountability Systems", *Yearbook of the National Society for the Study of Education*, Vol. 104, Issue 2.

Looney, J. (2011), "Integrating Formative and Summative Assessment: Progress toward a Seamless System?", *OECD Education Working Paper No. 58*, OECD Publications, Paris.

Meckes, L. and R. Carrasco (2006), *SIMCE: Lessons from the Chilean Experience in National Assessment Systems for Learning Outcomes*, Cartagena de Indias Conference.

Mons, N. (2009), "Theoretical and Real Effects of Standardised Assessment", Background paper to the study: National Testing of Pupils in Europe, Eurydice Network, *http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/111EN.pdf*, accessed 12 July 2011.

National Center for Education Statistics (NCES) (2011a), *Mapping State Proficiency Standards onto the NAEP Scales*, U.S. Department of Education, *http://big.assets.huffingtonpost.com/naepproficiency.pdf*, accessed 23 August 2011.

National Center for Education Statistics (NCES) (2011b), *NAEP Item Development Process*, *http://nces.ed.gov/nationsreportcard/contracts/item_dev.asp*, accessed 24 August 2011.

Newton, P. (2007), *Evaluating Assessment Systems*, Qualifications and Curriculum Authority, *http://orderline.qcda.gov.uk/gempdf/1445900599/Evaluating_Assessment_Systems1.pdf*, accessed 12 July 2011.

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

Nusche, D., G. Halász, J. Looney, P. Santiago and C. Shewbridge (2011), *OECD Reviews of Evaluation and Assessment in Education: Sweden*, OECD, Paris, available from *www.oecd.org/edu/evaluationpolicy*.

Perie, M. *et al.* (2007), *Key Elements for Educational Accountability Models*, paper commissioned by Council of Chief State School Officers, Washington, D.C.

Popham, W.J. (1999), "Why Standardized Tests Don't Measure Educational Quality", *Using Standards and Assessments,* Vol. 56, No. 6, pp. 8-15.

Popham, W.J. (2006), "All about Accountability. Phony Formative Assessments: Buyer Beware!", *Educational Leadership*, Vol. 64, No. 3, pp. 86-87.

Ravitch, D. (2010), *The Death and Life of the Great American School System*, Basic Books, New York.

Rosenkvist, M.A. (2010), "Using Student Test Results for Accountability and Improvement: A Literature Review", *OECD Education Working Paper No. 54*, OECD, Paris.

Santiago, P., G. Donaldson, J. Herman and C. Shewbridge (2011), *OECD Reviews of Evaluation and Assessment in Education: Australia*, OECD, Paris, available from *www.oecd.org/edu/evaluationpolicy*.

Shewbridge, C. (forthcoming), "Evaluating Educational Systems in the OECD Countries: A Review of Country Practices and Related Literature", *OECD Education Working Paper Series*, OECD, Paris.

Shewbridge, C., E. Jang, P. Matthews and P. Santiago (2011), *OECD Reviews of Evaluation and Assessment in Education: Denmark*, OECD, Paris, available from *www.oecd.org/edu/evaluationpolicy*.

Stiggens, R. (2005), "From Formative Assessment to Assessment FOR Learning: A Path to Success in Standards-Based Schools," *Phi Delta Kappan*, Vol. 87, No. 4, pp. 324-328.

Toch, T. (2006), *Margins of Error: The Education Testing Industry in the No Child Left Behind Era,* Education Sector Reports, *www.educationsector.org/publications/margins-error-testing-industry-no-child-left-behind-era*, accessed 12 July 2011.

Visscher, A., S. Karsten, T. De Jong and R. Bosker (2000), "Evidence on the Intended and Unintended Effects of Publishing School Performance Indicators", *Evaluation & Research in Education*, Vol. 14, No. 3, pp. 254-267.

Volante, L. and S. Ben Jaafar (2008), "Educational Assessment in Canada", *Assessment in Education: Principles, Policy and Practice,* Vol. 15, No. 2, pp. 201-210.

Wang, L., G. Beckett and L. Brown (2006), "Controversies of Standardized Assessment in School Accountability Reform: A Critical Synthesis of Multidisciplinary Research Evidence", *Applied Measurement in Education*, Vol. 19, No. 4, pp. 305-328.

Zhang, L. (2008), "Validity of Comparing Test Scores on State Assessments with the Results of the National Achievement of Educational Progress", paper presented at the AERA Annual Conference, New York.

Zucker, S. (2003), *Fundamentals of Standardized Testing*, Pearson Education, *www.pearsonassessments.com*, accessed 12 July 2011.

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.