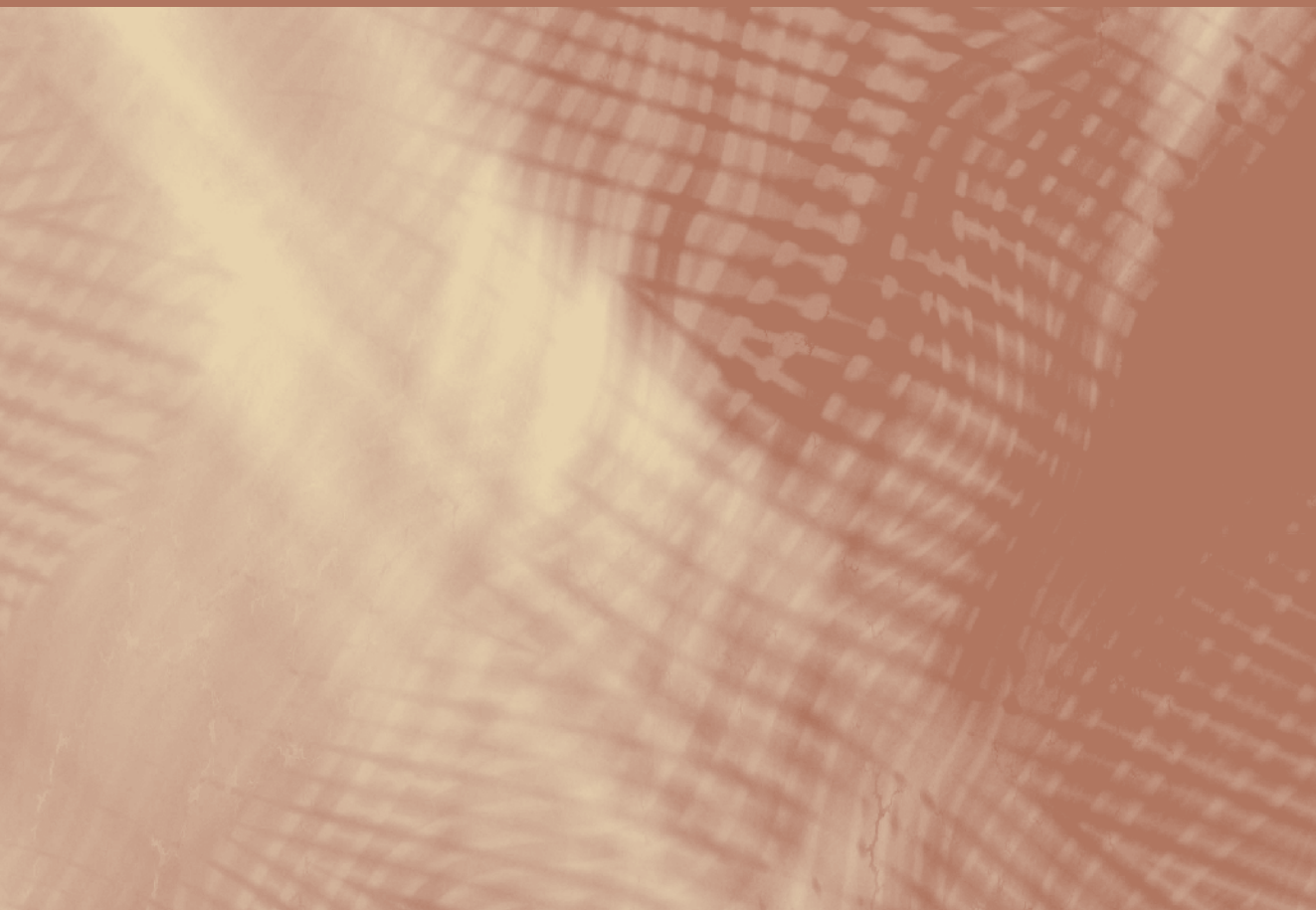
NCCA
An Chomhairle Náisiúnta Curaclaim agus Measúnachta
National Council for Curriculum and Assessment

Research Report No. 12

Standardised Testing In Lower Secondary Education

Gerry Shiel, Thomas Kellaghan, Gráinne Moran



Boyd Freeman Design
www.boydfreeman.ie
boyd@boydfreeman.ie
+ 353 (0)87 2442506

Standardised Testing In Lower Secondary Education

Gerry Shiel

Thomas Kellaghan

Gráinne Moran

Educational Research Centre

St Patrick's College, Dublin

May 2010

Educational Research Centre
Foras Taighde ar Oideachas

Research conducted on behalf of the
National Council for Curriculum and Assessment

© NCCA 2010

ISSN 1649-3362

National Council for Curriculum and Assessment

24, Merrion Square, Dublin 2.

www.ncca.ie

Acknowledgements

The Educational Research Centre thanks the NCCA for commissioning and supporting this project. In particular we thank John Hammond, Deputy Chief Executive, and John Halbert, Director, Curriculum and Assessment, for their helpful advice in compiling this report.

We also thank colleagues at the Educational Research Centre who supported us in carrying out this research, including Peter Archer, Director, Mary Rohan, Administrator, and Hilary Walshe, who provided clerical assistance.

Finally, we thank ministry officials and/or their nominees in Denmark, Finland, France, Norway, the Netherlands and New Zealand, who completed our questionnaire on assessment at lower-secondary level, and patiently responded to our follow-up questions.

Contents

STANDARDISED TESTING IN LOWER SECONDARY EDUCATION

Preface

In July 2009, the Educational Research Centre was commissioned by the NCCA to conduct a desk-based study into current practices in standardised testing in lower secondary schools in a number of countries. The terms of reference of the study asked us to report on:

- the nature of the testing that takes place;
- how the outcomes of testing are recorded and reported to students and their parents/guardians; and
- the impact of testing on teaching and learning in schools.

In relation to the first of these, we were asked to describe the purposes for which testing is carried out; the point of lower secondary schooling at which testing takes place, and whether there is discretion as to when it happens; and the range of competences or areas of student achievement tested. We were also asked to describe the test instruments used and any validation research underpinning their use, and the implications of testing for such areas as professional development, operational issues, and the role of the teacher in administering, marking and reporting.

In relation to the second, we were asked to describe how test results are used, consider relationships between standardised test outcomes and other school-based tests and external examinations at secondary level, indicate any links between standardised tests and large-scale international assessments of student achievement at secondary level, outline protocols governing access to test results, and report on the issue of stakes, by describing the consequences linked to the outcome of tests.

Finally, we were asked to identify whether standardised tests are

administered in the language of instruction, and, where there is more than one such language, what provisions are built into the system in recognition of this.

We used a questionnaire to seek information about standardised testing from education ministries in a number of countries – Denmark, Finland, France, Norway, the Netherlands, Scotland and New Zealand. We wish to acknowledge the help of those who responded. Information on standardised testing in these and in other jurisdictions (Canada (Ontario), England, Northern Ireland, the United States) was also obtained from journal articles, research reports and ministry websites.

While we have made every endeavour to cross-check the information provided in this report, this has not always been possible. Systems of assessment change on an ongoing basis, and an article published a few years ago may no longer present a true picture of the situation in schools and classrooms, while a website updated a year or two ago may now be out of date. We endeavoured to ensure that respondents to our questionnaire understood what we meant by a standardised test, but language differences and assessment traditions in different countries mean that we cannot be sure that they interpreted our questions in the way we had intended.

We were not asked to provide recommendations in this report. Instead, we have endeavoured to provide some options that policy makers can consider as they look at ways in which assessment can be strengthened in lower-secondary schooling.

CHAPTER 1

INTRODUCTION

It is difficult to envisage a description of teaching that does not accord assessment an essential role. Teachers need to continually collect, synthesise, and interpret information about their students' learning. They need to know the state of knowledge and skills of their students before they can begin to plan instruction and they need evidence as instruction proceeds that students are, or are not, learning. This evidence is based for the most part on teachers' own observations and monitoring of students in the classroom (e.g., the quality of students' written work, their responses to questions) and is used for a variety of purposes: to plan future instruction; to adapt teaching to learning styles, skills, interests, and motivations of students; to provide feedback and incentives; to place students in instructional groups; and to diagnose problems that students may be experiencing (see e.g., Airasian, 2001; OECD, 2005a).

While much assessment activity by teachers is ongoing and often intuitive, there is a long history in some countries of providing additional information on student achievement obtained from externally devised standardised tests. As these tests usually provide norm-referenced information, they allow teachers to compare the achievements of their own students with those of a reference group outside the school. Some tests also provide information which indicates the extent to which students are achieving curriculum targets or information that identifies particular problem areas in students' achievements.

Over the past two decades, there has been considerable interest in examining how the assessment capacity of teachers might be enhanced to improve student learning (see, e.g., Black & Wiliam, 1998; Gipps & Stobart, 2003). This interest was often accompanied by an effort to shift teacher dependence for assessment information from standardised tests based on psychometric models to other forms of assessment (e.g., 'authentic' performance-based assessment,

portfolios, student self-assessment). However, government investment in several countries (in particular, the United States and the United Kingdom) over this time has not been to support such activity, but rather to extend the ways in which information derived from standardised tests can be used and to privilege the information such tests provide.

This development is illustrated in a number of features of recent reforms involving assessment. Firstly, the administration of tests is mandated by an agent outside the school, usually a national government. Secondly, testing is controlled or monitored by an agent outside the school. Thirdly, the assessment is primarily concerned with obtaining summative information about student achievement that can be aggregated to provide a basis for a judgment about the quality of education at the level of the school, state, or national education system. Fourthly, the assessment exercise is expected to not just obtain information about education systems, but to be a lever of reform. Thus, on the basis of assessment findings, policy decisions may be made to adjust standards, to review curricula, or to provide additional resources to schools.

In Ireland, the Curriculum and Examinations Board (1986) recommended that schools be provided with appropriate assessment techniques, tests, and support services in recognition of the important role that assessment plays in promoting student learning. Specific reference was not made to standardised tests. At the time, and into the 1990s, policy relating to standardised testing focused on use at the primary school level. In the green paper on education, *Education for a Changing World* (1992), it was proposed to extend standardised testing to all primary schools as a diagnostic aid. Its primary purpose would be to support efforts by teachers to identify students in need of special assistance and the nature and extent of the assistance needed. It was anticipated that it would provide a further safety net

to those who might be experiencing basic literacy or numeracy problems. Tests at ages 7 and 11 were considered to be most appropriate for this purpose (p. 175). In the white paper, *Charting Our Education Future* (1995), influenced in part by concerns raised in the Report on the National Education Convention (1994) where the problem of under-performance in schools was raised, it was stated that

> All primary schools will be required to develop a policy on assessment within the framework of the school plan. The policy should ensure uniformity and continuity of approach between classes and within the school. Under the direction of the school principal, students will be assessed by their teachers at the end of first and fifth classes in order to evaluate the quality of their learning and to identify any special learning needs that may arise (p. 28).

In the 1998 Education Act, in which it was stated that the 'principal and teachers shall regularly evaluate students and periodically report the evaluation to the students and their parents' [Section 22(2)], no specific reference was made to standardised tests. However, the DES (2006) circular (0138/2006) to primary schools identified standardised tests as one of several tools that a school should use in meeting its obligations under Section 22 of the Act. The circular requested schools, beginning in the 2007 calendar year, to administer standardised tests to students in two curriculum areas, English reading and mathematics, at the end of first class/beginning of second and the end of fourth/beginning of fifth class. The primary purposes of testing were identified as informing parents of students' progress and assisting in the identification of students who may require support. Funding was provided to schools to purchase tests and ancillary materials. The results of tests were to be maintained by the school,

and made available to DES officials, though inspectors in their reports could not make reference to test data that might facilitate school comparisons or the compilation of league tables. The outcomes of testing were to be reported to parents in respect of their own children, with effect from the 2007/08 school year, in accordance with a reporting template developed by the National Council for Curriculum and Assessment (NCCA).

The NCCA (2007) guidelines on assessment in primary schools identified standardised testing as one of eight methods of assessment[1]. Key terms such as 'standardised test', 'standard score', and 'percentile rank' were defined. Suggestions on ways in which standardised test scores could be reported to parents were provided. Templates placed on its website by the NCCA included strategies for reporting the results of standardised tests and other assessments to parents.

During the time that these developments occurred, policy vacillated somewhat between the use of assessment to ensure greater openness and accountability and maximising parental involvement, on the one hand, and endorsing a model of assessment that prioritised its formative purposes and the central role of the teacher on the other hand (Hall, 2000).

In communications with the Minister for Education and Science, the NCCA undertook to extend its focus on assessment practices beyond the primary to the post-primary sector. In pursuit of this objective, it proposed gathering information on international practice on testing for students in post-primary schools (aged 12 to 15 years) with a view to advising on the implications of introducing standardised tests at one further point during the course of compulsory education. The study described in this paper was carried out in response to a request to the Educational Research Centre from the NCCA to obtain the required information.

1 The other methods were identified as self-assessment, conferencing, portfolio assessment, concept mapping, questioning, teacher observation, and teacher-designed tasks and tests.

To set our study in context, we begin with a description of a standardised test (Chapter 2). On the basis of the international literature, we describe the criteria that have to be met if a test is to be considered standardised. Two key concepts (validity and reliability) which merit consideration in deciding on the appropriateness of an assessment in any situation are considered. The extent to which the validity or reliability of a procedure needs to be established will depend on the seriousness of the decision which follows an assessment.

We consider the use of standardised tests in three contexts: classroom use by teachers in which the achievements of individual students are of primary concern; use to obtain information that describes the achievements of students in the education system as a whole (national assessment); and use to obtain information that allows a comparison of the achievements of students in a number of countries (international assessment).

Following this, still with the context of our study in mind, we provide a brief outline of the history of the development of standardised tests, and of growth in their use (Chapter 3).

In recognition of the fact that the use of standardised tests has for many years been a topic of controversy, we outline perceived advantages and disadvantages of their use in Chapter 4. We also report the findings of a study carried out in Irish schools relating to seven frequently expressed statements about the effects of standardised testing.

In Chapter 5, we present the results of our enquiry into the use of standardised tests in selected countries. Information was obtained in a questionnaire about seven education systems (Denmark, Finland, France, the Netherlands, Norway, and New Zealand). Information was not sought from England or the United States, partly because

considerable information was already available in the literature but, of greater significance, because the kind of high stakes testing being carried out in those countries did not seem appropriate, or acceptable, in an Irish context. Finally, information on the use of standardised tests in Northern Ireland, Scotland and Ontario (Canada) was obtained from published and web-based sources.

In Chapter 6, we explore the utility of international studies and describe the results of research that attest to their value in identifying issues in national education systems that merit the attention of policy makers and school personnel.

In Chapter 7, we present a range of options relating to the introduction of standardised testing at lower secondary level.

CHAPTER 2

WHAT IS A STANDARDISED TEST?

Tests (or examinations) take a variety of forms, ranging from informal quizzes in the classroom to formal assessment, which may be written, oral, or practical, in a public examination. Most tests involve sampling some aspect of a test taker's knowledge or skills, on the basis of which an inference is made about his/her probable performance in the domain (the body of knowledge or set of skills) from which the sample was drawn. The inference, in turn, may be used to describe or make decisions about an individual or group of test takers (see Anastasi, 1954; Crocker & Algina, 1986; Ebel, 1972; Madaus, Russell, & Higgins, 2009; Osterlind, 1989).

Tests vary in a number of ways, in particular in the extent to which

- the domain being assessed is clearly described;
- the domain being assessed is adequately sampled;
- conditions for administration are identical for all test takers;
- scoring is not influenced by the person administering the test;
- guidance on interpretation of the test taker's performance is available.

In this chapter, we describe a form of test usually referred to as a standardised test that attempts to meet all the conditions.

Definition

A standardised test is a procedure designed to assess the abilities, knowledge, or skills of individuals under clearly specified and controlled conditions relating to (i) construction, (ii) administration; and (iii) scoring, (iv) to provide scores that derive their meaning from an interpretative framework that is provided with the test. Some definitions specify only administration, scoring and interpretation (e.g., NCCA, 2007; Popham, 1995). However, aspects of construction

are also important, particularly in the context of establishing validity. Standardised tests differ from other forms of student evaluation in one or more of these characteristics. They were in fact developed in the early years of the 20th century to address the perceived shortcomings of tests and examinations in use at the time (in particular, essay-type examinations).

Test Construction

The first requirement in the construction of a standardised test is to describe the domain or construct (ability, body of knowledge, set of skills) that is to be assessed. In the case of an achievement test, this will most likely involve a review of curriculum documents, instructional materials, and textbooks. Following the review, the domain may be represented in a table of specifications or a blueprint consisting of a matrix in which content (specific subject matter) is crossed with process (what the student can do with the subject matter) (see Bloom, Hasting, & Madaus, 1971). Table 2.1 presents an example of a table of specifications, in this case one developed for a third/fourth grade mathematics test (Educational Research Centre, 2007).

Skills	Content Strands				
	Number	Algebra	Measures	Shape & Space	Data
Understanding & Recalling	5		1	2	
Implementing	8		2	1	
Integrating & Connecting		2	1	1	4
Reasoning	6	4	5	9	5
Applying & Problem Solving	8		11		

As a test can contain only a small sample of the knowledge and skills that students are expected to acquire in a curriculum area, it is extremely important that the tasks/questions selected for the test

provide an adequate representation of the curriculum. Otherwise, it will not be possible to infer from a student's performance on the test his/her achievement in the entire domain being assessed. Table 2.1 identifies the number of items in each cell of the content by skills matrix for a third/fourth grade mathematics test. Perceived importance of the content/skills is reflected in the number of items in each cell.

The next step in the construction of a standardised test is to field-trial items in a small sample of students that spans the variation in achievement of the students for whom the test is intended. A larger number of items than will be included in the final test is required for this exercise as some items will, inevitably, be found to be unsuitable. Traditionally, the results of item analysis based on classical test theory were used to select items for the final form of a test. The criteria used were the difficulty level of items (the proportion of students in the sample who got the item right) and their discriminating power (the relationship between performance on an individual item and performance on the test as a whole). Since classical test theory does not adequately model answers to individual items, item response modelling, which is based on the assumption that a single trait underlies performance, and specifies how the probability of answering a specific item correctly depends on the attribute being measured, is increasingly used.

The final version of a test is administered to a representative sample of the population for whom the test is intended (e.g., fourth grade students) to establish norms (e.g., average performance, relative frequency of varying degrees of deviation from the average).

Administration

Standardised tests require uniformity of procedure in their administration. The materials used, instructions to test takers,

preliminary demonstrations, and ways of handling queries are all clearly specified.

Furthermore, the conditions under which a test is administered relating to comfort, lighting, freedom from distraction, and student interest, co-operation, and motivation should be the same for all examinees.

Deviations in administration or in the conditions of testing will affect interpretation of examinees' performance.

Scoring and Aggregation of Scores

The precise instructions for scoring in the manual accompanying a test must be followed exactly.

Discretion on the part of the examiner is eliminated when selection-type items are used in which the examinee is required to select one correct option from a limited number of options (e.g., in multiple-choice items). Tests with this type of item can be, and frequently are, scored by machine, increasing the speed and reducing the cost of the operation. A further advantage of selection-type items is that they allow wide sampling of a domain since responding to an item requires very little time.

Since selection-type items may not provide a full measure of the knowledge and skills represented in the domain on which the assessment is based, supply-type items may be included in a test. Such items require the test taker to supply an answer (involving recall, analysis, synthesis of information, evaluation), usually in an essay or short written response. These items are considered more appropriate to elicit higher order thinking skills (involved in analysis, synthesis, evaluation). While preset criteria to evaluate responses will be provided, scoring will not be as 'objective' as in the case of selection-type items, giving rise to problems of reliability.

Interpretation of Test Scores

Standardised tests are presented with one of two interpretative frameworks. In the first, it is possible to locate the relative position of an examinee's score in a distribution of scores. In this case, the standard used in interpreting test performance is a relative one, and the score given to an examinee is called a *norm-referenced measure*.

An alternative interpretative framework is provided when performance on a test describes the degree to which the performance of an examinee meets an established standard, criterion, or proficiency level (Glaser, 1963). For example, if a simple test of addition facts consisted of 50 items chosen randomly from all possible items, a test taker's proportion-correct score could be considered to be an estimate of his/her knowledge of addition facts. Interpretation in this case does not require information on how other test takers performed. The proportion correct score is called a *criterion-referenced measure*, which is sometimes used to classify test takers as having achieved 'mastery' or not having achieved mastery.

A variety of score conversions are provided in test manuals to facilitate inter-individual comparisons when norm-referenced tests are used (Crocker & Algina, 1986). These include:

- *percentile rank* (the percentage of examinees in the norm group scoring at or below a given raw score)
- *derived or standard score* (linear transformation of z-scores to an arbitrary mean (e.g., 100) and standard deviation (e.g., 15)
- *scaled score* (reflects an examinee's score relative to the norm group and the location of that norm group's distribution in relation to that of other group distributions, often examinees at a higher grade)

A description of student performance in terms of proficiency levels, which combine aspects of norm-referencing and criterion-referencing, is increasingly being used to present the results of national and international assessments (see, e.g., OECD, 2005b, Chapter 16). Division of a continuum of achievement into levels involves scale anchoring which has two components: a statistical component that identifies items that discriminate between successive points on the proficiency scale using specific item characteristics (e.g., the proportions of successful responses to items at different score levels) and a consensus component in which identified items are used by curriculum specialists to provide an interpretation of what groups of students at, or close to, the related points know and can do (Beaton & Allen, 1992).

Validity and Reliability

Two related concepts, validity and reliability, need to be considered in evaluating all assessment activity, including standardised testing.

Validity

Validity, according to Crooks, Kane, and Cohen (1996) is 'the most important consideration in the use of assessment procedures' (p. 265). According to *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), it refers to

> the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of the test…it is the interpretation of test scores required by proposed uses that are evaluated, not the test itself. When test scores are used to interpret in more than one way, each intended interpretation must be validated. (p.9)

Part of the test validation process involves providing a conceptual

framework for the test by 'delineating the knowledge, skills, abilities, processes, or characteristics to be assessed' (p. 9). Central to this is the concept of construct validity, in which the construct being measured (such as mathematical achievement) is clearly distinguished from other related constructs (see Messick, 1989). As issues such as construct underrepresentation (the failure of a test to capture important aspects of the construct), and construct irrelevant variance (the degree to which test scores are affected by processes that are irrelevant to the intended construct) are examined, the process of validation may lead to revisions to the test as well as the underlying conceptual framework. Validity is seen as being a joint responsibility of the test developer and the test user. According to the *Standards*, 'when use of the test differs from that supported by the test developer, the test user develops special responsibility for test validation' (p. 11).

Several types of evidence can be drawn on to support test validity. These include

- *Evidence based on test content,* such as analyses of the relationship between test content and the construct (domain) it is intended to measure. Expert judgment of the appropriateness of test content is one type of evidence that might be provided.

- *Evidence based on internal structure,* including the extent to which test items and test components conform to the construct on which test score interpretations are based. Evidence of the unidimensionality of a test would contribute to this, as would information on differential item functioning (i.e., if different groups of examinees with similar overall scores have systematically different average responses to an item).

- *Evidence based on the relationship of performance on a test to other variables*, such as some criterion the test is expected to predict

(predictive evidence) or not predict (discriminant evidence). It can also include information on relationships between performance on a test and a measure designed to assess the same domain (evidence of concurrent validity).

- *Evidence based on the consequences of testing*, such as the effects of placing students in a learning support programme or special education class. The use of test scores can be shown to be valid if participation in the programme benefits students. The effects of other uses of testing, such as to increase accountability, also need to be assessed.

The idea that the consequences of testing should be taken into account in validating a test is relatively new, and is not universally accepted (e.g., Lissitz & Samuelsen, 2007). In considering this form of evidence, it is useful to distinguish between an intended consequence (e.g., achievement improves after a period of time in the instructional group to which students were assigned on the basis of their performance on a test) and an unintended consequence (e.g., when a decision based on test performance leads to 'labelling' of students or affects their self-concept negatively).

As estimation of validity is dependent on human judgment, it is often very difficult to do. Drawing on work relating to the identification of sets of criteria (see, e.g., Frederickson & Collins, 1989) and a 'validity argument' proposed by Cronbach (2000), Crooks et al. (1996) identified threats to the interpretation and use of assessment data for eight components of the assessment process. While the threats are most likely to be considered in the context of standardised tests, they merit consideration in any type of assessment or evaluation.

1. *Administration of assessment tasks/tests.* For example, some students may receive inappropriate help; others may not be motivated to respond to tasks.

2. *Scoring of students' performance on tasks/tests.* A threat will be present if a scoring rubric takes account of some qualities of performance, ignoring others (e.g., in an oral language test, vocabulary span is credited, but fluency or pronunciation is not).

3. *Aggregation of scores on individual tasks/items to produce one or more aggregated (total or subscale) scores.* For example, the weights given to tasks/items in an assessment do not reflect the relative importance of the tasks in the domain being assessed, as occurs when differences in score variance for different tasks are not recognised in calculating total scores.

4. *Generalisation from the particular tasks on which an aggregate score is based to the whole domain of similar tasks*. If the size of the sample (number of items) drawn from the assessed domain is too small, it will not be possible to generalise from the student's score to his/her universe score in the assessed domain.

5. *Extrapolation from the assessed domain to a target domain containing all tasks relevant to the proposed interpretation*. If no tasks are included from some substantial sections of the target domain (resulting in construct under-representation), it will not be possible to extrapolate from a universe score for the assessed domain to a universe score for the target domain. This will be the case if adequate attention in the assessment is not accorded the content coverage, content quality, and cognitive complexity represented in a curriculum.

6. *Evaluation of the student's performance.* Inappropriate judgments on the basis of assessment information will be made if the person evaluating it does not understand the information or the limitations arising from its relative nature or the particular arrangements used to collect it.

7. *Decision on actions to be taken in light of judgments.* Threats to validity arise if standards used in making decisions are inappropriately high or low, if inappropriate pedagogical decisions are made, or if inappropriate feedback is provided to students.

8. *Impact on the student and other participants arising from the assessment process, interpretations, decisions, and consequences of assessment.* Threats would arise if, as a the result of the assessment, a teacher neglected important curriculum areas to align her/his teaching with the demands of the test, if the teacher formed inappropriate expectations for students, if student motivation was reduced, or if teaching and learning focused on the acquisition of factual knowledge at the expense of higher-level cognitive outcomes.

Reliability

According to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), reliability refers to consistency of the measurement when a testing procedure is repeated on a population of individuals or groups. Central to this is the concept of measurement error – the unsystematic error that arises because a student is tested on a particular set of items in a particular context. Such error may also be due to inconsistencies in scoring open-ended items. Systematic error (e.g., error because one form of a test is easier than another, and the two forms have not been properly equated) is not regarded as measurement error. More formally, 'the hypothetical difference between an examinee's observed score on any particular measurement and the examinee's true or universal score for the procedure is called measurement error' (AERA, APA, NCME, 1999, p. 25).

In providing evidence to support the reliability of test scores, test users are expected to identify 'the major sources of error, summary statistics bearing on the size of such errors, and the degree of

generalisability of scores across alternate forms, scorers, administrations and other relevant dimensions' (p. 27). Statistics such as standard error of measurement (the standard deviation of a hypothetical distribution of measurement errors) should be produced and reported. This may be based on an internal consistency coefficient, an alternate forms coefficient, or a test-retest coefficient (see Feldt & Brennan, 1989). If item response modelling is used, the test information function (an average precision of measurement at each level of a trait, based on a set of items) should be reported.

A key issue in interpreting standardised test scores relates to proficiency levels, and whether students close to a cut score (the dividing point between two adjacent levels) belong to one level or the other. An analogous situation occurs in the case of a student on the borderline between an A and a B grade in an examination. While the incorrect assignment of a 'borderline' student to a proficiency level will have no consequences for the student in a national or international sample survey, incorrect assignment could have significant consequences on a test designed to allocate the student to a course or programme of study.

It should be noted that the scores derived from some standardised tests may not provide reliable estimates of achievement at the individual student level. The individual scores achieved by students in sample-based national or international assessments, while suitable for generating reasonably accurate population estimates (e.g., overall mean scores for a country, overall mean scores for male and female students), often cannot be used to report on individual student performance. One reason for this is that students may be tested on a small part of the domain of interest, and may not attempt enough items to yield a reliable estimate of performance across that domain. This occurs in the Programme for International Student Assessment (PISA), where students taking a two-hour test measuring

achievement in several domains may respond to only 10 to 15 mathematics items.

The Contexts in Which Standardised Tests are Used

We will consider the use of standardised tests in three contexts throughout the rest of this report (in describing the history of testing and in our investigation of the use of tests in other countries):

- administration by teachers of tests in their classrooms to support student learning (classroom assessment);
- administration of a national assessment;
- administration of an international assessment.

Classroom Assessment. Tests designed to provide information to teachers which, in conjunction with other sources, can be used in a variety of activities relating to teaching and learning are sometimes referred to as formative assessment instruments (see OECD, 2005a). The information they provide may be used to monitor student progress, to diagnose student learning difficulties, to adapt teaching to student needs, and to allocate students to instructional groups. Formative assessment can be contrasted with summative assessment which has as its primary goal grading or certifying students, judging the effectiveness of a teacher, or comparing curricula (Bloom et al., 1971). Bloom et al. (1971) actually distinguished between diagnostic and formative evaluation. The former refers to determining the presence or absence of prerequisite skills, students' level of mastery, and underlying causes of learning difficulties. The latter refers to the process of providing feedback on a student's progress during instruction.

National Assessments. Over the past twenty years, there has been a dramatic increase in the number of countries that are using

standardised tests in what have come to be known as national assessments, to provide an overview of the extent that students in the education system as a whole have acquired knowledge and skills. Although a national assessment requires the participation of individual students, the focus of interest is on the aggregation of data collected from the students, not on the performance of individual participating students. Specific questions addressed in a national assessment include: (a) How well are students learning with reference to general expectations, the aims of the curriculum, or preparation for life? (b) Is there evidence of particular strengths or weaknesses in students' knowledge and skills? (c) Do particular subgroups in the population perform poorly? (d) What factors are associated with student achievement? (e) Do the achievements of students change over time? (Kellaghan & Greaney, 2001).

The findings of a national assessment are intended to be used primarily as a basis for formulating education policy and as a means of improving the management of the education system at its varying levels. Today, standardised assessment of educational achievement is considered to be an essential component of a comprehensive educational assessment system.

National assessments are either sample-based or census-based. In a sample-based assessment, students in schools are selected to be representative of the specified grade or age levels that are the focus of the assessment. In a census-based assessment, all (or nearly all) schools and students, usually at specific grade or age levels, participate. Two purposes related to the design of a national assessment can be identified. In the first, which may be termed *diagnostic monitoring*, an attempt is made to identify problems in the education system, following which efforts will be made to address such problems. A variety of resources (new programmes, new educational materials, inservice for teachers) may be provided. An alternative purpose may

be termed *performance monitoring*. In this approach, based on principles of microeconomics, the focus is on organisational outcomes and the objective is to improve student achievement primarily through competition. No specific action may be required beyond the publication of information about performance (e.g., in league tables), though inducements for improved performance may also be provided. For example, schools and/or teachers may receive money if students reach a specified target (e.g., if 85% of students reach a satisfactory level of proficiency). Whether a national assessment can be used for diagnostic or performance monitoring depends on its design. If based on a sample of schools/students, it can be used only for diagnostic purposes, and then only for diagnosis at the system level, or, if the sample is sufficiently large, for subpopulations in the system (e.g., urban and rural students, students in different regions, students attending different types of school). Census-based assessments, on the other hand, may be used for both diagnostic and performance monitoring (Kellaghan, 2003).

International Assessments. International assessments of student achievement are designed to provide information on standards of student achievement in a number of countries, and individual countries can compare the performance of their students against average international performance or against the performance of students in other countries. They share many procedural features with national assessments, though they also differ from them in a number of respects, most obviously in the fact that they have to be designed to allow administration in more than one country. As in national assessments, standardised tests are developed in international assessments to assess students' knowledge and skills. However, instead of representing the curriculum of only one education system, the tests have to be considered appropriate for use in all participating countries. The age or grade at which tests are to be administered has to be agreed, as have procedures for selecting schools and students.

International studies have all been based on samples of students (see Beaton et al., 1999).

International assessments have been carried out for the past half century, during which the number of participating countries has grown dramatically, especially during the last decade.

Conclusion

The development of standardised tests represents a serious effort to make student assessment more objective. Such tests were not intended to replace other forms of assessment, which teachers need to use in the day-to-day practice of their pedagogy in the classroom.

Aspects of test construction are highly technical but they need not concern the user. What is important for the user is to develop the competence to select an appropriate instrument, to be aware of the conditions that should obtain during administration, to learn how to interpret and report scores, and to be aware of the limitations of tests and the undesirable, if unintended, consequences that can follow their use. Students and parents are likely to need assistance in the interpretation of scores.

CHAPTER 3

HISTORY OF STANDARDISED TESTING

The origin of standardised tests as we now know them can be traced back to a number of features of education and psychology in the late 19th and early 20th centuries: written essay-type examinations (introduced to select students for university and government personnel); early psychological testing, mostly designed in the context of the study of individual differences to measure sensation, discrimination, and reaction time (associated with Francis Galton and James McKeen Cattell); the development of statistical methods, in particular correlation methods (associated with Karl Pearson); and testing to diagnose mental retardation (associated with Alfred Binet and Theophile Simon) (Du Bois, 1970). The tests of Binet and Simon were particularly germane to future developments as they consisted of a wide range of separate items, using different types of material, and were designed to assess higher mental processes, such as memory span, problem solving, and judgment. In the selection of items for inclusion in a test, consideration was given to their difficulty level and to independent criteria relating to their appropriateness (the age of the testee and judgments of his/her intelligence) while detailed instructions were provided for administration and interpretation.

While the tests of Binet and Simon were individually administered and focused on intelligence, it seemed only a matter of time until tests of achievement that could be administered to groups would be developed. Some efforts were made to develop such tests in the early decades of the 20th century. However, it was not until the need for large-scale testing arose during World War 1 for the selection and placement of personnel in the U.S. army that the first group test was developed. Development of the test was facilitated by the invention, attributed to Frederick J. Kelly, of the multiple-choice format in 1914. Soon after this, the first group test (of intelligence) which made extensive use of multiple-choice items that could be scored objectively (using stencils) was developed by Arthur Otis, which then became the prototype for the Army Alpha test. A parallel nonverbal

group test (Army Beta) was developed for use with individuals with literacy problems or whose first language was not English.

After the war, and through the 1920s, tests of achievement in a variety of curriculum areas (arithmetic, English composition, spelling, handwriting) were developed for use in American schools. Tests were designed primarily to assess individual students, but test data were also aggregated to assess curricula and later to evaluate the efficiency of teachers and school systems in delivering the curriculum. This use declined in the 1930s when tests were used extensively, but almost exclusively, to make judgments about individual students – to assign grades, to diagnose learning difficulties, and to place students in instructional groups.

A rapid and dramatic growth in objective testing in the United States followed the Second World War (Lindquist, 1969). National and state programmes of testing were facilitated by the availability of new technologies, in particular high speed data processing devices and optical scanning. These developments relieved teachers and school administrators of clerical burdens (e.g., hand scoring, converting scores), provided fast turnaround, and allowed more detailed analysis of test data (e.g., tabulating score distributions and responses to sets of items for classes or groups of students for diagnostic use in improving instruction or for curriculum development).

The extent of standardised testing in the U.S. school system in 1967 is evident in the fact that over 68 million test booklets were bought for a school population of 48 million students (Gardner, 1969). Test results were used by teachers to compare the performance of their students with normative data; to identify curriculum areas that might be in need of particular attention; and to compare the end of year performance of students with their beginning of year performance to determine growth and particular areas of effectiveness and non-effectiveness.

In subsequent years, the number of students sitting standardised tests in the U.S. increased while the functions of testing expanded. The increase can be attributed to a growth in the number of states authorising statewide assessments and minimum competency testing. Rather than using off-the-shelf tests as previously had been the case, state testing programmes were more likely to establish a contract with a company to build a test battery to specification (Madaus & Raczek, 1996). Haney, Madaus & Lyons (1993) estimated that as many as 395 million tests were administered annually in the education sector in the 1990s. The use of aggregated standardised test data to make judgments about school systems, which was a feature of testing in the 1920s, was revived in the closing decades of the 20th century in national and international assessments of student achievements which are now a feature of a great many education systems throughout the world.

Information on the use of standardised testing in European countries is difficult to come by. Limited information for the 1960s for a number of countries is available in the proceedings of an international conference on educational measurement held in Berlin from May 16 to 25, 1967 (Ingenkamp, 1969a). Among the countries represented at the conference, the most extensive use seems to have been in France and Sweden. In France, group tests of aptitude (verbal, numerical, spatial) and of achievement were administered in the fifth year of schooling at the point of entry to secondary education and at the end of the ninth year. It was estimated that between a third and a half of students in the relevant grade levels had sat for the tests (Bacher, 1969). Tests were administered by psychologists and counsellors and were used for educational and vocational guidance. The use of standardised tests by teachers did not seem to be a feature of the system.

The use of standardised tests in Belgium was very similar to use in

France. A battery of group tests of aptitude and achievement were administered by centres for educational and vocational guidance, not school authorities. Test results were used for guidance in the last grade of primary school or in the first grade of secondary school, and again in the last grade of secondary school. It was estimated that in 1965, nearly half the school population in the relevant grades sat standardised tests (Stinissen, 1969).

The situation in Sweden, where standardised tests in basic curriculum areas were available to teachers, can be contrasted with the situation in France and Belgium. Use was not mandatory, but most teachers used the tests to diagnose students' readiness to commence school at age 7, to assess students' reading comprehension at grades 4 and 7, to diagnose reading difficulties in lower primary grades, and to provide guidance in curriculum choice at the end of grade 6. To obtain maximum teacher co-operation and to avoid coaching or other unwanted effects of testing, there was no requirement to report test results to anyone (students, other teachers, principal teachers, parents) (Henrysson, 1969).

Despite some efforts to introduce standardised tests in the 1920s and again by the U.S. military government after World War 2, standardised tests were not used to any extent in Germany. It seems that the ethos of German schools was not hospitable to what might be regarded as 'empirical' data (including psychometric data) (Ingenkamp, 1969b).

Standardised Testing in Ireland

Up to the 1960s, standardised tests had been used in a number of research studies in Ireland (e.g., Kelly & McGee, 1967; Macnamara, 1966). They had also been used in some schools, in particular special schools, for the diagnosis of learning problems and in post-primary schools where the Differential Aptitude Tests were used for educational and vocational guidance. Two particular drawbacks

associated with the use of such tests were recognised: the fact that achievement tests might not reflect the content of curricula in Irish schools, and the absence of normative data based on the performance of Irish students. The latter situation was associated with a finding that Irish teachers regarded the general progress of a large proportion of their students as unsatisfactory, suggesting that in the absence of norm-referenced information, teachers held unrealistic standards for students (Kellaghan, Macnamara, & Neuman, 1969).

This situation might be interpreted as indicating a need to develop standardised tests in Ireland, both for teacher use and for research purposes. The latter need was recognised when the Department of Education supported the establishment of the Educational Research Centre in 1966. However, before embarking on a programme to develop tests for research and, in particular, for use in schools, the Centre took advantage of an interest (particularly in the United States) in resolving some of the issues surrounding the use of standardised tests, and in particular their effects. Funds to support a randomised controlled field study, designed by the Educational Research Centre and Boston College, were obtained from a number of philanthropic foundations (Kellaghan, Madaus, & Airasian, 1982). Some of the findings of the study are reported in Chapter 4.

Development of Standardised Tests for Use in Classrooms

The study of the effects of standardised tests (Kellaghan et al., 1982) required the development of a range of tests designed to support teaching and learning in the classroom. A large test development programme, funded by the Department of Education, commenced in the early 1970s. The tests spanned primary grades 2 to 6 and the first three grades of post-primary schooling. Tests were designed to assess student achievement in Mathematics, Irish, and English (except at 2nd class primary where there was no Irish test).

In the late 1980s, two new test series were published – the MICRA-T (Reading) and SIGMA-T (Mathematics). The publication of these tests, which were developed by the Curriculum Development Unit at Mary Immaculate College, Limerick, meant that schools and teachers had a choice when it came to selecting tests. The tests differed from the Drumcondra tests developed in the 1970s in a number of ways, including the use of short-answer as well as multiple-choice items, the availability of procedures for converting scores to reading ages, and the use of cloze procedures to assess reading comprehension. No new tests of Irish reading were published.

In the early 1990s, the Educational Research Centre revised its reading and mathematics tests for primary schools. The Drumcondra Primary Reading Test (for classes 1-6) was published in 1994-95 and the Drumcondra Primary Mathematics Test (also for classes 1-6) in 1997. An Irish-language version of the Mathematics Test was also produced.

Following the introduction of the revised Primary School Curriculum (DES/NCCA, 1999), the Drumcondra Primary Reading and Mathematics Tests and the MICRA-T and the SIGMA-T were revised and renormed. The revised tests included some new features designed to make them more useful to schools and teachers. For example, test-wide scales accompanied the revised Drumcondra Primary Reading and Mathematics Tests (Educational Research Centre, 2007, 2008), making it possible to track the performance of students over several years, while the revised Drumcondra Primary Mathematics Test also included test-wide and class-level proficiency levels, allowing teachers to access a description of the skills that students at different levels of performance were likely to possess.

Unlike at primary level, where the most widely available group-administered standardised tests were revised in line with curriculum

change, there have been no such developments at post-primary level. The Drumcondra Attainment tests in English, Irish, and Mathematics (Levels IV-VI) have not been revised since the late 1970s, although they continue to be used in some schools. However, some work has been done on the development of ability tests. The Drumcondra Verbal Reasoning Test (Educational Research Centre, 1968) was replaced in the late-1990s by the Drumcondra Reasoning Test (Educational Research Centre, 1998). The test, which includes subtests of verbal reasoning and numerical ability, was normed on students in sixth class in primary schools, and first and second year in post-primary schools, and is used by schools to assess students in transition from primary to post-primary schooling. Other tests, such as the Differential Aptitude Test, which is used for educational and vocational guidance, have been re-normed in Ireland.

Development of Standardised Tests for Use in National Assessments

National assessments in Ireland have been a feature of the education system since the 1970s, but only at the primary school level. Practice was endorsed in the white paper, *Charting Our Education Future* (1995), which advocated a system of monitoring student achievement standards based on the regular assessment of the performance of a representative sample of schools. From their inception up to the early 1990s, the main tests used to assess reading in these assessments, which were conducted by the Department of Education, had been standardised in Britain (in particular, the NS6) and tended to focus on word- and sentence-level understanding. These allowed the Department to track standards over time and to compare the performance of students in Ireland with students in Britain (mainly test standardisation samples) (see e.g., Department of Education, 1991; Mulrooney, 1986). The tests, it should be noted, were designed to assess individual student achievement, not for system monitoring. In 1993, and in subsequent national assessments of English reading, a

new test (TARA, Tasks for the Assessment of Reading Achievement) developed at the Educational Research Centre was used. TARA was influenced by trends in test development in other English-speaking countries (e.g., the work of the Assessment Performance Unit in Britain), and allowed students to demonstrate a broad range of reading skills across a variety of text and question types (Cosgrove, Kellaghan, Forde, & Morgan, 2000). Irish norms were established for the test, which was used again in 1999, and in modified form in 2004, to reflect changes in emphasis brought about by the 1999 Primary School English Curriculum.

The earlier national assessments of mathematics achievement, administered between 1977 and 1984, used criterion-referenced tests. Students were asked to respond to test items based on key curriculum objectives, and an objective was said to have been mastered if a student answered two out of three items correctly. Average percentage mastery scores were reported for key mathematics content areas. When the series was resumed in 1999 in fourth class, a new norm-referenced test, based on the 1999 Primary School Mathematics Curriculum, was developed and was used for a second time in 2004. The report on the 2004 assessment included proficiency levels, allowing a criterion-referenced description of performance as well as a norm-referenced one (Shiel, Surgenor, Close, & Millar, 2006).

A new series of national assessments of reading and mathematics in 2nd and 6th classes was launched in 2009, and have been used in separate surveys for schools in general and for Irish language schools (Scoileanna Lán-Ghaeilge and schools located in Gaeltacht areas). The surveys use specially-developed standardised tests of reading and mathematics, based on the 1999 Primary School Curriculum, and are designed to monitor standards across sectors of the education system and among key at-risk groups.

The Use of Standardised Tests in Irish Schools

Standardised testing has been widespread in Irish primary schools for many years. Four out of five principal teachers reported that their schools had a policy of administering standardised English reading tests in 1993. This figure had increased to 97% in 1998 (Cosgrove et al., 2000). In 2004, teachers reported that 95% of students in first class and 96% in fifth class were assessed using standardised tests of English reading at least once a year (Eivers, Shiel, Perkins, & Cosgrove, 2005). Use of standardised tests of mathematics was less widespread but still extensive. In 1999, 55% of students in fourth class were taught by teachers who said that they administered standardised tests of mathematics at least once a year (Shiel & Kelly, 2001). By 2004, that figure had risen to 84% (Shiel et al., 2006). A number of factors may have contributed to this increase in use, including the use of tests to identify students who may be in need of learning support (DES, 2000) and encouragement by inspectors to provide test results in the context of Whole School Evaluation (WSE).

A number of recent policy initiatives require the use of data from standardised tests. For example, the establishment of school-level targets in literacy and numeracy as proposed in the blueprint for DEIS (*Delivering Equality of Opportunity in Schools*), the action plan to tackle educational disadvantage and bridge the gap in achievement between children in disadvantaged communities and their non-disadvantaged counterparts, will require test information (DES, 2005). Earlier policy initiatives, such as setting a national target of halving the proportion of students with serious literacy difficulties by 2006 (*National Action Plan against Poverty and Social Exclusion,* 2003), were also premised on the use of standardised tests of achievement, since performance on a test at or below the 10th percentile was specified as an indicator of low achievement (Eivers, Shiel, & Shortt, 2004). The 10th percentile has also been used as a cut-score for access to learning support (DES, 2000).

Precise data on the use of standardised tests in post-primary schools are not available. However, it is known that standardised testing has featured in the activities of post-primary schools for many years. Tests are used for a number of purposes, including (i) screening students before or shortly after entry; (ii) monitoring the eligibility, needs, and progress of students with special education needs; and (iii) gathering information to use in providing guidance and counselling. Use seems to be particularly prevalent at the point of student entry. Smyth, McCoy and Darmody (2004) reported that schools administered 26 different tests, either prior to school entry, or immediately afterwards, including standardised tests developed for use in primary schools, standardised ability tests, and tests developed by the schools themselves. There is also widespread use of individual and group standardised tests of achievement and ability by guidance counsellors (see, e.g., DES, 2009) and by resource/support teachers.

The use of standardised tests in post-primary schools differs in a number of respects from their use in primary schools. First, in the former, tests are often administered by specialist or guidance teachers rather than by subject teachers. It is not known if subject teachers draw on the outcomes of tests to inform their teaching. Secondly, those responsible for test administration in post-primary schools are likely to have specialist training in the administration and interpretation of tests beyond that available to primary school teachers. Thirdly, there is a paucity of Irish-normed tests in post-primary schools that could be used to assess and monitor the achievements of students, even in key learning areas such as reading and mathematics (Junior Certificate School Programme, 2006). Finally, the purposes for which standardised test information are used are likely to differ in primary and post-primary schools. In particular, the use of test information to allocate students to classes is much more likely to occur in post-primary schools.

Irish students also have had experience of standardised tests when they participated in international studies of student achievement at both primary and post-primary levels (Appendix A). A decision to participate in such a study is usually taken by, or in consultation with, a country's ministry of education, since the assessment must be funded, and access to schools may be required. Between 1989 and 1995, Ireland participated in several surveys, including the Third International Mathematics and Science Study (TIMSS) in 1995, which was administered in third and fourth classes in primary schools and first and second years in post-primary schools. Since 1995, Ireland has not participated in any international assessment at primary level. By contrast, at post-primary level, the country has participated in four cycles of the Programme for International Student Assessment (PISA), administered to 15-year olds drawn from second, third, transition, and fifth years. While earlier international assessments (including TIMSS) were curriculum-based, the focus in PISA is on 'real-life knowledge and skills' that are not based on school curricula.

The use of standardised tests in national or international assessments was unlikely to have had much impact on participating schools as tests were administered in only a sample of schools and results of student performance were not returned to schools. However, participation in international assessments, in addition to contributing to capacity building at national level regarding the development of tests, their administration, and analysis of finding, also raised issues about standards in the Junior Certificate Examination (see Chapter 6). The inclusion in PISA 2009 of an optional computer-delivered assessment of reading, and, in PISA 2012, a compulsory computer-delivered test of cross-curricular problem-solving skills, will serve as a basis for future development of computer-based standardised tests.

Conclusion

There has been remarkable growth in the use of standardised tests in recent years, much of it in the context of accountability. This growth has mostly been in the United States, though in Europe, there is also a tradition of their use in France and Sweden. Other European countries have also shown an increase in use, while world-wide, the implementation of national and international assessments of student achievement has resulted in widespread use.

In Ireland, the use of standardised testing is firmly established in primary schools. At post-primary level, with the exception of aptitude testing, it is largely confined to the point of entry to schools. Although tests in Irish, English, and Mathematics were developed and standardised for the first three years of post-primary school in the 1970s, little use has been made of them beyond the first year.

CHAPTER 4

ISSUES IN THE USE OF STANDARDISED TESTS

The use of standardised tests has for many years been a topic of controversy, particularly in the United States (see Kellaghan et al., 1982). A variety of advantages have been attributed to the practice. First, tests provide more objective and reliable information than the impressionistic measurement of student learning which is subject to a variety of biases. Secondly, tests can identify important curriculum objectives which teachers can use as instructional targets. Thirdly, tests provide teachers with information on how their students' achievements compare with those of students in other schools. Fourthly, tests can provide more detailed and systematic information on students' strengths and weaknesses, errors and misunderstandings, than a teacher is likely to be able to do for all students in his/her class. Fifthly, information based on test performance, when given to students and parents, is a potential source of motivation and accountability.

The use of standardised tests has also been strongly criticised. First, most tests do not provide information on what a student has learned, only how he/she stands relative to other students. Secondly, tests put pressure on teachers to teach to the test, leading to a narrowing of the curriculum. Thirdly, tests encourage a competitive atmosphere in the classroom. Fourthly, when standardised test results are used to select and classify students, they lead to labelling, which, in turn may be associated with the perpetuation of distinctions based on race, gender, or socioeconomic status. Even when not consciously used to classify students, it has been argued that test information can influence teachers' expectations.

Criticism of tests often failed to distinguish between different types of tests and different uses of tests. It also failed to appreciate that negative effects (e.g., labelling) may ensue from a variety of forms of evaluation. At a more fundamental level, little empirical evidence was available either to support or to challenge the value of standardised

tests. In an effort to provide such evidence, Rosenthal and Jacobson (1968) in a much cited study, *Pygmalion in the Classroom*, sought to determine if test information could influence teachers' perceptions of student ability which, in turn, could lead to changes in students' cognitive performance. The study, however, was judged to be severely deficient on design and statistical grounds (see, e.g., Gephart, 1970; Snow, 1969).

In this chapter, we review evidence relating to the effects of using standardised tests in two contexts: when low stakes are attached to test performance and when high stakes are attached. It should be noted that whether or not tests are standardised is not the crucial factor. In fact, most of the evidence on the consequences of testing relates to essay-type examinations, not standardised tests.

Effects of Standardised Tests in Low Stakes Contexts

Evidence relating to the use of standardised tests when low stakes are attached to performance comes from a four-year study carried out in the 1970s in a sample of Irish primary schools, some of which were randomly assigned to treatment (testing) groups and some to control (no testing) groups[1]. The study sought information on a wide range of effects of standardised tests on schools, teachers, students, and parents (Kellaghan et al., 1982). Here we report findings relating to seven frequently expressed statements about the effects of standardised testing (Kellaghan et al., 1980).

1. Testing limits teaching, putting pressure on teachers to teach to the test, thus leading to a narrowing of the curriculum.

Overall, fewer than 30% of teachers agreed that tests create pressures

1 There were four groups of teachers in this research: those who had tested and received norm-referenced results only; those who had tested and received norm-referenced results and diagnostic information; those who had tested but who had not received results; and those who had not tested at all.

to teach to the test. Furthermore, teachers who had experience with testing and with the use of test information tended to perceive tests as having less influence than teachers who had not had access to test results. With regard to classroom practices, quite a large number (about 40%) of teachers indicated that the achievement tests influenced, to at least some extent, the content they covered in class. A somewhat smaller number (about 30%) indicated that their teaching methods were affected, at least to some extent. Teachers' responses to these questions were practically identical after two and four years experience with testing. Thus, familiarity with testing gained throughout the study did not affect teachers' practices to any great extent

2. Testing leads to rigid grouping practices either at school or class level.

If test results were used to stratify students, we would expect that classes would become internally more homogeneous in terms of ability and/or achievement in those schools which had access to test results. On the other hand, classes would, under these conditions, become more heterogeneous with respect to each other. Analyses of test data revealed no effect which could be attributed to testing.

To examine the effects of test information at class level, teachers were asked the basis on which they grouped students within the classroom. Somewhat over 60% of teachers grouped their students for instructional purposes. While the availability of standardised tests and test information did not lead to more grouping of students according to ability and/or achievement, where teachers already operated such a procedure there was a tendency to use the results of standardised tests of intelligence in its operation.

3. Testing lowers student achievement.

Analysis of student test scores revealed that test experience had a

differential effect on performance on tests of ability and achievement. On ability tests, there was some evidence of an effect that was attributable to practice, but a further effect associated with the provision of test information was also identified. On most achievement tests, on the other hand, there was no evidence of positive effects of practice or of test information. An exception to this occurred in the case of the group which received diagnostic as well as norm-referenced information. There was a general and significant tendency for students in this group to score higher than students in all other groups. Thus, it would appear that the availability of diagnostic test information enhanced students' performance on achievement tests over the availability of norm-referenced test information alone.

4. Tests lead to labelling students. Further, teachers form expectations for students on the basis of test scores and students conform to these expectations.

An important corollary of this position is that if test scores underestimate a student's ability and/or achievement, as they are likely to do in the case of students of low socioeconomic status, then students may perform less well scholastically than they might have done if teachers did not have access to test scores (the 'self-fulfilling prophecy' or 'Pygmalion effect').

At the beginning of the school year in the investigation, at about the time that students sat for a battery of standardised ability and achievement tests, teachers rated each student in their class for general progress on the variables measured by the tests (e.g., mathematics computation, English reading). In the case of teachers who were to receive test results this was done before the receipt of results. The testing and rating procedures were repeated at the end of the school year.

When test information was made available to teachers, their subsequent ratings of their students' intelligence and scholastic achievement and students' actual test performance tended to move into line with the test information. If, on the other hand, test information was not available, students' subsequent test performance tended to move into line with the initial teacher perceptions of their intelligence and achievement. The inference from these findings is that test information disrupts teachers' perceptions, and that an expectancy process based on test information operates in classes. But an expectancy process also operates if teachers do not have access to test information. In that case, students tend to conform more in their scholastic performance to teachers' perceptions of them than do students whose teachers had access to test information.

For the majority of students (about two-thirds), there was no difference between teachers' beginning and end-of-year ratings. Students for whom ratings did change were more likely to be in the group in which teachers received test information than in the group whose teachers did not receive test information. Teachers who had received test information were more apt to raise their ratings of students than were their colleagues without such information.

While there was some evidence that the relationship between teacher perceptions and test performance was affected by group membership, the relationship was not consistent across subject area, grade, or socioeconomic group. However, the relationship was more likely to operate at higher grade levels, than lower ones, and to involve students from middle socioeconomic groups than students from higher or lower groups. Thus, there was little support for the claim that test information is likely to be most effective in the case of students from low socioeconomic backgrounds.

5. Testing increases fear, anxiety, and competitiveness among students.

Following four years of testing experience, sixth grade students were asked to respond to a series of questions dealing with their perceptions of, and reactions to, the standardised tests they had taken. While the majority of students expressed favourable attitudes towards the tests and reported no adverse emotional reaction to taking them, there was a significant minority who approached the testing situation with some trepidation. More students in the group whose teachers did not receive test results enjoyed sitting for the tests while more students in the test information group felt afraid in taking the tests. Thus, anticipation of test information being available to teachers affected students' feelings about taking the tests. Thirty percent of teachers said the tests increased competitiveness.

6. Testing may damage a student's self-concept.

Used properly, it is claimed that test information might enhance a student's self-concept (e.g., Bloom, 1969; Tyler, 1968). If, on the other hand, the information obtained from the test has strong negative overtones for the student, it may prove damaging to self-concept. Almost 60% of teachers thought that test results affected a student's self-concept. Over most analyses, however, it was not possible to demonstrate such a relationship.

7. Test scores have no direct positive usefulness in guiding instruction.

Teachers were asked their opinion on the usefulness of standardised tests in the classroom after having four years experience in using the tests. A majority (60%) thought that tests provided teachers with important information about students that was not generally obtainable from classroom observation. The specific classroom purposes for which teachers saw test results as being useful were the

grouping of students within a classroom for instruction (73%), the diagnosis of individual students' needs and abilities (68%), and the counselling of individual students regarding educational plans (50%). Thus, the majority of teachers did not agree that test scores had no direct positive usefulness in guiding instruction.

Effects of Standardised Tests in High Stakes Contexts

The research so far considered in this chapter was concerned with standardised tests in a context in which they were used for the first time, and the stakes for schools, teachers and students were low. We now shift the focus to high stakes standardised testing. Two countries stand out as users of standardised tests to hold states, regional educational authorities, schools, and teachers accountable for student achievement. In England (and in other parts of the UK until devolution), standardised tests have been administered at the end of Key Stages 1 (age 7), 2 (age 11) and 3 (age 14)[2], while students take an examination for the General Certificate of Secondary Education (GCSE), at age 16 (the end of lower-secondary education). In the United States, tests have been mandated in individual states for many years, with, in some cases, high stakes attached to performance for districts, schools, teachers, and students. The present situation is that state tests are administered at the end of grades 3 to 8 in the context of the requirements of No Child Left Behind (NCLB) (2002) legislation. Schools are required to demonstrate annual yearly progress (AYP), by steadily increasing the percentage of students who achieve at the 'proficient' level or higher so that, by 2014, all students will be reading at a proficient level.[3] An important requirement of NCLB is that, in addition to increasing overall achievement, schools are responsible for raising the achievement of students in various

2 From 2009, end of Key Stage 3 tests in England are optional.

3 It should be noted that states vary in terms of how they define 'proficiency' (i.e., at what point on an achievement scale the cut-point for proficiency is set). In the United States, the National Assessment of Educational Progress at fourth and eighth grades also uses proficiency levels, but these differ from those used in individual states.

subgroups, such as students from low-income families, ethnic and racial minorities, students learning English as a second or third language, and students who have a disability.

Some of the positive effects of standardised testing in high-stakes contexts that have been observed include

- constructive discussion of testing within schools through a collegial approach that can have a positive impact on students' self-efficacy (Harlen & Deakin Crick, 2002) and the emergence of greater co-operation in professional interactions (Demailly, 2001);

- some improvement in test scores, albeit often confined to the first few years after high stakes testing is introduced and soon reaching a plateau (Wyse & Torrance, 2009). Moreover, 'improvements' may not replicate themselves on other external measures of achievement (Mons, 2009);

- a stronger emphasis on higher-level thinking, but only if such thinking is emphasised in tests (e.g., state writing tests in the US) (Stecher, Barron, Chun & Ross, 2000);

- use of the results of high-stakes tests to plan instruction and to provide students with feedback (IGEN-IGAENR, 2005; Pedulla et al., 2003).

A series of negative effects of standardised tests in high stakes contexts have also been documented, including:

- a narrowing of the curriculum to closely resemble the content sampled by the test (Boyle & Bragg, 2006; Madaus, 1988; Madaus & Kellaghan, 1992), with less emphasis placed on non-tested subjects such as the fine arts, social studies and science (Pedulla et al., 2003; Smith et al., 1991);

- a progressive narrowing of the skills measured by tests over time, with tests in England requiring fewer higher-order reading skills such as inferencing and deduction (Hilton, 2001);

- teaching in ways that contradict teachers' ideas of sound instructional practice (Pedulla et al., 2003), with some adopting a teaching style emphasising transmission of knowledge at the expense of a more active and creative learning experience (Harlen & Deakin Crick, 2002)

- decreased teacher autonomy (Pollard et al., 1994);

- increased stress, anxiety and fatigue among teachers (Barksdale-Ladd & Thomas, 2000) and lower levels of teacher morale (Pedulla et al., 2003), with some teachers leaving the field (Hoffman, Assaf, & Paris, 2001);

- increased stress and anxiety among students (Webb & Vulliamy, 2006), lower levels of self-esteem among low achievers (Harlen & Deakin Crick, 2002), and more competitive classroom environments (Reay & Wiliam, 1999);

- increased dropout rates among lower achievers, placing minority students, students with disabilities, English as a second language learners, and low-SES students at greater risk (Haney, 2000);

- exclusion of lower-achieving and learning disabled students from testing (Haney, 2000);

- a stronger focus on summative and accountability purposes of testing, with less focus on developmental possibilities of providing feedback to teachers, parents, and students (Daugherty, 1995; Torrance, 1995, 2003);

- a tendency for teachers' own assessments to be more summative rather than formative (Harlen & Deakin Crick, 2002).

A common theme in the research on the effects of standardised tests is that the negative effects tend to be weaker in secondary schools than in primary schools. For example, in their national survey of teachers on the effects of state-mandated testing programmes on teaching and learning, Pedulla et al. (2003) found that high-school teachers (teaching grades 9 to 12) were less familiar with reports based on standardised tests, and less likely to agree that reports provided useful information, than were elementary or middle-school teachers. Furthermore, high school teachers reported fewer negative psychological effects of testing on students. High school teachers also felt less pressure from parents to bring about improvements. One reason for these differential effects may be that high school teachers, who are often content specialists and teach a small number of subjects, are already very familiar with the content standards in their subjects (on which tests are based), and have emphasised key content and processes in their teaching in the past. Given that many high school teachers do not teach the subjects usually targeted in high stakes assessments (i.e., mother tongue, mathematics, and sometimes science), they may not feel the responsibilities associated with high stakes testing to the same extent as teachers at primary level who work with the same students across a range of subjects, including those assessed using standardised tests.

Conclusion

The research reviewed in this chapter indicates that the stakes that are attached to test performance have a major role in determining the consequences that can be expected to ensue. When low stakes were attached to performance, as when tests were administered as a component of normal classroom procedures and the information they yielded was entirely under the control of the teacher, the information did not have a negative impact on what teachers taught or on how they organised their classrooms. Furthermore, the most

useful information came from tests that provided diagnostic information about student performance. In fact, students whose teachers were in receipt of diagnostic information, when later tested, achieved at a higher level than students whose teachers had not received such information. Analyses relating to the effects of test information on teachers' expectations for student performance indicated that teachers form expectations whether or not test information is available. However, teachers with test information were more likely to raise their expectations than teachers who did not have this information.

When high stakes are attached to test performance, the impact is likely to be much stronger. While there is some evidence of an associated improvement in test scores and a stronger emphasis on higher-order thinking if this is a feature of the test, a variety of negative consequences can also be anticipated: a narrowing of the curriculum, limited active and creative learning opportunities, differential treatment of students leading to increased dropout, increased stress on teachers and students, and lower levels of self-esteem among low achievers.

Assessmant Update: Australia

Australia has recently introduced a National Assessment Programme: Literacy and Numeracy (NAPLAN) for students in four grade levels including Years 5 (12-13 year olds). All students at the target year levels are assessed on reading, writing, language conventions (grammar and punctuation, spelling), and numeracy. NAPLAN results are reported nationally through the Summary and National Reports, and at the student level. Results are available for use by education systems, schools and parents. See http://www.naplan.edu.au/

CHAPTER 5

INTERNATIONAL PRACTICE: THE FINDINGS OF THE CROSS-COUNTRY STUDY

In this chapter we present the results of our enquiry into the use of standardised tests in selected countries. The review is based on responses to a questionnaire administered for this study and other evidence where available. The content of the questionnaire was determined by the terms of reference of the study and by feedback provided by the NCCA. Two considerations merit attention in reading this review. In the first place, education systems differ in their structure, with the result that what is regarded as secondary education varies from country to country. For example, primary education lasts five years in France while in Denmark what might be regarded as primary and lower secondary education are combined in the nine grades of the *Folkeskole*. A judgment had to be made in some cases about what to regard as constituting lower secondary education (e.g., the higher grades of the Danish *Folkeskole*, although decisions about students in this situation are likely to be different from those for students of the same age who have transferred to a different sector of the education system).

A second consideration to be borne in mind in reading our review is that traditions of assessment vary from country to country. In some (e.g., Denmark, Germany, Norway), teachers' assessments have long been privileged, even to the extent that they played a major role in the certification of students at the end of secondary school. There was little interest in more 'objective', but what might be considered narrower, forms of assessment, as we saw in the section on the history of testing was the case in Germany (Chapter 3). In Denmark, until recently, a student could pass through the education system up to the last month of grade 9 without ever having taken a test or formal examination. A consequence of differences in experience with standardised testing is that we cannot be sure that all respondents to our questionnaire had in mind the characteristics of such tests as described in Chapter 2.

A related issue to be borne in mind is that external assessment instruments were not confined to standardised tests. For example, in New Zealand, a wide spectrum of resources suitable for classroom assessment, including exemplars and item banks, are available. In France, in addition to the formal tests used in national assessments, a bank of assessment tools that may be used on a voluntary basis in the lower secondary school is available. In interpreting the responses provided in questionnaires, it was not always possible to distinguish between standardised tests and other resources.

The questionnaire (Appendix B) included the following sections:

- A general section, which asked about the grade levels (Grades 7, 8, 9) in lower-secondary schools at which standardised tests were administered; the particular abilities and curriculum areas assessed by standardised tests; whether tests used at more than one grade level were linked; the grade levels at which the administration of standardised tests was compulsory for schools; the use of standardised tests to certify students' achievements; the time of year at which tests are administered and who decides this; whether schools have a choice in the tests they use; who determines the purposes of the tests; the main interpretation attached to standardised tests at each grade level; and whether tests used at lower secondary level were linked to tests used at primary level.

- A section on test administration, which asked how tests are delivered (paper and pencil and/or computer-based); who administers and scores the tests; if administration is monitored by an external agency; categories of students excluded from testing; accommodations made for students with home languages different from the language of instruction; and who supports standardised testing financially.

- A section on use, interpretation, and dissemination, which asked how test results are used; how results are reported, and to whom; restrictions (if any) placed on the use of test results; the types of support provided to teachers in interpreting the outcomes of standardised tests; how parents are supported in interpreting test results; and how results are presented to the public.

- A section in which respondents could identify other sources where information on their assessment systems might be obtained (e.g., journal articles, websites).

The questionnaire was sent to representatives on the PISA Governing Board for the following countries: Denmark, Finland, France, the Netherlands, Norway, and New Zealand. In some cases, the representatives who received the questionnaires (mainly officials in State Departments of Education) completed them themselves; in other cases, they forwarded them to colleagues in the same department with the relevant knowledge or to outside organisations or individuals. Information was obtained from published sources on use of standardised tests in Northern Ireland, Scotland, and one Canadian province, Ontario.

Some respondents to our questionnnaire expressed difficulty with the term 'standardised test', and, even when this was clarified for them, still had difficulty answering some parts of the questionnaire. In one case, the respondent indicated that the questionnaire was not relevant to the situation in his/her country, and provided instead a description of the assessment system that was in place.

In considering the responses provided by respondents and information gleaned from the literature, it should be noted that, in most countries, assessment systems are constantly changing, and that therefore, the information in this chapter may soon be dated. This also presented problems when attempts were made to cross-check

responses to questionnaire items with other sources of information, such as the International Review of Curriculum and Assessment Frameworks Internet Archive (INCA, www.inca.org.uk) and a recent EU study on national testing in Europe (Eurydice EACEA, 2009).

At different points in the chapter, we present short vignettes – descriptions of approaches to testing and/or reporting outcomes in some of the education systems we examined. These are intended to complement the more general descriptions in the text.

Purposes for which Standardised Testing Is Carried Out

Three main purposes can be identified in the use of standardised tests which coincide with the contexts for the uses described in Chapter 2: to support teachers' assessments of their students (classroom assessment); to provide information on standards of achievement in the education system (national assessments); and to provide information on student achievements in the education system relative to the achievements of students in other education systems (international assessments). All three purposes are in evidence in the countries in which we examined the use of standardised tests. (Information for individual countries regarding national and international assessments and student certification examinations is contained in Appendices E, F and G).

It should be noted that national and international assessments are to be distinguished from public/certification examinations, which also are a feature of many European systems of education, although the Eurydice EACEA (2009) report on 'national testing' of students does not maintain this distinction. Such examinations are held at the end of lower secondary education in Denmark (final year of the *Folkeskole*), France, New Zealand, Norway, and Scotland. In Northern Ireland, students complete the General Certificate of Secondary

Education (GSCE) at the end of Key Stage 4 (age 16). Examinations are not held until a later point in upper secondary education in Finland, the Netherlands, and British Columbia and Ontario in Canada. In Ontario, the compulsory state-wide literacy tests at grade 10, while designed to provide an objective measure of students' literacy levels, also serves as a surrogate public examination, since passing the test is a graduation requirement (see Country Vignette 1).

In the Netherlands and New Zealand, the main purpose of testing is to support teaching and learning in the classroom (see Country Vignettes 2 and 3). While this purpose is also articulated in other countries, other purposes also are pursued (e.g., monitoring the performance of schools) which may not be entirely compatible with the support of classroom learning (see Country Vignette 4).

Information from sample-based national assessments is broadly used to inform policy about teaching and learning and to devise policy to promote equity in the system (see Country Vignette 5). When the assessment is census-based, there are additional opportunities for impacting more directly on teacher behaviour.

In countries with census-based assessments, such as Denmark and Norway, the function of providing information for guidance at the classroom level is combined with the function of providing information at national (or sub-population) level in a single system of assessment. Feedback information on the performance of schools is provided to teachers, while data are also aggregated to describe performance at municipal, county, and national levels. Assessment systems that have dual functions are census-based.

This is also the case in France where the system is elaborate and unique. A census-based 'diagnostic' assessment of French and mathematics, designed to provide guidance for teachers, is administered in all classrooms in the first year of secondary education,

and the results from a random sample of participating schools are used to compile a national report providing information on the system as a whole (see Country Vignette 6). In the fourth year of lower-secondary schooling, there is a rolling programme of sample-based national assessments (see Country Vignette 7).

By contrast, in Scotland, the external system of assessment to support classroom teaching and learning is separate from the system to monitor the education system. Assessment materials (including standardised tests) are made available to schools, but their use is not mandatory. Commercially prepared tests in mathematics, reading (diagnostic) and spelling, covering mainly grades 2 to 9, are available for teacher use in Denmark. Information for national monitoring is obtained using specially designed tests in sample-based surveys.

International assessments such as PISA and TIMSS, which use standardised tests, are quite separate from external assessments designed to support classroom teaching or to provide information on the performance of the education system.

In all countries, standardised tests were perceived to provide information that could be used in a formative way by teachers. Not only that, test information was clearly intended to have a role that is subsidiary to teachers' judgments in Denmark, New Zealand, and Scotland. Even when an assessment provided summative information on national performance, as in Norway, the assessment results were perceived as having a formative role in the classroom.

Country Vignette 1: Ontario's Census-based National Assessment

In addition to the Ontario Secondary School Literacy Test at Grade 10, the Education Quality and Accountability Office (EQAO) implements census-based national assessments involving standardised tests at the end of grades three and six (reading, writing and mathematics) and grade nine (mathematics only). Each year, separate provincial reports are published for both English and French-speaking students taking the tests. The annual provincial reports cover performance across the three grade levels. Performance is also reported by school-board area, and by school. In cases where the number of students in a school or school board is less than 15, results are only available to school staff and the school board. Although the EQAO has stated it is opposed to ranking schools, the Ontario Ministry of Education mandates that school level data be publicly released, leading to the ranking of schools in local newspapers. In 2004, the EQAO introduced the Education Quality Indicators Framework (EQIF) to provide information on a range of factors influencing achievement, such as linguistic background and socioeconomic status, to encourage a more contextual interpretation of results.

Students and their parents receive an Individual Student Report (ISR) for each assessment. EQIF data is publicly available in the provincial report, but failure to send the information directly to parents may negate any benefits in terms of public interpretation of results, as a 2005 study (Mu & Childs) revealed only 13.5% of parents visited the EQAO website. Results of the assessments of reading, writing and mathematics are reported with respect to four achievement (proficiency) levels.

School, board and provincial reports contain: overall results for each subject at school, board and province levels; longitudinal data at each level so that changes in the performance of cohorts can be tracked over time; overall jurisdictional results for each subject by gender and other characteristics (such as ESL/ELD learners and students with special needs); areas of strength within the curriculum and areas for improvement; and contextual data. Individual student achievement results for all students in the school and board are also contained in school and board reports. In addition to all this, schools receive summary item statistics (e.g., percent correct scores) for the school, school board and province, and item results for each student, which may be useful for diagnostic purposes.

Results are also reported to schools as individual profiles that explain students' assessment results in relation to provincial standards. The profiles provide a strategy for teachers to use exemplars to talk to parents and students about how the assessment information fits with the provincial curriculum expectations and with other information about the student.

Schools and districts must compile reports consisting of interpretation of assessment results and action plans for improvement, based on the information provided by the EQAO. Thus assessment results are intended to feed back into the teaching and learning process in the classroom as well as informing system planning.

Source: www.eqao.com

Country Vignette 2: Optional Standardised Testing for Diagnostic Purposes in the First Two Years of Lower-Secondary Schooling in The Netherlands

In the Netherlands, control of testing is largely exercised at school level, reflecting the high degree of autonomy granted to Dutch schools generally. Schools may opt to make use of a monitoring and evaluation system for students (LVOS) that covers the first two years of secondary education, which is provided by CITO, an independent testing agency. LVOS at lower secondary level consists of an entrance test, a test after the first year, and a test after the second year of lower-secondary schooling. The tests are not compulsory, and testing is financed solely by schools. Decisions regarding testing (e.g., when tests are to be administered, which skills are to be tested, and in which order the tests are to proceed) are left to school principals and class teachers. Individual teachers also determine what purposes the tests will serve. In practice, results are almost exclusively used formatively, that is to adapt education to suit the needs of the individual student. Ultimately, tests results can contribute to the decision that a student should go to a different type of school, but the test results play only a minor role in this, being considered alongside other forms of school-based assessment

LVOS tests are available in reading comprehension in the language of instruction (Dutch), reading comprehension in a foreign language (English), and Mathematics. There is also an aptitude test in study skills, which can be administered at any time. All of the tests are available at three different levels of difficulty, each of which serves two of the six levels of secondary education in the Netherlands.[1] Schools decide whether to include students with SEN. Test norms are available for both the beginning and end of the school year, and administration is carried out by class teachers. Tests are linked to allow tracking of the progress of individual students over time.

CITO provides an electronic scoring service, and reports results to schools at student, class and school levels. Tests at different levels of difficulty are reported on the same scale. Schools may report individual student results to parents. Schools may also report results at class level to the school board, the local community and external bodies.

CITO provide training courses and written materials to assist teachers in interpreting and using the results for diagnostic purposes. The report received by parents provides some information to help in interpreting the scores, and additional information for parents is available on the internet. Ultimately it is the responsibility of the school to ensure that parents can interpret results.

Lower secondary schools also have access to their students' results on the 'CITO tests' – optional standardised tests in language, mathematics, study skills and (where selected) environmental studies, taken by almost all Sixth grade students in the February prior to entry to lower-secondary schooling. These tests are intended to provide independent information to assist schools in arriving at decisions about intake. Test results are available to parents as well as to secondary schools.

Sources: Eurydice EACEA, 2009; http://www.cito.com Questionnaire responses.

1 Generally, all students within a school sit tests at the same level of difficulty. In the first grade of secondary education, students of different levels may be combined in the one grade. In such a case, schools may administer different tests to students within a grade, although this is rare in practice.

Country Vignette 3: Tools for Classroom Assessment in New Zealand

In New Zealand, the Ministry of Education provides a number of tools for classroom-based assessment – Assessment Resource Banks (*ARBs*), Assessment Tools for Teaching and Learning (*asTTle*) and National Exemplars – all of which are free to schools, and all are available in English and Maori. The tools are intended to provide externally referenced assessment information to assist teachers in making valid, reliable and nationally consistent judgments about the work and progress of their students. The tools have not all been standardised in a formal sense, nor are steps taken to ensure that administration and scoring is consistent across schools. Nevertheless, the tests enable teachers to diagnose how their students are performing, give feedback to them about progress, and jointly establish goals for learning. At school level, information may be aggregated and used to evaluate teaching programmes and inform strategic planning.

The Assessment Resource Banks (*ARBs*) are an online collection of 2868 curriculum-based assessment resources in English, mathematics and science, designed for students working at levels 2-5 (up to age 15) of the New Zealand national curriculum (see http://arb.nzcer.org.nz/sample.php for examples). Assessment tasks and items may be combined to form tests for class or school-wide use, or customised sets for formative and diagnostic assessment. Each resource includes an assessment task, a scoring guide, and information on how the resource relates to the national curriculum.

The Assessment Tools for Teaching and Learning (*asTTle*) are for assessing reading, writing and mathematics in years 4-12 (8 to 16 years of age). Students can take tests in paper-and-pen format or online. Graphic reports allow teachers to analyze the achievements of individual students and groups against curriculum levels, curriculum objectives, and population norms. Future learning needs are also specified. Workshops, online tutorials and videos are provided to inform teachers on technical and interpretative aspects of *asTTle*.

National Exemplars covering Levels 1-5 of the New Zealand curriculum provide teachers and students with annotated examples of work that show progression in selected areas of each subject, allowing them to make decisions about the quality of individual learning, achievement and progress. Features of learning that teachers need to watch for, collect information about, and act on to support progress in learning are highlighted. There are 75 exemplars for English writing, covering poetic writing-character, poetic writing-personal experience, transactional writing-character, and transactional writing-personal experience. Exemplars relating to visual language and oral language, mathematics, health and physical education, science, social studies, technology, and the arts (dance, drama, music and visual arts), are also provided.

With such a broad range of tools available for classroom assessment, teachers can also access a selector that allows them to draw comparisons across different tools, and select the one most appropriate to their needs. The selector covers English, social studies, the arts, cross-curricular, mathematics, health and PE, information skills, science, technology, and student engagement in learning. No tools are provided for two aspects of the New Zealand curriculum: key competences and values.

Sources: www.inca.co.uk , http://assessment.tki.org.nz/

Country Vignette 4: Compulsory Standardised Test of Basic Literacy Skills in Ontario (Canada) Secondary Schools

The Ontario Secondary School Literacy Test (OSSLT) is a compulsory, state-wide literacy test, which is administered in grade 10 (age 15-16). The test has been administered annually on a census basis since 2000/2001. Its function is 'to determine whether a student has the literacy (reading and writing) skills required to meet the standard for understanding reading selections and communicating in a variety of writing forms expected by the Ontario Curriculum across all subjects up to the end of Grade 9' (the end of lower secondary schooling). The assessment is intended to provide an objective measure of the literacy levels of graduates of Ontario's high schools for the assurance of students, parents, post-secondary institutions and employers.

Although results do not count towards students' grades in any subject, passing the test is a graduation requirement since the 2001/2002 school year. Testing takes place two years before graduation in order that students who fail may receive additional help and re-sit the test. Students who repeatedly fail the test may take the Ontario Secondary School Literacy Course in its place, as completion of this course also fulfils the graduation requirement. According to Volante (2006), the OSSLT is responsible for an increase in early school leaving in Ontario, as lower-achieving students who do poorly become discouraged.

As is common throughout Canada, teachers are involved in the development, administration and scoring of the test. Unlike the other provinces, in Ontario this process is supervised by an independent agency, the Education Quality and Accountability Office (EQAO).

Class teachers administer the test in March/April. EQAO quality monitors are sent to a random selection of schools. Tests contain both open-ended and multiple-choice style items. Multiple-choice items are machine-scored, and written responses are scored in a central location by teachers from across the province. One version of each assessment is developed for English-language students and another for French-language students. If the Individual Education Plan (IEP) of a student with special education needs states that they are not working towards an Ontario Secondary School Diploma, they may be excluded from testing. Students with IEPs are allowed the accommodations that they would normally receive.

Results are reported at two levels only: pass and fail. The EQAO publishes an annual report on results at provincial level on its website. Results at school and school board level are also publicly available through the site, except in cases where the number of students is fewer than 15. Schools and school boards receive data files with individual student achievement results for all students in the school and the board. Schools also receive individual item results for each unsuccessful student. School boards receive additional data files with detailed results for each school, the board and the province. Parents and students receive an Individual Student Report. Student scale scores and feedback are provided to unsuccessful students.

Source: Educational Quality and Accountability Office. (2009).

Country Vignette 5: Monitoring of Educational Outcomes through Periodic Sample-based National Assessments in Finland

Unlike Ireland, students completing compulsory (basic) education in Finland (Grade 9; equivalent to Third year in Ireland) do not take an external examination for certification purposes. Rather, individual schools are responsible for certifying satisfactory completion of basic education. In assessing students for certification purposes, teachers may compose their own examinations, use tests that accompany text books, or draw on exam papers provided by subject teachers' associations.

National assessments have been implemented by the Finnish National Board of Education (FNBE), an agency of the Ministry of Education, at the end of Grades 6 and 9 since the early 1990s. Their purpose is to ascertain how well the objectives set in the national curricula have been achieved, and to monitor equality of outcomes by gender, region, social group, and language group. A different subject or cluster of subjects is assessed each year, with mother tongue (Finnish or Swedish) and mathematics being assessed most often. Other subjects are assessed according to national priorities. Physical education was assessed in 2003, with students being graded on their ability to perform standardised physical tasks by their teachers.

The national assessments are administered to representative samples of schools and students. Schools not selected to participate may purchase the tests. Students, subject teachers and principal teachers complete questionnaires. A feature of the student questionnaire is the inclusion of questions on attitudes and learning styles, and the treatment of these as outcomes alongside achievement.

In Grade 9 mathematics, three aspects are assessed: basic mathematics (multiple-choice only), mental calculation (with items presented orally or in writing), and problem-solving (open-ended only). Thirty minutes is allocated to basic mathematics, and one hour to problem solving.

The outcomes of the national assessments are used for a variety of purposes. Schools (those sampled, and those that purchase tests) use them for their own development purposes; at national level, learning outcomes are used in making decisions about

- support measures to promote equity across social and other groups
- standards for student assessment (grades assigned to students by their teachers are compared to their performance on the national assessment tests)
- teaching and learning (e.g., allocation of time to various subjects).

Achievement of learning outcomes is monitored over time, through the inclusion of 'anchor' items on the tests.

Although the FNBE does not issue results to regional education authorities, some authorities compile results for schools in their region.

There has been intense media pressure to publish school rankings, based on performance on the national assessments, but the national consensus in the ensuing debate was against publishing test results for schools.

Source: Finnish Board of Education http://www.oph.fi/english; Eurydice EACEA, (2009); Questionnaire responses.

Country Vignette 6: Compulsory Census-based Diagnostic Testing at the Beginning of Lower Secondary Schooling in France

Compulsory mass diagnostic national testing of students in the first year of upper primary schooling (age 8) and the first year of lower secondary schooling (age 11) was introduced in both public- and private-sector schools in France in 1989, though, since 2007, such testing is no longer compulsory at primary level. Standardised tests in French and mathematics, designed by teams of teachers and researchers, are provided to schools by the Directorate of Evaluation, Planning and Performance (DEPP) of the Ministry of Education. The tests are administered by students' form teachers during the first two weeks of September, and are also scored by them, with administration and scoring each taking two hours. The tests are adapted for students with special education needs (e.g., braille format for the visually impaired).

The primary goals of mass diagnostic testing are:

- to provide teachers with a tool to gauge their students' progress, strengths and weaknesses.
- to assist teachers in choosing the teaching activities most suited to the students' needs.
- to assist teachers in planning their teaching of the curriculum accordingly.

After testing has taken place, teachers can investigate further to establish the thought processes used by students to reach certain answers. To help them with this, the DEPP provides tables specifying objectives and competencies, and a coding system for categorising students' errors or incorrect answers. Computer software is provided for calculating student scores and summarising error patterns. Students' performance can be categorised as below basic, basic, good, or above average, with basic regarded as a minimum for success in lower secondary schooling. Results are discussed with students' parents with a view to determining which students need to make use of the additional/optional two to three hours per week allowed in the school timetable for the consolidation of areas of weakness.

Indicators of student achievement in French and mathematics are published annually by the DEPP on the basis of data collected from a representative national sample of schools. However, since test content changes from year to year, no trend data are provided. Results for individual schools or regions are not published.

To complement or enhance the diagnosis conducted during mass diagnostic testing, teachers can draw on a bank of assessment tools in French and mathematics that is made available on the Internet (www.educ-eval.education.fr).

Teachers have found that the results of mass diagnostic testing at ages 8 and 11 serve as a starting point for discussions with parents, as the nature and timing of these national assessments convince parents that the results are objective and that their child's individual needs are being taken into consideration. In this way, parents are aware of the need for any remedial action that may be necessary and can be encouraged to involve themselves with their child's learning.

Teachers of lower-secondary students also have access to results of a census-based national assessment completed towards the end of primary schooling. These are intended for information purposes. They may also be used nationally or locally to plan in-career development activities for teachers.

Sources: www.educ-eval.education.fr and http://www.inca.org.uk/france-assessment-mainstream.html

Point at Which Testing Takes Place

External standardised tests that schools are required to administer and that provide information to teachers to support their own assessments are administered at the beginning of the school year in the first grade of secondary education in Norway. Tests (diagnostic) are also administered in the first year in France. Administration early in the careers of students in secondary schools means that information is available to guide instructional practice from the beginning.

Tests are available for the first two grades of secondary education in Denmark (grades 7 and 8 in the *Folkeskole*), where tests are computer-based, and testing is required by law, but teachers decide on the most appropriate time. In New Zealand, testing is not compulsory and teachers decide whether or not to use tests, and when to use them, while in Scotland, again testing is not compulsory, but most students participate.

A sample-based national assessment which cannot provide diagnostic information to individual schools is administered in lower secondary schools in Finland and Scotland. In Scotland, it is carried out in the second year of secondary education. As the national assessment in Finland is primarily designed to provide information for policy purposes on the achievements of students in the education system, it is administered in the final year of lower secondary education towards the end of the school year (in March-April). The national assessment designed to provide information on achievement in the education system in France is sample-based and is carried out at the end of the secondary school cycle.

PISA tests are administered every three years to 15-year olds in all the countries included in our survey. They are administered between March and May in the northern hemisphere, except in the UK and US where they are administered in November, and between October

and November in the southern hemisphere. This is the only assessment in which age, not grade, is the criterion for participation. All countries in our survey, with the exception of Finland, France and Northern Ireland, also participated in TIMSS 2008 (grade 8), which is administered in April to June in northern hemisphere countries, and in November-December in southern hemisphere countries, every four years.

Achievements Assessed

The provision of tests to assess achievement in students' mother tongue or language of instruction is a feature of all the education systems investigated in this review. In some countries, this involves only one language. In others (Finland, Scotland), more than one language is involved.

The main focus in language tests is on reading. However, New Zealand has a listening test. Scotland has, in addition to reading, assessment in listening, talking, and writing, though how to assess listening and talking is left to teachers. Norway included a writing test in an earlier assessment but this has been dropped.

Mathematics or numeracy features in all assessment systems at lower secondary level with the exception of Denmark where its (computer-based) assessment system is in the course of development. At present, mathematics tests are available at grade 6, but will be available at grade 7 in the future (Wandall, 2009).

Other curriculum areas are included in the formal assessment systems of some countries. In Denmark, the Netherlands, and Norway, English reading is assessed. In Denmark, there is also provision for the assessment of Danish as a foreign language. In France, all subjects taught in lower secondary school are assessed in a six-year cycle in a national assessment (see Country Vignette 7). Denmark also has

assessments in a range of curriculum areas (Physics/Chemistry, Biology, Geography). Finland has assessments in foreign language, science, and technology.

In a number of countries, an attempt is made to assess areas of achievement that are difficult to measure using standardised tests. In Northern Ireland, there is a strong focus on assessing cross-curricular competencies (communication, using mathematics, using ICTs) (see Country Vignette 8). In Finland, the areas are cross-curricular abilities, problem-solving ability, learning strategies/skills, and the ability to work in groups. In Scotland, the areas are communication, using ICT, problem-solving ability, and working with others. Not all of these would be amenable to assessment in a test that meets all the criteria associated with standardisation.

The absence of assessments in science is noteworthy. That lacuna may, however, be addressed in PISA which assesses the achievements of 15-year old students in all countries in reading literacy, numerical literacy, and scientific literacy in a three-year cycle. Science is also assessed in TIMSS, in which Denmark, the Netherlands, New Zealand, Norway, and Scotland participate.

Country Vignette 7: Rotating Programme of Sample-based National Assessments at the End of Lower Secondary Schooling in France

There are three assessment strands at the end of lower-secondary schooling in France: an examination (the diplôme national du brevet, of French, mathematics, history/geography and civics, foreign language and ICT skills, and, from 2011, art history) taken by most students, which certifies successful completion of lower secondary schooling (the collège); occasional national assessments of French and mathematics involving representative samples of schools, classes and students; and a rotating national assessment programme covering these and other subjects, and also involving representative samples. The national assessments use standardised tests. In the case of the brevet, the outcomes of school-based assessments are combined with examination results, and scoring and interpretation of outcomes vary by region. Here, we consider the rotating national programme. The table shows the domains (competences) assessed each year since this programme was initiated in 2003.

Subject Domains Assessed Since 2003 in Rotating National Assessment Programme at End of Lower-Secondary Schooling in France	
Year	Domain
2003	Written and oral comprehension (French)
2004	Foreign Language (English, German, Spanish)
2005	Attitudes toward life and society
2006	History, Geography and Civic Education
2007	Science (Life and earth sciences, Physics, Chemistry)
2008	Mathematics
2009	Written and oral comprehension (French)

The purposes of the rotating programme are to monitor the education system at national level, and to compile an objective report on the competencies and knowledge of students in key subjects. The monitoring function is fulfilled by assessing the same domains every six years. The outcomes of the assessments are used to regulate educational policy at national level, to modify curricular content, to inform the definition of competencies, to review the structure of academic courses and pedagogical organisations, and to address the needs of certain school populations (e.g., low-SES students).

During testing, students answer different clusters of questions, ensuring broad coverage of the assessment domain. Testing takes two hours. Participation of students with special educational needs is optional, and school principals decide whether or not such students can take the test under the same conditions as other students. The tests are supplemented with background information gathered from principal teachers, class teachers and students. Tests are scored centrally by the DEFF and scaled using item response theory methodologies (IRT).

Outcomes of the national tests are reported in terms of proficiency levels and mean scale scores (aggregated, by gender, and by school type), and national reports are compiled and published (see http://educ-eval.education.fr/bilan2.htm). Performance is not aggregated or reported by region or school.

A similar rotating programme of national assessments operates at the end of primary schooling, allowing for some comparisons in attitudes and knowledge between students at the end of primary and lower-secondary levels.

Source: Eurydice EACEA (2009); http://www.education.gouv.fr/

Country Vignette 8: Standardised Testing in Northern Ireland

Up until recently, students in Northern Ireland reaching the end of Key Stage 3 (Age 14, Year 10) were required to sit standardised tests in English/Irish, mathematics, science and technology. Following the phased introduction of a revised curriculum beginning in 2006, standardised testing at the end of KS3 is no longer mandatory. Instead, schools are required to conduct and report on the outcomes of teacher assessments that are linked to curriculum levels. Scores of students at the end of Years 4, 7 and 10 on language and literacy (English or Irish as appropriate) and on mathematics and numeracy must be reported to the Council for Curriculum, Examinations and Assessment. Schools are also required to enter individual student outcomes in all subjects on a Student Profile which is sent to a student's parents. Over time, it is expected that teacher assessments will be supported by more formal, computer-based tests.

Another significant change has been the discontinuation of the 11+, a centrally-administered standardised test used to determine the post-primary schools to which students would transfer. For 2010 entry, post-primary schools are advised not to use academic criteria, such as results on a standardised test, but are not precluded by the Department of Education from doing so.

Taken together, these changes represent a strong shift away from standardised testing towards teacher-based assessment, which is sometimes moderated.

An important development in curriculum in Northern Ireland is a renewed focus on key skills or cross-curricular competencies. For Key Stage 3, these are communication, using mathematics, and using ICTs. Hence, in addition to assessing traditional subject domains, teachers will be required to assess students on these key skills using a seven-level framework containing descriptive criteria. Criteria for assessing progress in "thinking skills and personal capabilities" are also under development.

Students in Northern Ireland continue to take the GCSE (General Certificate of Secondary Education) examination at age 16 (Year 12), marking the end of Key Stage 4 and compulsory schooling.

Sources: http://www.nicurriculum.org.uk;
http://www.deni.gov.uk;
http://www.rewardinglearning.org.uk/

Language of Tests

In bilingual countries, assessment instruments are provided in two languages: Finnish and Swedish in Finland, English and Gaelic in Scotland, and English and Irish in Northern Ireland. Regulations vary in officially monolingual countries. In France and the Netherlands, no accommodation is permitted for students whose home language is not the official language. In other countries, there is provision for assisting students whose home language differs from the language of instruction. In Denmark, teachers can decide how much support students may need, and provide that support. In New Zealand, students can also be given assistance.

In Finland, students whose home language is neither Finnish nor Swedish, and who are considered not to be able to take the tests in one of these languages, are exempt from testing. In Scotland, students whose first language is neither English nor Gaelic should only attempt reading and writing tasks when the results of continuous assessment indicate they will attain targets independently of language support. Language support may be provided in mathematics, but when it is, this should be recorded and reported.

Format of Tests

Tests are presented in both paper-and-pencil and electronic forms. In Finland and New Zealand, the paper-and-pencil format is used. However, there are also internet-accessible resource banks available for use at primary level in New Zealand.

In Denmark, a system of computer-based adaptive testing is being developed which will automatically generate reports for parents and teachers (see Country Vignette 9).

Country Vignette 9: Introduction of Census-based Computerised Adaptive National Testing in Denmark

In Spring 2010, Denmark will introduce compulsory computer-adaptive national testing for students in public primary and lower secondary schools (the Folkeskolen). The introduction of the national tests is designed to establish a stronger assessment culture in Danish schools, and hence improve standards. The table below shows the subjects to be assessed at each grade level. Each subject is further divided into three dimensions, with separate results to be generated for each dimension (for example, the dimensions of Danish/reading are understanding language, decoding and text comprehension) as well as for overall performance.

Grade Levels at Which Computer-Adaptive National Tests Are To Be Administered in Danish Schools in 2010, and Corresponding Subjects

	Grade Level							
Subject	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Danish/reading								Folkeskolens afgangsprove (Leaving Examination)
Mathematics					**			
English								
Geography								
Biology								
Physics/Chemistry							**	
Danish as 2nd language								

*Grades 7-9 (13-15 years of age) can be viewed as being equivalent to Lower-secondary Schooling
**Compulsory testing of mathematics in Grade 6, and of Physics/Chemistry in Grade 8 was also conducted in 2007.

All tests will be offered on computer over the internet, free of charge to schools. The tests are adaptive in that the items administered to an individual student are selected with reference to the student's ability (e.g., after the first few items, a student with 'high' ability would not be expected to respond to easier items, thus allowing for a more accurate estimation of his/her achievement). The test administration window is February 1st to April 30th, and schools will be required to book testing time, as only 60,000 students nationally can be tested at any given time. Although the time allowed for each test is 45 minutes (during which students are asked to respond to 50-80 questions drawn from a pool of 500), teachers may extend the testing time for an individual student. Similarly, teachers will decide which tools students are to use during testing, and which accommodations to make for students with disabilities. Scoring will be done centrally, by computer, with reports issued for individual students.

The new tests are intended to be 'low stakes'. Schools and municipalities will be allowed to access results on different levels, while class results will be available to class teachers, and parents will be provided with reports by the school on their child's performance. Finally, national results will be used to generate a national profile of performance, with attention to differences in performance from year to year. Five proficiency levels are identified for each subject and each dimension within a subject: Level 5 (top 10%), Level 4 (next 25%), Level 3 (middle 30%), Level 2 (next 25%) and Level 1 (bottom 10%). It is planned to publish national results on an annual basis.

Clearly, the planned testing programme for Denmark is innovative and is worth examining further as it evolves. It is computer-based, and hence can be expected to minimise the time required for scoring and generating reports. But some drawbacks are apparent. Only multiple-choice items are used to assess student performance (see http://evaluering.uvm.dk for examples), and this may restrict the range of processes that are assessed. There have been technical problems in administration of the test, leading to the introduction of compulsory testing being postponed in 2009. Finally, even though there is flexibility with the arrangements for testing (teachers decide which accommodations to provide), it is nevertheless planned to use results to track progress of the system, and of individuals, over time.

Other education systems combine paper-and-pencil and electronic means of presentation of tests. In the Netherlands, tests are available in both forms. In Norway, the paper-and-pencil form is used for Norwegian reading, while tests of mathematics and English reading are computer-based. In Scotland, teachers access tests on a website, but tests are distributed to students in print form.

Administration and Scoring

Tests are administered by classroom teachers in most jurisdictions. Teachers also score tests, though in some countries an electronic scanning service is available (Finland, Netherlands, New Zealand). Tests used for the sample-based national assessments in France and Scotland are scored externally.

A variety of supports are in place to support teachers in administering tests, in scoring and interpreting test performance, and in communicating results to stakeholders. The supports include:

1. written materials relating to testing (directions for administration, scoring, interpretation, analysis, and use for diagnostic purposes)
2. specific directions for scoring paper-based tests (including, e.g., coding or marking guides)
3. a central scanning service (New Zealand)
4. electronic scoring and analysis service (Finland, Netherlands)
5. provision of computer software for analysis of results (France)
6. automatic scoring of computer-based tests (Denmark, Norway)
7. information on websites including description of test instruments (Denmark), 'best practice' items (Denmark), how to use assessment information to improve learning (Scotland)

8. a variety of forms of inservice including on-site courses for schools (Denmark), professional development workshops (New Zealand), and support services to schools on request (New Zealand).

In addition to support for teachers, guidance is also provided in the Netherlands for parents in interpreting test scores. Some information is also available on the internet. However, ensuring that parents are adequately informed is considered to be the responsibility of the school. In Scotland, a website with advice on using assessment information to support learning is intended to be of use to local authorities, parents, and students, as well as teachers.

Test Development

Tests in Finland and France are developed by a government agency. In other jurisdictions the task is contracted to a specialist agency: the National Institute for Educational Measurement (CITO) in the Netherlands, the New Zealand Council for Educational Research, and a university in Norway.

Vertical Linking of Performance

Tests are vertically linked to allow the tracking of student progress over time in Denmark, the Netherlands, New Zealand, and Norway.

Students with Special Educational Needs

Practice varies widely in regulations regarding the testing of students with special educational needs.

In Denmark and France, students with special educational needs should participate in assessment programmes. It is also recommended that they be included in Scotland, but that they should be provided with the support they normally receive in the classroom.

In the Netherlands and New Zealand, the inclusion or exclusion of students with special educational needs in testing is a matter for individual school policy. This fits with the general policy of leaving it to the discretion of the school whether or not it uses assessment procedures developed outside the school.

In Finland and Norway, students with special educational needs are exempt from testing. In Norway, it is recommended that the decision not to include such students be based on the agreement of parents and teachers that testing would not be of benefit to the student.

Reporting the Results of a Standardised Test

When schools administer tests on their own initiative or in a census-based national assessment, policy in all countries indicates that test results are primarily for teacher use. In some countries, the results may comprise detailed diagnostic information which may be accompanied by a range of strategies to address identified student learning difficulties.

Students and their parents are also brought into the information network in all countries. Students may be informed orally or in written form. Parents too may be informed in writing and/or at a teacher-parent meeting. While students may get numerical data (e.g., in the Netherlands they are provided with a scale score, raw score and percentile rank), reports to parents attempt to be less technical. For example, in Denmark, student performance is described as 'well below average', 'below average', 'average', 'above average', or 'well above average'. Reports may be issued in the language spoken at home by students.

Test results may also be aggregated to the level of their school and reported to various stakeholders and this clearly indicates that testing has a role beyond teacher support. In most countries, information on

school performance is provided to the school board and, in some cases, to the local community. School-level data may also be available to the inspectorate (e.g., in the Netherlands). Some local authorities, (e.g., in Finland) aggregate the results for the schools in their jurisdiction.

A number of countries have faced the issue of publication of results of testing that permits comparisons to be made between the performances of schools. This, of course, does not arise in countries where test results are considered confidential and publication (except data aggregated to national level) is prohibited by law (as in Denmark). Elsewhere, even if not supported by legislation, the publication of school results has encountered resistance. In Finland, media pressure to publicise league tables was resisted by state and local education authorities. In Norway, following media publication of league tables some years ago, access to the website containing results has been restricted so that only individual schools can have access to their own results. In the Netherlands, school-level data are considered the property of the school and can only be made public with the agreement of the school.

Conclusion

Our review of the use of standardised tests and related assessments in several countries shows a broad range of practices. Across all countries, however, a key purpose of assessment is to provide information that will support teaching and learning. In general, the tests do not have high stakes attached to performance and are not used for strong accountability purposes. This suggests that they are less likely to give rise to some of the negative effects associated with high stakes testing described in Chapter 4, such as restricting the implemented curriculum to what is tested, raising levels of stress and anxiety among teachers and students, and reducing emphasis on

formative assessment. However, as Mons (2009) noted in her recent review of the effects of standardised assessments, relatively little research has been conducted into the effects of national tests in European countries. Mons does not distinguish between standardised tests (e.g., those used in TIMSS and PISA) and public examinations. Furthermore, she seems unaware of research carried out in Ireland relating to standardised tests (Kellaghan et al., 1982) and public examinations (Madaus & Greaney, 1985).

There are some commonalities across the countries we examined. A majority conduct survey sample national assessments as least once during lower secondary schooling, even though most also had a public curriculum-based examination from which some data on standards might be gleaned. With the exception of France, the focus in sample-based national assessments is usually on mother tongue, mathematics, and, sometimes, a foreign language. In France, each subject on the curriculum is assessed in a six-year cycle. This model would seem appropriate if the goal is to obtain an overview of standards in all aspects of the curriculum and information on trends, particularly if such information is not available from other sources. The assessment of cross-curricular competencies in Finland, Scotland and Northern Ireland is also interesting in that it may serve to focus the attention of schools and teachers on competencies such as problem solving, use of learning strategies and skills, and the ability to work in groups. However, it is unclear how reliable the scores assigned to students on these competencies are, or indeed how schools and teachers use any information that might be gleaned from the assessment.

Several of the systems we examined implement school-based assessments designed to provide teachers with diagnostic information to inform student learning. The implementation of a diagnostic test at the beginning of lower secondary schooling in France is well

established. Standardised tests or related measures that provide diagnostic information are also used in Denmark, Scotland and New Zealand. The degree of autonomy enjoyed by teachers in Scotland is notable, in that teachers decide on the most appropriate tests for students, based on the likelihood that they will be successful. In Denmark, the use of computer-based testing means that students take the items most suited to their level of ability, leading to more accurate estimates of their achievement (adaptive testing). Given the potential of the tests administered in these countries to provide diagnostic information to teachers, their use should be considered as part of a broader range of supports for teachers and students.

Finally, our review indicates that, in general, the reporting of test scores to parents is uncontroversial. Where reliable individual student scores are available, they are typically reported, sometimes in meetings with teachers. Moreover, supports such as websites are available to parents in some countries, though they are not always widely accessed. The practice of reporting outcomes in the media is controversial, but, in the case of school-based tests with a strong formative purpose, it is generally avoided.

CHAPTER 6

THE UTILITY OF INTERNATIONAL ASSESSMENTS

International comparative assessments of student achievement grew out of a consciousness in the late 1950s and early 1960s of the lack of internationally valid standards with which individual countries could compare the performance of their students. As well as providing data for comparisons, it was envisaged that the studies would capitalise on the variability that exists across education systems, exploiting the conditions that 'one big educational laboratory' of varying school structures and curricula provides, not only to describe conditions, but to suggest what might be educationally possible (Husén & Postlethwaite, 1996). While international studies may not have fulfilled the dreams of their early pioneers in identifying factors associated with high performance that could be transported from one education system to another, they do provide evidence that merits the attention of policy makers. In this chapter we examine data to answer four questions relating to the utility of international assessments:

1. Do tests used in international studies measure the same constructs as a national system of assessment (e.g., the Junior Certificate Examination)?

2. Why might performance on an international assessment differ from performance on a national system of assessment (e.g., the Junior Certificate Examination)?

3. What can an international assessment tell us about standards of student achievement?

4. What can an international assessment tell us about the stability of standards of student achievement over time?

Do tests used in international studies measure the same constructs as a national system of assessment?

This question may be rephrased to ask: is the domain of achievement

(e.g., mathematics, science) construed in the same way in different systems of assessment? It is of interest to policy makers to know if, for example, the achievement assessed in the Junior Certificate Examination is very similar to that agreed by international experts and assessed in PISA. Or it might be even more interesting to know that it is not. Such a finding should surely prompt a review of the domain measured in the Junior Certificate Examination, following which an examination and its associated syllabus might, or might not, be adjusted to conform more to international standards. Indeed, this and other concerns prompted an international review of the mathematics curricula in post-primary schooling (Conway & Sloane, 2005), leading to the development of a new mathematics curriculum (Project Maths) for post-primary schools.

If two assessments measure very similar domains of achievement, students' performance on one assessment should closely parallel their performance on the other. This issue was investigated when the performance of students on PISA 2003 was correlated with their performance on the Junior Certificate Mathematics Examination taken in either 2002 or 2003 (Cosgrove et al., 2005). The correlation between Junior Certificate Examination performance on mathematics and overall performance on PISA mathematics was found to be .75. Correlations between Junior Certificate Examination performance and PISA mathematics content areas ranged from .68 (Space & Shape) to .74 (Uncertainty). Similar results were obtained when performance on the Junior Certificate Examination in science was correlated with performance on PISA 2006 science (r=.70) (Eivers, Shiel, & Cunningham, 2008). A similar correlation (.68) was also found between TIMSS 1995 mathematics scores in first year post-primary education and performance on the Junior Certificate mathematics examination at the end of third year, even though in this case there was a two-year interval between the two assessments (Sofroniou & Kellaghan, 2004).

These findings are similar to those of studies carried out in other countries. For example, in England, statistically significant correlation coefficients were found between students' Key Stage 3 level scores in English (at age 14) and PISA 2000 scores (at age 15) for reading (r=.73). Relationships were somewhat stronger for KS3/PISA mathematics (r=.82) and KS3/PISA science (r=.83) (Micklewright & Schnepf, 2006). In Iceland, a correlation of .60 was obtained between performance on the Icelandic Language Test (taken one month after PISA) and PISA reading literacy (Mejding, Reusch, & Anderson, 2004).

The values of the correlations revealed in these studies indicate that there is considerable overlap between performance on international assessments and on local assessments. However, the overlap is not sufficiently large to support the inference that precisely the same domain is assessed in the two assessments. Mathematics (or Science) as construed by PISA is not the same as Mathematics (or Science) as construed in the Junior Certificate Examination or in other countries' national assessments. This should prompt investigation of the type addressed in our next question.

Why might performance on an international assessment differ from performance on a national system of assessment (e.g., the Junior Certificate Examination)?

Two approaches were adopted in attempting to answer this question. In the first, PISA assessment instruments were judged on their likely familiarity to Irish students. In the second, the frameworks of PISA and of the Junior Certificate syllabus were compared to determine degree of overlap.

The study of the familiarity of PISA to Irish students involved experts (experienced teachers involved in setting and/or marking Junior Certificate Examinations) examining each PISA item to make

a judgment about how likely it was to be familiar to students in terms of (i) assessed processes/concepts; (ii) the context in which the item was embedded and its applications; and (iii) the format of the item (e.g., multiple-choice, constructed response). Judgments were made separately for students taking Foundation, Ordinary, and Higher Levels.

Here the ratings for mathematics in 2003, when it was a major assessment domain in PISA, are considered. Table 6.1 shows that the concepts underlying almost 70% of PISA items were judged by the expert raters to be 'somewhat' or 'very' familiar to students taking Higher level. The corresponding estimates for Ordinary and Foundation levels were 65% and 48% respectively. Two-thirds (66%) of the contexts/applications underlying items were expected to be unfamiliar to students taking Higher level, and 80% to students taking Foundation level. The formats underlying 63% of items at Higher level and 80% at Ordinary Level were deemed to be unfamiliar. These figures reflect the fact, firstly, that many PISA mathematics items are embedded in real-life contexts, whereas many Junior Certificate Examination questions in mathematics tend to be context-free, and secondly, that the multiple-choice format is not used to the same extent in the Junior Certificate as in PISA.

A familiarity rating was computed for each PISA booklet and correlations between booklet familiarity and performance on PISA mathematics were calculated. Correlations were 0.21 ($p. <.001$) for familiarity with contexts/applications, 0.28 ($p. < .001$) for familiarity with formats, and 0.37 ($p. <.001$) for familiarity with concepts. Hence, students' expected familiarity with concepts was more strongly associated with performance than was familiarity with either contexts/applications or format, even though several PISA mathematics items were presented in contexts and formats not found in the Junior Certificate mathematics examination.

2003 (N = 85 items)	Not Familiar	Somewhat Familiar	Very Familiar
Concept			
Higher	30.6	24.7	44.7
Ordinary	35.3	29.4	35.3
Foundation	51.8	25.9	22.4
Context/Application			
Higher	65.9	22.4	11.8
Ordinary	70.6	20.0	9.4
Foundation	80.0	16.5	3.5
Format			
Higher	62.4	24.7	12.9
Ordinary	72.9	20.0	7.1
Foundation	83.5	14.1	2.4

Source: Cosgrove et al. (2005), Table 6.14.

In our second approach to attempting to determine why performance on an international assessment might differ from performance on a national system of assessment, the frameworks of PISA and the Junior Certificate mathematics syllabus were analyzed and compared. The results indicated that 30% of PISA items did not appear in the Junior Certificate syllabus at Higher level, while 50% were not found at Foundation level (Cosgrove et al., 2005). Moreover, the PISA Space and Shape items that were on the Junior Certificate syllabus were more likely to be found in Applied Arithmetic and Measure than in Geometry. Close (2006) compared the two assessments in the opposite direction. He used the PISA framework to classify items on the 2003 Junior Certificate mathematics examination at Higher, Ordinary, and Foundation levels (190 items in all) with reference to the content areas and processes (competency clusters) assessed. The vast majority of Junior Certificate

items were found to fall in the PISA Reproduction competency cluster (indicating that they assessed more lower-order mathematics processes), while no Ordinary or Foundation level items, and just a handful at Higher level, were categorised as Reflection items.

What can an international assessment tell us about standards of student achievement?

One of the initial purposes envisaged for international assessments was to provide standards with which individual countries could compare the performance of their students. This has been done most frequently simply by comparing mean scores of education systems in league tables.

In Ireland, average performance on PISA reading literacy has been well above the corresponding OECD country average in all three PISA cycles (2000, 2003 and 2006). Performance in mathematics has not been significantly different from the OECD average, while performance on scientific literacy has been just above the OECD average.

A number of studies have recently been carried out in which standards (explicit or implicit) on local assessments were compared with standards on an international assessment. Cosgrove et al. (2005), for example, found that in 2003, while only 8% of Ordinary level Junior Certificate students were awarded a grade E or lower in the Junior Certificate Mathematics examination, 22% achieved at level 1 or below on PISA mathematics. Furthermore, about 14% of students at Ordinary level had very low achievement (Level 1 or below) on PISA, even though they achieved a grade D or higher on the Junior Certificate Examination. Clearly the 'standards' on the Junior Certificate Examination are lower than on PISA. Indeed, quite a number of students who were awarded a grade D or higher could be considered on the basis of their PISA performance to have achieved

a level of mathematical literacy that would be insufficient to meet their future needs in education and later life.

Cartwright et al (2003) reported very different results in their study of student performance on an annual Foundation Skills Assessment (FSA) in British Columbia and on the PISA combined reading literacy scale (Figure 6.1). The threshold of the highest FSA performance level ('exceeds expectations') was set well above the threshold for PISA Level 5 (the highest level of PISA reading literacy). While 9% of students scored at the highest FSA reading level, almost 18% scored at the highest PISA level.

When the performance of selected countries in PISA 2000 was projected onto the FSA (British Columbia) scales, it was found that while 19% of students in Finland (the highest scoring country in PISA 2000 reading) scored at the 'not yet meeting standards' benchmark for British Columbia, only 7% performed at or below Level 1 on PISA.

In the United States, Phillips (2009) used a broadly similar method to that used by Cartwright et al. to establish links between performance on mathematics at Grades 4 and 8 in the 2007 US National Assessment of Educational Progress (a sample-based national assessment conducted at regular intervals) and in TIMSS 2008. Using a grade-based system (A, B, C, D and BD – below D), Phillips placed state-level performance on NAEP on the TIMSS[1] proficiency scale. Identifying Level B as the level at which US states and large school districts should seek to perform[2], he found that, at Grade 8, only Massachusetts approached this average level of performance[3]. As the

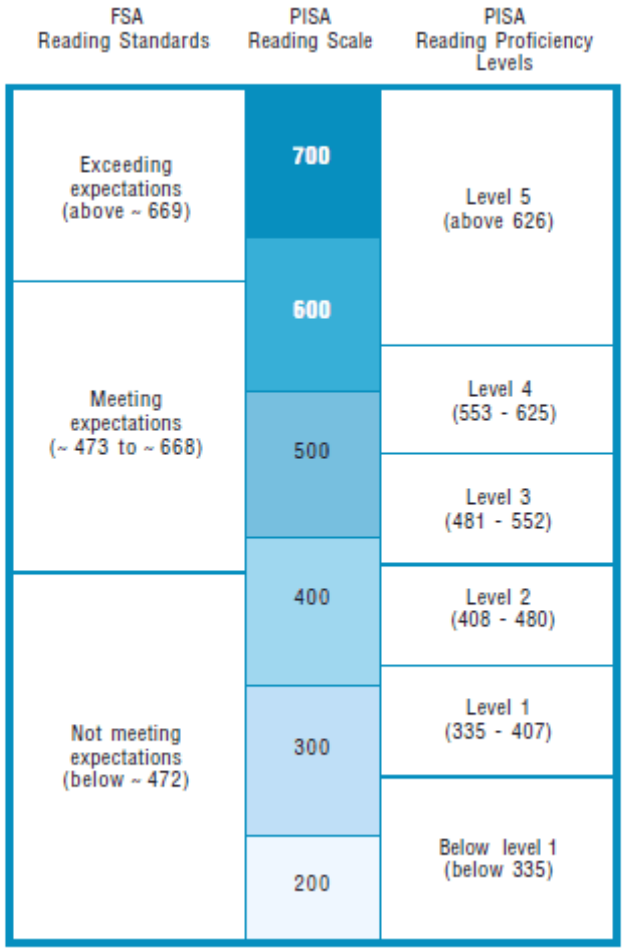
1 Prior to 1999, TIMSS was known as the Third International Mathematics and Science Study. From 1999 onwards, it is known as Trends in International Mathematics and Science Study. Ireland participated in TIMSS 1995 (Grades 4 and 8), but not in subsequent TIMSS assessments.

2 Level B is equivalent to 'proficient' on the NAEP scales.

3 A grade with a plus or minus was used if a state or country mean was more than halfway between the midpoints of adjacent benchmarks (proficiency levels).

OECD average on TIMSS was identified as Grade C, countries close to this average (e.g., England) were also identified as performing below a level of performance regarded as proficient on NAEP. Phillips argued that states in the US (and, by implication, countries performing at Grade C or lower) would need to make substantive rather than incremental progress if they were to achieve Grade B, a standard already achieved in a number of Asian countries.

Figure 6.1: FSA (British Columbia) Reading Standards Projected onto the PISA Reading Proficiency Scale



Source:Cartwright et al. (2003, Figure 5)

A very different approach to the use of international data to reflect on the national situation was adopted by Cosgrove et al. (2005) when they used PISA data to throw light on the appropriateness of students' placement in a curriculum track. When they examined student performance on PISA, they found that some students who had taken Ordinary level Mathematics in the Junior Certificate Examination outperformed students who had taken Higher level. While 10% of students who had taken the Ordinary level examination in 2003 achieved at Level 4 on PISA, 9% of students who had taken Higher level only achieved at Level 2 (Table 6.2). Such findings clearly have implications for educational guidance and the placement of students in curriculum tracks.

	Percent of Students at PISA Proficiency Levels				
	At or below Level 1	Level 2	Level 3	Level 4	Levels 5 and 6
Higher	1.5	9.0	28.8	35.8	24.9
Ordinary	21.9	36.2	30.4	9.9	1.6
Foundation	71.9	22.5	5.5	0.0	0.0

Source: Cosgrove et al. (2005, Table 6.19)

What can an international assessment tell us about the stability of standards of student achievement over time?

The charge is frequently made that Leaving Certificate Examination results have been subject to 'grade inflation' over the years. The Junior Certificate Examination has received less attention in this context, no doubt because less significant consequences are attached to performance for most students. A problem in interpreting an increase in the proportion of high grades being awarded in either examination is that the content of examinations changes from year to

year. However, this is not the case in international assessments, and if it changes, performances can still be linked across assessments.

It is clear from Tables D1 to D3 (Appendix D) that some changes are in evidence in the percentage of higher grades awarded in Junior Certificate Examinations since 2000 (the first year in which PISA was administered). For example, in 2000, 71% of students achieved a grade C or higher on Higher level English. This had increased to 78% by 2006 (Table D1). However, performance on PISA reading literacy did not change significantly between 2000 and 2006 across domains or assessment cycles, either in terms of average scores or scores at key benchmarks such as the 5th and 95th percentiles, except that students scoring at the national 90th percentile did less well in 2003 than in 2000 (Eivers et al., 2008). Thus, it would appear that a factor or factors other than enhanced reading literacy was responsible for the increase in the percentage of high achievers on the Junior Certificate Examination. The percentage achieving a grade C or higher on Higher level mathematics also increased from 2000 to 2006 (from 66% to 78%), during a time when overall PISA average scores, and scores for students at key PISA benchmarks (percentile points), did not change. However, it seems that Junior Certificate Higher-level mathematics was particularly difficult in 2000, since percentages range from 73 to 80 for all other years listed in Table D2. A similar pattern is evident for Foundation level mathematics. While the percentages achieving grade C or higher on Higher level science were virtually the same in 2000 and 2006 (70% and 71% respectively; Table D3), the general trend over the period is for more students to achieve higher grades. In 2007, for example, 78% achieved grade C or higher. Again, this occurred during a period in which no changes were recorded on the PISA science test, although a revised Science curriculum, examined for the first time in 2006, was introduced in 2003. It should be noted, however, that grades seem to have stabilised somewhat since the establishment of the State Examinations Commission in 2003.

Conclusion

The achievements that international assessments construe differ somewhat from those of national systems of education. This is a disadvantage in that an international assessment may not provide an accurate assessment of how well students have learned the content of national curricula. It may, however, also be an advantage if it causes national authorities to review their curricula in light of students' performance on the international assessment.

Information from an international assessment (PISA) described in this chapter also raised questions about standards of achievement represented in the grades of the Junior Certificate Examination, as well improvement in achievements over time, as indicated by an increase in the proportion of high grades awarded in the examination.

In general, links between standardised tests used in national assessments and those used in international assessments have been established by comparing the performance of students who have taken part in both types of assessment at around the same time or who belong to equivalent groups (for example, representative samples at the same grade level). A step beyond this is to incorporate test items from an international assessment in a national assessment, as has been done in a number of countries. In proposing the introduction of a national sample survey to replace Key Stage 3 tests in England, the Expert Group on Assessment (2009) recommended that, 'where possible, test items should be linked to international comparison surveys in which England already participates (e.g., TIMSS)' (p. 35). In Ireland, the revised Junior Certificate science syllabus introduced in 2003 (DES, 2003), and examined for the first time in 2006, makes a number of references to PISA.

CHAPTER 7

CONCLUSIONS AND OPTIONS

This section draws together the information presented in previous chapters to form some broad conclusions about current practices in standardised testing. Conclusions are organised into the following sections: development of standardised testing; organisation of assessment practices; defining standardised testing; areas assessed; functions of testing; control of testing; reporting to parents; the issue of stakes; innovations in assessment practices; the utility of international assessments; and the utility of national assessments.

Development of Standardised Testing

Our outline of the development of standardised tests indicates a considerable increase over time in the use of such tests. While we do not have detailed comparative data for the education systems considered in our review, there is evidence that all systems are adding standardised procedures to their suite of assessments. This reflects the situation in Ireland where most activity has been concentrated at the primary school level.

Organisation of Assessment Practices

In all the education systems considered for this review, formal procedures (including standardised tests) now play a role in their systems of assessment. There is, however, considerable variation in how those procedures are organised and, in particular, in their relationship to the informal assessment practices involved in classroom assessment. In some, a single assessment system serves the dual function of providing information for classroom use and information about the performance of the system (Denmark, Norway and France). In others, support for classroom assessment (in the form of standardised tests, item banks, 'best practice' items, assessment case studies, and self-assessment toolkits for schools and teachers to audit their own practices) is separate from procedures to monitor the performance of the education system.

Defining Standardised Testing

There also appears to be variation in how the term *standardised test* is interpreted. Some of this may be due to different experiences in the use of such tests. It is, however, surprising that in some systems, considerable teacher discretion is allowed in administration. In Scotland, teachers assess listening and talking, and can provide support to students with special educational needs. In Denmark, it is left to teachers to decide how much support to give students whose first language is not Danish. These provisions clearly violate standards for test administration set out in Chapter 2.

Areas Assessed

In all systems, provision is made for the assessment of students' basic language (usually reading) and numeracy skills. There was some variation in the additional constructs or curriculum areas that were assessed. It is of interest that in the Netherlands and Norway, English, as well as the national language, is assessed. Also notable is the general absence of science among the curriculum areas for which formal assessment procedures were specified or available.

Functions of Testing

In all countries, the primary function of formal assessment procedures was stated to be to support teaching and learning in the classroom by, for example, providing evidence that teachers could use in adapting teaching to the needs of individual students, in allocating students to instructional groups, in diagnosing student learning difficulties, in identifying students in need of further investigation, and in deciding whether to retain or promote students. It was envisaged that decisions would not be based on test information alone. Rather, test information should be considered as just one element of information that was relevant to any pedagogical decision.

Control of Testing

There is some evidence across countries of a shift in emphasis to achieve balance between internal school assessments and assessments that are external to the school. In the Netherlands, for example, which has a long tradition of external testing, efforts are being made to accord a greater role to teachers' judgments, for both formative and summative purposes, in the assessment of students. In Finland, on the other hand, where the tradition has been to accord teachers major responsibility for assessment, government is currently strengthening an external evaluation system.

Despite the claim that the judgments of teachers are accorded priority in making assessment decisions, whether on the basis of informal procedures or evidence from externally devised tests, there is also evidence, even if not formally recognised, of a concern with issues of accountability, standard monitoring, the use of performance indicators, and quality assurance, all of which are associated with a corporatist approach to administration, and are significant features of education policy in England and in the United States. A number of features of the assessment systems we considered support this view. Making testing compulsory, as is the case in Denmark, Norway and France, would tend to suggest that teacher judgement is not entirely to be trusted. On the other hand, a situation in which the use of externally devised assessment procedures is entirely voluntary and left to the discretion of teachers, as is the case in Finland, Scotland, Netherlands, and New Zealand, would tend to support the view that the teacher's role in assessment is preeminent. Similarly, schools in Northern Ireland may now opt into national tests at the end of Key Stage 3 (age 14).

Another feature of an assessment system that has implications for whether tests are used for formative purposes (under the control of teachers) or for summative purposes is the time at which tests are

administered, and the associated issue of the time of year for which norms are provided. When most schools test towards the end rather than the beginning of the year (as in the Netherlands and Scotland), this suggests a summative rather than a formative function for the tests.

Reporting to Parents

The emphasis on reporting to parents, which was a central feature of all the assessment systems we examined, is also indicative of a concern with accountability and quality assurance. This position may be contrasted with that which obtained in Sweden in the 1960s, where reporting to parents was not encouraged because it might have led to coaching or other undesirable practices (Chapter 3). However, while reporting to parents might be nothing more than a recognition of the important role that parents play in their children's education, many commentators would also regard it as an important component of an accountability system.

The Issue of Stakes

Among the countries we investigated, only Ontario (Canada) seemed to attach high stakes to assessment information in ways that are common in England and the United States, where information on the performance of schools is published in league tables. Indeed, in Denmark, the publication of any results, except data aggregated to the national level, is prohibited by the same legal framework that protects national and military secrets, with potential imprisonment as punishment. In Finland, proposals to publish school-level outcomes on national sample-based assessments met with objections from the general public. Whether or not high stakes are attached to testing is a crucial consideration when devising an assessment system.

Assessments can also be high-stakes if there are serious consequences for teachers and students. When sanctions are attached to student

performance, negative, if unintended, consequences can ensue. Teachers will tend to react by aligning their teaching to the knowledge and skills assessed in the test ('teaching to the test'), while neglecting curriculum areas (e.g., art, social studies) that are not assessed. They will also tend to emphasise rote memorisation, routine drilling, and a passive approach to learning, rather than an approach that stresses higher-order reasoning and problem solving skills (see Chapter 4). It should be noted that these effects are likely to ensue whatever the nature of the assessment instruments. In fact, most of the evidence relating to them comes from observations on public (essay-type) examinations, not standardised tests.

Our review of issues in the use of standardised tests was important in this context (Chapter 4). There we saw that when a testing programme is under the control of teachers and sanctions are not attached to student performance, either for students or teachers, the negative effects outlined above are not in evidence. There are dangers attached to any evaluation programme, including one in which standardised tests feature. Evaluation information may be used inappropriately to determine the subject matter that is taught or to allocate students to grades or curriculum tracks. However, there are also benefits attached to the information provided by standardised tests. For example, while test information disrupts teachers' perceptions in creating an expectancy process, teachers create their own expectancies in the absence of information provided by the tests. Furthermore, as noted in Chapter 4, expectancies based on test information resulted in more favourable shifts than expectancies based on teachers' perceptions which were not informed by test results.

Innovations in Assessment Practices

Our survey of assessment practices in other countries points to a number of innovations which we might expect to become more

common in time: item banking, computer-based testing, and linking of the performance of individual students vertically on a series of tests through item response modelling to allow an estimate to be made of their scholastic progress (Netherlands, New Zealand).

The Utility of International Assessments

Given the cost and imposition on schools of national and international assessments, it is reasonable to ask if the information they provide outweighs their disadvantages. Our review of the utility of international assessments provides evidence that the findings of an international assessment can have important implications for national policy (Chapter 6). In particular, we saw that standards on the Junior Certificate Examination are lower than on an international assessment (PISA) when we compare the proportions of students awarded low grades on the two assessments. We also saw that much higher levels of achievement than were attained in Ireland are possible. Other findings indicated that recent increases in the award of higher grades on the Junior Certificate Examination in language, mathematics, and science were not matched by an improvement in performance on PISA. Finally, the PISA results raised issues that have implications for educational guidance and the placement of students in curriculum tracks.

The Utility of National Assessments

A question that policy makers need to consider is whether a national assessment should be administered in post-primary schools in addition to an international assessment. An argument in favour would be that PISA is age-based and is not designed to reflect national curricula, though it does, of course, provide the opportunity to evaluate national curricula in the light of international experience. An argument against a national assessment would relate to cost. Clearly, if a decision were to be made to carry out a national

assessment at post-primary level, the decision should be based on a careful analysis of how the information it would provide would complement the information obtained from PISA and other sources, and how it would be used for policy.

Options

In this section, we present a series of options relating to the implementation and use of standardised tests in lower secondary schooling in Ireland. The areas in which options are outlined are: implementing standardised testing in schools; reporting outcomes of standardised tests to parents; reporting outcomes to students; using technology to support assessment; developing classroom-based assessments; developing teachers' assessment skills; establishing a sample-based national assessment; and planning for development in assessment.

Implementing Standardised Testing in Schools

Option 1: Standardised tests of achievement in literacy (English/Irish) and numeracy with Irish norms are developed for the three years of lower secondary schooling, and made available to schools to be administered when considered appropriate, to support monitoring the progress of students ('the Netherlands model').

Option 2: Standardised tests are developed and mandated for use at one point in lower-secondary schooling, such as the first term of first year ('the French model', but without central reporting), or the end of Second/beginning of Third year, when results might be used for guidance purposes (e.g., advising on the level at which to study Junior Certificate subjects).

Option 3: The outcomes of standardised tests are presented as summative information (i.e., a student's overall performance, using, for example, proficiency levels), diagnostic information (i.e., information designed to support schools and teachers in developing students' learning), or some combination of summative and diagnostic information.

On the basis of our review, we see evidence of a need for standardised testing in lower secondary schooling, to assist teachers in diagnosing student learning difficulties and in establishing learning programmes to address those difficulties. This is common practice in most countries whose assessment systems we reviewed. The strong emphasis attached to target setting in literacy and numeracy in the DEIS blueprint (DES, 2005), and the need for individual schools in the School Support Programme under DEIS to establish targets as part of their DEIS development plan[1], also point to value in implementing standardised testing on a more formal basis in lower secondary schools.

A problem at the present, however, is that there are no group-administered standardised tests of achievement with current Irish norms available to post-primary schools. In this situation, it seems that many schools use tests that have been normed at primary level or tests with British norms to assess the achievements of incoming students. It would seem important that tests with current norms be made available to schools, and that the tests be revised and/or re-normed every 5 to 7 years.

There are a number of other issues that arise from the options outlined in this section. One is whether the results of standardised tests should be available to the Department of Education and Science (perhaps in summary form) as occurs at primary level in the context of Whole School Evaluation, or whether results would be used only by schools as part of their own school development planning or in planning by individual teachers.

Another issue is whether schools should be required to use a specific standardised test that had been developed centrally (a practice in many countries, especially when the standardised test is part of a

1 School Development Planning provide a template for a DEIS Three-year plan at http://www.sdpi.ie/SDPI_DEIS_Docs/DEIS_Planning-Action_Plan_DES_Approved.doc

national assessment) or should be allowed to select from a range of available tests (the current practice at primary level in Ireland). The latter option would seem to preclude the use of standardised test results at an administrative level higher than the school.

A third issue that arises when tests are administered at more than one point in time is whether scales should be established which would allow schools to monitor progress over time. Although standardised tests could be implemented at one point in time (for example, the beginning of first year), there may be value in developing tests and establishing scales which would allow schools to monitor progress in relation to student- and school-level targets over time (e.g., between the beginning of first year and the end of second year).

A fourth issue relates to the cost to schools of purchasing standardised tests and related services (e.g., paper and pen tests, online testing, electronic scoring, computerised reports). Currently, schools receive an annual grant for the purchase of tests. This may need to be increased.

Reporting Outcomes of Standardised Tests to Parents

Option 4: Support to parents in interpreting their child's scores on standardised tests is given in written reports that include explanations of what the test scores mean and a description of the implications of the scores for their child's learning.

Option 5: Information is given to parents in face-to-face meetings with teachers, or through a combination of written reports and face-to-face meetings. There may be some value in providing web-based support to parents who may need additional information.

Option 6: The information given to parents would be limited to normative information.

Option 7: The information provided to parents would, in addition to normative information, include information on proficiency, information on progress, and diagnostic information.

Following the practices of a number of countries at lower secondary level, it would seem important to provide parents with the results of their child's performance on a standardised test, taking care to ensure that the information reported is comprehensible, and that appropriate support in interpreting results is provided. Such support could take the form of written explanation, guidance on where to obtain additional information, and face-to-face meetings with students' teachers as appropriate. In some cases, three-way conferences involving teacher, parent and student may be appropriate.

Reporting Outcomes of Standardised Tests to Students

Option 8: Results are reported to students in summary form only, without reference to context or specification of future learning needs.

Option 9: The results of standardised tests are reported to students, along with an explanation of what they mean, how they relate to other assessments completed by the student, and steps that need to be taken to improve learning.

Option 10: Feedback is provided by subject/form teachers.

Option 11: Feedback is provided by guidance counsellors/support teachers.

There are advantages in having students' subject/form teachers report the results of testing as this tends to emphasise links between student performance and classroom teaching and learning. However, it may be an established tradition in some schools to have specialised teachers or guidance counsellors provide students with results. In either case, it would seem important that students reflect on the results they achieve, and relate them to self-assessments of their learning.

Using Technology to Support Assessment

Option 12: Standardised tests would continue to be administered in paper-and-pencil format, and scored electronically or by hand.

Option 13: Standardised tests would incorporate recent advances in administering and scoring tests electronically, and in generating reports that would be useful to schools, teachers and parents.

Technological advances in testing include the use of computer software or the internet to deliver tests, the development of item banks (pools of items from which a test developer or teacher can draw questions as needed), the use of adaptive testing principles during testing, electronic scoring of tests, and generation of reports electronically. Over time, some of these developments could be incorporated into standardised testing at lower secondary level. Indeed, the delivery of tests in electronic format is now standard in countries such as Denmark and the Netherlands, and is also a part of international assessments such as PISA.

These developments clearly raise issues about the cost of test development and the maintenance of an assessment system. Investment should be made in this area only if there is clear evidence that the proposed developments will lead to change that is beneficial and cost-effective. Furthermore, one can expect a variety of problems in implementing such a system. The development of computer-based adaptive tests in Denmark took considerably longer than expected because of technology-related problems, and some capacity problems persist (e.g., the numbers of students that can be assessed at a given point in time is limited).

Developing Classroom-based Assessments

Option 14: Standardised testing would be introduced without any link to classroom assessment procedures.

Option 15: In parallel with the introduction of standardised testing, schools and teachers would be facilitated in using a broader range of classroom assessments, both electronic and paper-and-pencil, to allow students' progress towards key learning targets to be monitored on an ongoing basis.

In several countries, including Scotland, the Netherlands and New Zealand, teachers are provided with standardised tests and other materials designed to support ongoing classroom-based assessment of their students. Clearly, teachers in lower secondary schools in Ireland could also be supported in this way, so that evidence-based assessment becomes a more prominent feature of teaching and learning. The provision of classroom-based assessment tools, such as item banks (i.e., clusters of test items that could be used by teachers on a needs basis to assess students' learning, for example at the end of a unit of study) could make a significant contribution to the support of student learning. This would be consistent with recent efforts by the NCCA (2005) to enhance the assessment skills of subject teachers at post-primary level. Initially, support for classroom assessment could be provided in the areas of literacy and numeracy. If, as in state-supported systems in Scotland and New Zealand, it is envisaged that classroom assessment will be linked to key learning targets and standards, it may be necessary to identify the key standards in a more precise way before proceeding with the development of instruments to assess achievement of the standards.

We see the development of classroom-based assessments as being important if teachers are to follow up effectively on student difficulties identified through standardised testing. However, the development of classroom-based assessments may require a somewhat

longer time span than the initial development of standardised tests. Hence, priority may need to be given to the development and administration of standardised tests, with a later emphasis on the development of classroom-based assessments, some of which could be technologically-based.

Developing Teachers' Assessment Skills

Option 16: The administration, scoring and interpretation of standardised tests would mainly involve specialists such as guidance counsellors and support teachers, with minimal input from subject teachers.

Option 17: Subject teachers would be enabled to access appropriate, ongoing in-career development in the administration, scoring and interpretation of standardised tests, and would be supported in using test results to inform teaching and learning.

Option 18: Support for teachers would be restricted to interpreting and using the outcomes of standardised tests.

Option 19: The assessment skills of subject teachers would be further strengthened by enabling them to access support on the use of a range of classroom-based assessments, as well as standardised tests.

In many post-primary schools, test administration, scoring and interpretation are carried out by guidance counsellors and support teachers, while subject teachers proceed with the business of covering the syllabus and preparing students for state examinations. This division of labour has arisen, in part, because of the special training required by guidance counsellors to administer and interpret the results of psychological tests such as the Differential Aptitude Tests. Subject teachers may be less familiar with standardised tests, or with the implications of test outcomes for teaching and learning.

If options to develop teachers' assessment skills are accepted, some development activities could be located within schools, drawing on the existing expertise of guidance counsellors and resource/support

teachers to support the work of subject teachers. It may be necessary, however, in some cases to call on external support, including support that involves coaching and mentoring.

A further key issue is whether to involve only those teachers whose curriculum areas are being assessed (perhaps literacy and numeracy/mathematics at first), or to involve all the teachers in a school. The latter option should serve to strengthen assessment in schools, to support the achievement of key targets (including literacy targets), and to promote the development of core competencies throughout the curriculum.

Establishing a Sample-based National Assessment

Option 20: A rotating programme of sample-based national assessments would be introduced, perhaps in the first term of third year, using standardised tests and other appropriate instruments. Over time, such a programme could fulfil some of the functions for which the Junior Certificate Examination may not be well suited, such as monitoring standards and the quality of teaching and learning.

It was noted in Chapter 5 that almost all of the countries whose assessment systems we examined carried out national assessments of educational achievement, even if they also held examinations at the end of lower secondary schooling and participated in international assessments. One reason for carrying out a national assessment relates to the fact that examinations are not likely to provide accurate trend data that allows for monitoring of standards over time, while international assessments may not be sufficiently sensitive to national curricula to allow for an evaluation of curriculum-based teaching and learning. The French system of assessing each subject over a six-year period is perhaps the most systematic system in place among the countries we reviewed and ensures ongoing review of a broad range of curriculum areas.

The present situation is that the Junior Certificate Examination places a heavy burden on the education system, with considerable assessment capacity at national and school levels being expended on preparing for the examination (e.g., mock exams), administering the examination, scoring students' work, and reporting results. Should the examination be modified (e.g., by reducing the number of subjects assessed, or extending provision for teacher-based assessment), it would seem important to proceed with a programme of sample-based national assessments. These could complement the work of the Inspectorate of the Department of Education and Science in evaluating teaching and learning in a variety of curriculum areas, and also help teachers to better align classroom assessments with national standards.

If national assessments are introduced in areas such as literacy, mathematics and science, the possibility of linking them to international assessments could be examined. This could be done, for example, by including items from an international assessment in a national assessment, or by projecting the performance of students on a national assessment onto the proficiency scales used in an international assessment. This exercise might raise interest in both national and international assessments.

Planning for Development in Assessment

Option 21: Bodies involved in policy and planning such as the DES and the NCCA would draw up a multi-year national plan for the development of school-based assessment at lower-secondary level. Such a plan would include a timeline for the implementation of its components, as well as procedures for evaluating the effects of implementation.

Finally, it would be for different organisations involved in assessment policy to establish a multi-year national plan for the implementation of new modes of assessment. This would ensure that new tests and

assessments were developed and rolled out in a systematic way, and that the effects of implementation could be carefully tracked.

Conclusion

Clearly, a case can be made for standardised testing in lower secondary education. Indeed, guidance counsellors and support teachers have been using standardised tests for many years. A disadvantage of the situation, however, is that standardised tests with Irish norms are not available. This situation might be expected to create problems as schools seek to establish learning targets and monitor progress, at school, class and individual student levels.

Policy makers have a number of options with respect to the implementation of standardised tests. A distinction can be drawn between the use of formal standardised tests at one or two points in time during lower secondary education and the use of a broader range of classroom assessments to inform teaching and learning on an ongoing basis. There may be value in supporting teachers in administering and interpreting both types of assessment, rather than focusing on standardised tests only.

Consideration needs to be given to whether new standardised tests of achievement might be delivered and scored electronically. While it would seem important to capitalise on emerging approaches, such as adaptive testing, the development of such tests may take some time (e.g., lessons are still being learned from the use of electronic tests in PISA; also see Scheuermann & Björnsson, J., 2009). In the meantime, there may be value in developing electronic tests for use in classroom assessments (i.e., to support teachers in assessing students after they complete a course unit), with a view to extending their use to more formal standardised testing over time. This would not preclude use of technology to score and report on the results of standardised tests in the meantime.

The value of introducing standardised tests of achievement may hinge on the uses to which teachers, parents and students put the results. It would seem important to ensure that subject teachers, as well as guidance/support teachers, are fully informed about the strengths and limitations of standardised tests, and of the relevance of test results to their work in teaching a range of subjects. If teachers are not fully informed, there is a risk that parents and students may not benefit either.

If implemented, some of the options in this report could result in a significant degree of change to existing assessment practices, including an increase in the responsibility that teachers have for administering both standardised and classroom assessments. Such change would need to be managed carefully and its effects considered at each stage. Hence, there is a need for a coherent, multi-year plan that maps out what it is hoped to achieve. Aspects of the plan that are implemented need to be evaluated to ensure that their objectives are achieved, and that unintended consequences are addressed. In particular, the effects of changes in assessment practices on at-risk groups would need to be tracked carefully.

A number of the options we presented in this chapter hinge on what happens over the next year or two with other aspects of assessment. For example, the need to introduce sample-based national assessments would intensify if substantive changes are made to the Junior Certificate Examination and information on the performance of students in each subject is no longer available on a regular basis. Similarly, changes to the structure of the Junior Certificate might create a need for exemplars of student performance that could be generated in the context of regular national assessments. For these reasons, it would be important to embed the introduction of standardised tests of achievement and other proposed changes in the context of a coherent assessment plan covering a period of several years.

REFERENCES

Airasian, P.W. (2001). *Classroom assessment* (4th ed.). New York: McGraw Hill.

Airasian, P.W., Kellaghan, T., Madaus, G.F., & Pedulla, J.J. (1977). Proportion and direction of teacher rating changes of students' progress attributable to standardised test information. *Journal of Educational Psychology,* 69, 702-709.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME) (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Anastasi, A. (1954). *Psychological testing*. New York: Macmillan.

Bacher, F. (1969). The use of tests in French schools. In K. Ingenkamp (Ed.), *Developments in educational testing* (vol. 1; pp. 59-74). London: University of London Press.

Barksdale-Ladd, M., & Thomas, K. (2000). What's at stake in high stakes testing? *Journal of Teacher Education*, 51, 384-397.

Beaton, A.E., & Allen, N.L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, 17, 191-204.

Beaton, A.E., Postlethwaite, T.N., Ross, K.N., Spearitt, D., & Wolf, R.M. (1999). *The benefits and limitations of international achievement studies*. Paris: UNESCO/International Institute for Educational Planning; International Academy of Education.

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

Bloom, B.S. (1969). Some theoretical issues relating to educational evaluation. In R.W. Tyler (Ed.), *Educational evaluation: New rules, new means. The Sixth-eight Yearbook of the National Society for the Study of Education, Part II.* Chicago: NSSE.

Bloom, B.S., Hastings, J.T., & Madaus, G.F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw Hill.

Boyle, B., & Bragg, J. (2006). A curriculum without foundation. *British Educational Research Journal*, 32, 569-582.

Cartwright, F., Lalancette, D., Mussio, J., & Xing, D. (2003). *Linking provincial student assessments with national and international assessments.* Report no 81-S9S-MIE 2003005. Ottawa: Statistics Canada.

Charting our education future. White paper on education. (1995). Dublin: Stationery Office.

Close, S. (2006). The Junior Cycle curriculum and the PISA mathematics framework. *Irish Journal of Education,* 38, 53-78.

Conway, P., & Sloane, F. (2005). I*nternational trends in post-primary mathematics education*. Dublin: National Council for Curriculum and Assessment.

Cosgrove, J., Kellaghan, T., Forde, P., & Morgan, M. (2000). *The 1998 National Assessment of English Reading.* Dublin: Educational Research Centre.

Cosgrove, J., Shiel, G., Sofroniou, N., Zastrutzki, S., & Shortt, F. (2005). *Education for life: The achievements of 15-year olds in Ireland in the second cycle of PISA*. Dublin: Educational Research Centre.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, CA:Wadsworth.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed.; pp. 443-507). Washington DC: American Council on Education.

Cronbach, L.J. (2000). Course improvement through evaluation. In D.L. Stufflebeam, G.F. Madaus, & T. Kellaghan (Eds), *Evaluation models. Viewpoints on educational and human services evaluation* (2nd ed; pp. 235-247). Dordrecht: Kluwer Academic.

Crooks, T.J., Kane, M.T., & Cohen, A.S. (1996). Threats to the valid use of assessments. *Assessment in Education* 3, 265-285.

Curriculum and Examinations Board. (1986). *In our schools: A framework for curriculum and assessment.* Dublin: Author.

Daugherty, R., (1995). *National curriculum assessment: A review of policy 1897-1994.* London: Falmer Press.

Demailly, L. (2001). Enjeux de l'évaluation et regulation des systèmes scolaires. In L. Demailly (Ed.), *Evaluer les politiques éducatives.* Bruxelles: Editions De Boeck Université.

Department of Education. (1991). *Report on the national survey of English reading in primary schools*. Dublin: Author.

Department of Education and Science. (DES). (2000). *Learning support guidelines.* Dublin: Stationery Office.

DES. (2003). *Junior Certificate science syllabus (Ordinary level and Higher level)*. Dublin: Stationery Office.

DES. (2005). *DEIS (Delivering equality of opportunity in schools): An action plan for educational inclusion.* Dublin: Stationery Office.

DES. (2006). Supporting assessment in primary schools. Circular 0138/2006. Accessed at http://www.education.ie/servlet/blobservlet/c10138_2006.doc

DES. (2009). *Looking at guidance: Teaching and learning in post-primary schools.* Dublin: Author.

DES/NCCA. (1999). *Primary school curriculum.* Dublin: Stationery Office.

Du Bois, P. H. (1970). *A history of psychological testing.* Boston: Allyn & Bacon.

Ebel, R.L. (1972). *Essentials of psychological measurement* (3rd ed.). Englewood Cliffs NJ: Prentice-Hall.

Education for a Changing World. Green paper on education. (1992). Dublin: Stationery Office.

Education Quality and Accountability Office. (EQAO). (2009). Ontario student achievement. Provincial report on the results of the 2008-09 Ontario secondary school literacy test. Toronto, Ontario: Author. Retrieved Jan 6, 2009 from http://www.eqao.com/pdf_e/09/CPRR_Xe_0609_WEB.pdf

Educational Research Centre. (1968). *Drumcondra Verbal Reasoning Test.* Dublin: Author.

Educational Research Centre. (1998). *Drumcondra Reasoning Test. Manual.* Dublin: Author.

Educational Research Centre. (2007). *Drumcondra Primary Mathematics Test-Revised. Levels 3-6. Administration manual and technical manual.* Dublin: Author.

Educational Research Centre. (2008). *Drumcondra Primary Reading Test-Revised. Levels 3-6. Administration and technical manual.* Dublin: Author.

Eivers, E., Shiel, G., & Shortt, F. (2004). *Reading literacy in disadvantaged primary schools.* Dublin: Educational Research Centre.

Eivers, E., Shiel, G., Perkins, R., & Cosgrove, J. (2005). *The 2004 National Assessment of English Reading.* Dublin: Educational Research Centre.

Eivers, E., Shiel, G., & Cunningham, R. (2008). *Ready for tomorrow's world. The competencies of Ireland's 15-year-olds in PISA 2006. Main report.* Dublin: Educational Research Centre.

Eurydice: Education, Audiovisual and Culture Executive Agency. (Eurydice EACEA). (2009). *National testing of students in Europe: Objectives, organisation and use of results.* Brussels: Author.

Expert Group on Assessment. (2009). *Report.* London: Department for Children, Schools and Families.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 105-156). New York: American Council on Education/Macmillan.

Frederiksen, J., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher,* 18(9), 27-32.

Gardner, E.F. (1969). Standardised testing in the United States. In K. Ingenkamp (Ed), *Developments in educational testing* (vol. 1; pp. 19-26). London: University of London Press.

Gephart, W.J. (1970). Will the real Pygmalion please stand up? *American Educational Research Journal,* 7, 473-475.

Gipps, C., & Stobart, G. (2003). Alternative assessment. In T. Kellaghan & D.L. Stufflebeam (Eds), *International handbook of educational evaluation* (pp. 549-575). Dordrecht: Kluwer Academic.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist,* 18, 519-521.

Hall, K. (2000). A conceptual evaluation of primary assessment policy and the education policy process in the Republic of Ireland. *Compare*, 30, 85-101.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives,* 8(41). Retrieved Nov 2, 2009 from http://epaa.asu.edu/epaa/v8n41/

Haney, W.M., Madaus, G.F., & Lyons, R. (1993). *The fractured marketplace for standardised testing.* Boston: Kluwer Academic.

Harlen, W.M., & Deakin Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning. In: *Research Evidence in Education Library.* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Henrysson, S. (1969). Educational testing in Sweden. In K Ingenkamp (Ed.), *Developments in educational testing* (vol. 1; pp. 80-86). London: University of London Press.

Hilton, M. (2001). Are the key stage two reading tests becoming easier each year? *Reading*, April 4-11.

Hoffman, J., Assaf, L., & Paris, S. (2001). High-stakes testing in reading: Today Texas, tomorrow? *Reading Teacher*, 54, 482-494.

Husén, T., & Postlethwaite, T.N. (1996). A brief history of the International Association for the Evaluation of Educational Achievement (IEA). *Assessment in Education*, 3, 129-141.

IGEN-IGAENR. (2005). *Les acquis des élevès, pierre de touché de la valeur de l'école?* Paris: Author.

Ingenkamp, K. (Ed.). (1969a). *Developments in educational testing*. London: University of London Press.

Ingenkamp, K. (1969b). The administration of school tests in Germany. In K. Ingenkamp (Ed.), *Developments in educational testing* (vol. 1; pp. 87-96). London: University of London Press.

Junior Certificate School Programme. (2006). *Room for reading. The Junior Certificate School Programme Demonstration Library Project. Research report 2005.* Dublin: Author.

Kellaghan, T. (2003). Local, national, and international levels of system evaluation. In T. Kellaghan & D.L. Stufflebeam (Eds), *International handbook of educational evaluation* (pp. 873-882). Dordrecht: Kluwer Academic.

Kellaghan, T., & Greaney, V. (2001). *Using assessment to improve the quality of education.* Paris: UNESCO. International Institute for Educational Planning.

Kellaghan, T., Greaney, V., & Murray, T.S. (2009). *Using the results of a national assessment of educational achievement*. Washington DC: World Bank.

Kellaghan, T., Macnamara, J., & Neuman, E. (1969). Teachers' assessment of the scholastic progress of students. *Irish Journal of Education,* 3, 95-104.

Kellaghan, T., Madaus, G.F., & Airasian, P.W. (1979). T*eachers' perceptions of test-taking behaviors of students.* Washington DC: National Institute of Education, US Department of Health, Education & Welfare.

Kellaghan, T., Madaus, G.F., & Airasian, P.W. (1980). *Standardised testing in elementary schools: Effects on schools, teachers, and students.* Washington DC: National Institute of Education, US Department of Health, Education & Welfare.

Kellaghan, T., Madaus, G.F., & Airasian, P.W. (1982). *The effects of standardised testing.* Boston: Kluwer-Nijhoff.

Kelly, S.G., & McGee, P. (1967). Survey of reading comprehension. *New Research in Education,* 1, 131-134.

Lindquist, E.F. (1969). Testing in the United States: Recent technological developments. In K. Ingenkamp (Ed.), *Developments in educational testing* (vol. 1; pp. 53-58). London: University of London Press.

Lissitz, R.W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437-448.

Macnamara, J. (1966). *Bilingualism and primary education.* Edinburgh: University Press.

Madaus, G. (1988). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum* (pp. 83-112). Chicago, IL: University of Chicago Press.

Madaus, G., & Greaney, V. (1985). The Irish experience in competency testing: Implications for American education. *American Journal of Education*, 93, 268-294.

Madaus, G.,& Greaney, V. (1991, January). *The effects of important tests on students: Implications for a national examination or system of examinations.* Paper prepared for the American Educational Research Association Invitational Conference on Accountability as a State Reform Instrument: Impact on Teaching, Learning, Minority-based Issues, and incentives for improvement. Washington, DC.

Madaus, G.F., & Kellaghan, T. (1992). Curriculum evaluation and assessment. In P.W Jackson (Ed.), *Handbook of research on curriculum* (pp. 119-154). New York: Macmillan.

Madaus, G.F., & Raczek, A.E. (1996). The extent and growth of educational testing in the United States: 1956-1994. In H. Goldstein & T. Lewis (Eds), *Assessment: Problems, developments and statistical issues* (pp. 145-165). Chichester: Wiley.

Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing. How they affect students, their parents, teachers, principals, schools, and society.* Charlotte NC: Information Age Publishing.

Mejding, J., Reusch, S., & Anderson, T.Y. (2004). Leaving examination marks and PISA results: Exploring the validity of PISA scores. In J. Medjing & A. Roe (Eds), *Northern lights on PISA 2003 – a reflection from Nordic countries* (pp. 215-228). Copenhagen: Nordic Council of Ministers.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 13-103). New York: American Council on Education/ Macmillan.

Micklewright, J., & Schnepf, S.V. (2006). *Response bias in England in PISA 2000 and PISA 2003*. Southampton: University of Southampton.

Mons, N. (2009). *Theoretical and real effects of standardised assessments.* Brussels: Eurydice: Education, Audiovisual and Culture Executive Agency.

Mu, M., & Childs, R. (2005). What parents know and believe about large-scale assessments. *Canadian Journal of Educational Administration and Policy, 37*. Accessed at http://www.umanitoba.ca/publications/cjeap/articles/childs.html

Mulrooney, V. (1986). National surveys of reading attainment in primary schools in Ireland. In V. Greaney (Ed.), *Irish papers presented at fourth European reading conference and the tenth annual conference* (pp. 187-200). Dublin: Reading Association of Ireland.

National Council for Curriculum and Assessment (NCCA). (1993). *Programme for reform: Curriculum and assessment policy towards the new century.* Dublin: Author.

NCCA. (2004). *Update on the Junior Cycle review.* Dublin: Author. Downloaded at http://www.ncca.ie/uploadedfiles/Publications/UpdateonJuniorCycleReview.pdf

NCCA. (2005). *Assessment for learning: Report on phase 2 of a developmental initiative.* Dublin: Author. Accessed at http://www.ncca.ie/uploadedfiles/JuniorCycleReview/InterimReportonAssessmentforLearning.pdf

NCCA. (2006). Interim report on the developmental initiative in assessment for learning in Junior Cycle. Dublin: Author. Accessed at www.ncca.ie/publications.

NCCA. (2007). *Assessment in the primary school curriculum: Guidelines for schools.* Dublin: Author.

National Education Convention. (1994). *Report*. Dublin: Stationery Office.

OECD (Organisation for Economic Co-operation and Development). (2005a). *Formative assessment. Improving learning in secondary classrooms.* Paris: Author.

OECD. (2005b). *PISA 2003 technical report*. Paris: Author.

OECD. (2007b). *PISA 2006: Science competencies for tomorrow's world, Volume 1: Analysis.* Paris: Author.

Osterlind, S.J. (1989). *Constructing test items.* Boston: Kluwer Academic.

Pedulla, J.J., Abrams, L.M., Madaus, G.F., Russel, M.K., Ramos, M.A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers.* Chestnut Hill, MA: Boston College, Lynch School of Education.

Phillips, G. (2009). *The second derivative: International benchmarks in mathematics for U.S. states and school districts.* Washington, DC: American Institute for Research.

Pollard, A., Broadfoot, P., Cross, M., Osborn, M., & Abbott, D. (1994). *Changing English primary schools? The impact of the education reform act at key stage one.* London: Cassell.

Popham, W.J. (1995). Classroom assessment. What teachers need to know. Boston: Allyn & Bacon.

Reay, D., & Wiliam, D. (1999). 'I'll be a nothing': Structure, agency and the construction of identity through assessment. *British Educational Research Journal,* 25, 343-354.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom.* New York: Holt, Rinehart & Winston.

Scheuermann, F., & Björnsson, J. (Eds.) (2009), The transition to computer-based assessment: *New approaches to assessment and implications for large-scale testing.* Ispra (VA), Italy: CRELL (Centre for Research on Lifelong Learning).

Shiel, G., & Kelly, D. (2001). *The 1999 National Assessment of Mathematics Achievement*. Dublin: Educational Research Centre.

Shiel, G., Surgenor, P., Close, S., & Millar, D. (2006). *The 2004 National Assessment of Mathematics Achievement*. Dublin: Educational Research Centre.

Smith, M., Edelsky, C., Draper, K., Rottenberg, C., & Cherland, M. (1991). *The role of testing in elementary schools.* (CSE Technical Report 321). Los Angeles: National Center for Research on Evaluation, Standards and Student Testing.

Smyth, E., McCoy, S., & Darmody, M. (2004). *Moving up: The experience of first year students in post-primary schools*. Dublin: Liffey Press.

Snow, R.E. (1969). Unfinished pygmalion. *Contemporary Psychology,* 14, 197-199.

Sofroniou, N., & Kellaghan, T. (2004). The utility of the Third International Mathematics and Science Study in predicting student's state examination performance. *Journal of Educational Measurement,* 41, 311-329.

Stecher, B., Barron, S., Chun, T., & Ross, K. (2000). *The effects of Washington state education reform on schools and classrooms.* (CES Technical Report 525). Los Angeles: National Center for Research on Evaluation, Standards and Student Testing.

Stinissen, J. (1969). Testing in Belgium. In K. Ingenkamp (Ed.), *Developments in educational testing* (vol. 1; pp. 103-108). London: University of London Press.

Torrance, H. (Ed.). (1995). *Evaluating authentic assessment: Issues, problems and future possibilities*. Buckingham: Open University Press.

Torrance, H. (2003). Assessment of the national curriculum in England. In T. Kellaghan & D. Stufflebeam (Eds)., *International handbook of educational evaluation* (pp. 827-842). Dordrecht: Kluwer Academic.

Tyler, R.W. (1968). Critique of the issue on educational and psychological testing. *Review of Educational Research,* 38, 102-107.

Volante, L. (2006). An alternative vision for large-scale assessment in Canada. *Journal of Teaching and Learning,* 4, 1-14.

Wandall, J. (2009). National tests in Denmark – CAT as a pedagogic tool. In F. Scheuermann & J. Björnsson (Eds), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 45-50). Ispra (VA), Italy: CRELL (Centre for Research on Lifelong Learning). Accessed at: http://crell.jrc.it/RP/reporttransition.pdf

Webb, R., & Vulliamy, G. (2006). Coming full circle: *The impact of New Labour's education policies on primary school teachers' work.* London: The Association of Teachers and Lecturers.

Wyse, D., & Torrance, H. (2009). The development and consequences of national curriculum assessment for primary education in England. *Educational Research,* 51, 213-228.

APPENDICES

Appendix A, International Assessments of Achievement in which Ireland Participated (1980-2009)

Year	Study	Areas Assessed	Population	Class Levels	Sector
1980-82	Second International Mathematics Study (SIMS)*	Mathematics	1st and 6th years	1st year, 6th year	P, PP
1989	International Assessment of Educational Progress I	Mathematics, Science	13-year-olds	6th class, 1st/2nd years	P, PP
1991	International Assessment of Educational Progress II	Mathematics, Science	9 and 13-year-olds	3rd/4th classes, 1st /2nd years	P, PP
1991	IEA Reading Literacy Study	Reading Literacy	9 and 14-year-olds	4th class, 2nd year	P, PP
1994	International Adult Literacy Survey (IALS)	Reading Literacy, Quantitative Literacy	Adults (16- to 64-year-olds)	---	Adult
1995	Third International Mathematics and Science Study (TIMSS)	Mathematics, Science	3rd/4th classes 1st/2nd years	3rd/4th classes 1st/2nd years	P, PP
2000	Programme for International Student Assessment (PISA)	Reading, Mathematical, and Scientific literacy	15-year-olds	2nd to 5th years	PP
2003	Programme for International Student Assessment (PISA)	Reading, Mathematical, and Scientific literacy, Cross-curricular Problem Solving	15-year-olds	2nd to 5th years	PP
2006	Programme for International Student Assessment (PISA)	Reading, Mathematical, and Scientific literacy	15-year-olds	2nd to 5th years	PP
2009	Programme for International Student Assessment (PISA)	Reading, Mathematical, and Scientific literacy, Computer-based Assessment of Reading	15-year-olds	2nd to 5th years	PP

* Ireland participated in the curriculum analysis component of SIMS. Achievement data were gathered only in the context of a follow-up study involving students in First year and were not analysed at international level.

Appendix B, Questionnaire to Countries

Date

Dear Colleague,

The Educational Research Centre, on behalf of the Irish Department of Education and Science and the National Council for Curriculum and Assessment, is currently conducting research into the use of standardised tests in lower secondary schools in a number of countries. We are asking for your help with this task by completing this questionnaire, which enquires about the use of standardised tests in your country.

Standardised tests of ability/aptitude and achievement, comprised of multiple-choice and sometimes open-response items, are a feature of education in many countries. However, there is considerable variation from country to country in the conditions under which tests are administered, the purposes of testing, and the ways in which test results are used.

Our interest in this questionnaire is in obtaining information about standardised testing in grades 7, 8, and 9, which in many countries constitute the lower grades of secondary education (i.e., ISCED 2) and in others are the final grades of basic education. The age range of children in these grades is typically 12 to15 years.

Furthermore, our focus is on the use of standardised tests in classrooms by teachers. In some countries, these tests may also be part of a national assessment.

We are not interested in tests administered by psychologists or counsellors for the purpose of assessing special educational needs or student guidance. Rather, our focus is on tests administered to provide information for such purposes as supporting teacher planning and informing students and parents of students' scholastic progress.

These tests may or may not be required or recommended by a national or state educational authority.

We would be most grateful if you could complete this questionnaire and return it to grainne.moran@erc.ie before September 30th, 2009.

Please feel free to contact Gráinne Moran (e-mail: grainne.moran@erc.ie ; tel: +353 1 806 5203) or Gerry Shiel (e-mail: gerry.shiel@erc.ie ; tel: +353 1 806 5227) if you have any queries about the content of the questionnaire. If you wish to return the questionnaire by ordinary mail, please send it to:

Gráinne Moran,
Educational Research Centre,
St. Patrick's College,
Dublin 9,
Ireland.

Thank you for taking the time to respond to this questionnaire.

Thinking about Standardised Tests that are administered and used by teachers of students in Grades 7-9...

General

1 At which grade level(s) are standardised test(s) administered?

Grade 7 ❒$_1$ Grade 8 ❒$_1$ Grade 9 ❒$_1$

2 What abilities/curricular areas are assessed by the standardised tests?

	Grade 7	Grade 8	Grade 9
Aptitude (e.g., Reasoning)	❒$_1$	❒$_1$	❒$_1$
Language of Instruction	❒$_1$	❒$_1$	❒$_1$
A foreign language	❒$_1$	❒$_1$	❒$_1$
Mathematics	❒$_1$	❒$_1$	❒$_1$
Science	❒$_1$	❒$_1$	❒$_1$
Technology (ICTs)	❒$_1$	❒$_1$	❒$_1$
Cross-curricular Problem Solving	❒$_1$	❒$_1$	❒$_1$
Learning Strategies/skills	❒$_1$	❒$_1$	❒$_1$
Ability to work in groups	❒$_1$	❒$_1$	❒$_1$
Other (1):	❒$_1$	❒$_1$	❒$_1$
Other (2):	❒$_1$	❒$_1$	❒$_1$

If an aptitude test is administered, please list the aptitudes that are tested: _____

3 If a test of a particular aptitude or curriculum area is administered at more than one grade level . . .

a) Are separate (different) tests administered at each grade level?

❒$_1$ Yes ❒$_1$No

b) If yes, are the tests linked from grade level to grade level so the progress of individual students can be tracked (e.g., with overlapping items)? ❒$_1$ Yes ❒$_1$No

If yes to 3b, please state how the tests are linked: _____

_____

_____

4 At which Grade level(s), if any, are the tests compulsory for schools? (please tick all that apply:)

Grade 7 ❒$_1$ Grade 8 ❒$_1$ Grade 9 ❒$_1$ None ❒$_1$

5 At which Grade level(s), if any, may students decline to take the test? (please tick all that apply:)

Grade 7 ❒$_1$ Grade 8 ❒$_1$ Grade 9 ❒$_1$ None ❒$_1$

6 At which Grade level(s), if any, are the tests used to certify student achievement? (please tick all that apply:)

Grade 7 ❒$_1$ Grade 8 ❒$_1$ Grade 9 ❒$_1$ None ❒$_1$

7 At what time of year are tests usually administered?

a) beginning of year ❒$_1$

b) end of year ❒$_1$

c) when teachers consider individual students to be ready ❒$_1$

d) varies from school to school ❒$_1$

e) other (please specify:)_____

8 Who decides when tests are administered?	Yes	No
a) National/State Ministry of Education	❒$_1$	❒$_2$
b) school principal	❒$_1$	❒$_2$
c) classroom teacher	❒$_1$	❒$_2$

9 Are the tests developed by:	Yes	No
a) a National/State Ministry?	❒$_1$	❒$_2$
b) an agency or contractor on behalf of the National/State Ministry?	❒$_1$	❒$_2$
c) a test development agency that produces the tests for commercial purposes (i.e., with no contract)?	❒$_1$	❒$_2$

d) other (please specify:)_____

	Yes	No
10 Is a bank of test items available to teachers to allow them to construct their own tests?	❒1	❒2
11 Do schools have a choice of tests, e.g., mathematics tests developed by different agencies?	❒1	❒2
12 If yes to 11, are the tests equated?	❒1	❒2
13 Are standardised tests that have been developed in other countries (e.g., U.S.A.) in use in schools?	❒1	❒2
14 If yes to 13: a) have the tests been standardised for local use?	❒1	❒2

b) what aptitudes/achievements do the tests measure? _____

_____

_____

_____

15 Are parallel forms of all/some tests available?	❒1	❒2

16 Who determines the purpose(s) of the test(s)?

a) National/State Ministry	❒1	❒2
b) schools	❒1	❒2
c) individual teachers	❒1	❒2

d) other (please specify:)_____

17 Tests can be used for norm-referencing (comparing the performance of a student with that of other students), criterion-referencing (identifying a student's mastery of curriculum content and processes), or diagnosis (identifying a student's learning difficulties).

Please indicate the main interpretation attached to the standardised tests at each grade level (please tick one box in each row:)

Grade 7: Norm-referenced ❒$_1$ Criterion-referenced ❒$_1$ Diagnostic ❒$_1$
Grade 8: Norm-referenced ❒$_1$ Criterion-referenced ❒$_1$ Diagnostic ❒$_1$
Grade 9: Norm-referenced ❒$_1$ Criterion-referenced ❒$_1$ Diagnostic ❒$_1$

18 Are test norms:	Yes	No
a) available for the beginning of the school year?	❒$_1$	❒$_2$
b) available for the end of the school year?	❒$_1$	❒$_2$
c) available but the time of year is unspecified (e.g., age-based)?	❒$_1$	❒$_2$

19 Has a relationship been established between the standardised tests used at lower secondary level and the following:

a) standardised tests at primary level _____

_____

_____

_____

_____

b) other examinations to certify student achievement _____

_____

_____

_____

_____

c) national assessments of student achievement _____

_____

_____

_____

_____

c) international surveys of student achievement (e.g., PISA, TIMSS)

_____

_____

_____

_____

Test Administration

20 How are the tests delivered?

	Yes	No
a) paper-and-pencil items only	❒$_1$	
b) computer-based items only	❒$_1$	
c) combination of paper-and-pencil and computer-based items	❒$_1$	

21 Who administers the tests?

a) students' own teachers	❒$_1$	❒$_2$
b) other teachers in the school	❒$_1$	❒$_2$
c) teachers from other schools	❒$_1$	❒$_2$

d) other (please specify:)_____

22 Is administration of the tests monitored by an external agency (e.g., National/State Ministry of Education)? ❒$_1$ ❒$_2$

23 If yes to 22, what form does monitoring take? _____

_____

_____

_____

24 Which categories of student (if any) are excluded from testing on the basis of having a special educational need? _____

_____

_____

_____

25 Please describe any accommodations that are made for students whose home language is different from the national language/ language of instruction: _____

_____

_____

_____

26 Is the testing in schools supported financially:	Yes	No
a) by a central authority (e.g., Ministry)?	❒1	❒2
b) from schools' own resources?	❒1	❒2
c) by students or their parents?	❒1	❒2

Scoring

27 Are tests scored:		
a) by students' own teachers?	❒1	❒2
b) by external scorers?	❒1	❒2
c) electronically (e.g., scanner)?	❒1	❒2

28 If yes to 27c;

a) is a central scoring service available to schools? ❒1 ❒2

b) what is the cost per test scored?_____

c) who pays for it?_____

Use, Interpretation, and Dissemination

29 Are test results used (please tick one box in each row):	Yes, this is required by the State	Yes, but this is not required by the State	No
a) to allocate students to classes/ courses (e.g., higher/ honours/ advanced, ordinary, foundation)?	❒1	❒2	❒3
b) to allocate students to instructional groups within a class?	❒1	❒2	❒3
c) to diagnose student learning difficulties?	❒1	❒2	❒3
d) to identify students in need of further investigation?	❒1	❒2	❒3
e) to retain in grade/promote students?	❒1	❒2	❒3

	Results are not reported	Individual results are reported	School-level results are reported
30 In what form, if any, are test results reported to:			
a) students?	❒1	❒2	❒3
b) students?	❒1	❒2	❒3
c) parents?	❒1	❒2	❒3
d) the school board?	❒1	❒2	❒3
e) the local community?	❒1	❒2	❒3
f) external bodies/individuals (e.g., inspector)?	❒1	❒2	❒3

31 Please indicate (where applicable) who reports test results to each of the following and how they are reported (e.g., orally, printed)

	Who reports results?	How are results reported?
a) students		
b) students' teachers?		
c) parents?		
d) the school board?		
e) the local community?		
f) external bodies/individuals (e.g., inspector)?		

32 Please describe restrictions, if any, placed on the use of test results:

_____

_____

_____

_____

33 Are procedures in place (e.g., in-service courses) for teachers to support them in interpreting and using standardised test results? Yes ❒1 No ❒2

34 If yes to 33, please describe: _____

_____

_____

35 Is documentation (e.g., manuals, guidelines) available to teachers to assist them in interpreting and using standardised test results? ❒1 ❒2

36 In what ways, if any, are parents supported in interpreting standardised test results? _____

_____

_____

37 Please summarise the consequences (if any) of doing well/poorly on a standardised test for schools, teachers and students: _____

_____

_____

38 Are the results of testing presented to the public in a way that allows comparisons to be made between schools (e.g. league-tables)? ❒1 ❒2

Please provide details of any published descriptions (e.g., websites, journal articles) of the system of standardised testing in your country that you are aware of:

_____

_____

Thank you very much for completing this questionnaire.

Please return to grainne.moran@erc.ie

Appendix C, Example of Proficiency Levels

Table C1: Proficiency Levels on the PISA 2006 Reading Scale

Level & score range	What students can typically do	IRL %	IRL SE	OECD %	OECD SE
5 625.6	Locate and possibly sequence or combine multiple pieces of deeply embedded information, some of which is outside the main body of the text. Infer which information in the text is relevant to the task at hand. Deal with highly plausible and/or extensive competing information. Critically evaluate or hypothesise, drawing on specialist knowledge. Deal with concepts that are contrary to expectation and draw on a deep understanding of long or complex texts. In continuous texts, can analyse texts whose discourse structure is not obvious or clearly marked, to discern relationships of specific parts of the text to its implicit theme or intention. In non-continuous texts, can identify patterns among many pieces	11.7	0.80	8.6	0.12
625.6 4 552.9	Locate and possibly sequence multiple pieces of embedded information, each of which may need to meet multiple criteria, in a text with familiar content or form. Infer which in formation in the text is relevant to the task. Use a high level of text-based inference to understand and apply categories in an unfamiliar context, and to construe the meaning of a section of text by taking into account the text as a whole. Deal with ambiguities, ideas that are contrary to expectation and ideas that are negatively worded. Use formal or public knowledge to hypothesise about or critically evaluate a text. Show accurate understanding of long or complex texts. Follow linguistic or thematic links over several paragraphs, in order to locate embedded information or to infer psychological or metaphysical meaning.	25.1	1.04	20.7	0.17
552.9 3 480.2	Locate, and in some cases, recognize, the relationship between pieces of information, each of which may need to meet multiple criteria. Deal with prominent competing information. Integrate several parts of a text in order to identify the main idea, understand a relationship, or construe the meaning of a word or phrase. Compare, contrast or categorise taking many criteria into account. Deal with competing information. Make connections or comparisons, give explanations or evaluate a feature of text. Demonstrate a detailed understanding of the text in relation to familiar everyday knowledge, or draw on less common knowledge. Use conventions of text organisation, where present, and follow implicit or explicit logical links such as cause and effect relationships across sentences or paragraphs in order to locate, interpret or evaluate information.	30.2	0.80	27.8	0.17
480.2 2 407.5	Locate one or more pieces of information, each of which may be required to meet multiple criteria. Deal with competing information. Identify the main idea in a text. Understand relationships, form or apply simple categories, or construe meaning within a limited part of the text when the information is not prominent and low level inferences are required. Make a comparison or connections between the text and outside knowledge, or explain a feature of the text by drawing on personal experience and attitudes. Follow logical and linguistic connections within a paragraph in order to locate or interpret information; or synthesise information across texts or parts of a text to infer the author's purpose.	20.9	0.93	22.7	0.17
407.5 1 334.8	Locate one or more pieces of explicitly stated information, typically meeting a single criterion, with little or no competing information in the text. Recognise the main theme or author's purpose in a text about a familiar topic, when required information in the text is prominent. Make a simple connection between information in the text and common, everyday knowledge. Can use redundancy, paragraph headings, or common print conventions to form an impression of the main idea of the text, or to locate information stated explicitly within a short section of text.	9.0	0.84	12.7	0.15
334.8 < 1	Students below Level 1 have a less than 50% chance of correctly answering Level 1 questions. Their reading literacy skills are not assessed by PISA.	3.2	0.55	7.4	0.14

Adapted from OECD (2007b) Figure 6.7, p.292-293

Appendix D, Trends in Junior Certificate English, Mathematics and Science Results (1999-2009)

	Candidates (Number)			Grades A-C (% of Candidates)		
English	Higher	Ordinary	Fndt	Higher	Ordinary	Fndt
1999	39079	20442	2644	74.0	75.7	71.2
2000	37548	20480	2411	71.2	75.2	71.0
2001	36875	20240	2380	72.2	74.2	76.3
2002	36973	19811	2806	76.5	80.1	81.5
2003	37023	19072	2621	78.0	80.1	82.8
2004	35593	18087	2537	77.2	79.6	81.9
2005	36172	17551	2302	75.8	79.1	80.3
2006	37145	17716	2264	77.6	78.6	80.2
2007	37740	16595	2339	77.2	79.1	79.9
2008	36938	16309	2048	78.5	78.9	79.7
2009	36574	16214	2074	76.5	79.4	77.4

	Candidates (Number)			Grades A-C (% of Candidates)		
Maths	Higher	Ordinary	Fndt	Higher	Ordinary	Fndt
1999	22240	31674	7831	76.0	67.6	73.5
2000	21926	30585	7508	66.4	66.8	76.8
2001	21113	30162	7909	77.0	68.4	73.2
2002	21821	29588	7886	74.1	67.7	78.4
2003*	23734	27383	7324	79.4	71.5	76.9
2004	23006	26347	6584	73.4	75.5	85.9
2005	23388	26518	5907	75.6	73.0	82.7
2006	24204	26820	5941	78.7	77.9	83.9
2007	23804	27094	5641	75.7	73.2	79.4
2008	23634	26384	5140	79.8	76.8	83.6
2009	23592	25930	5186	77.6	74.7	80.0

* Revised syllabus tested for first time in 2003

Candidates (Number)				Grades A-C (% of Candidates)		
Science	Higher	Ordinary		Higher	Ordinary	
1999	34952	19435		72.4	62.2	
2000	33802	18996		70.2	77.1	
2001	30784	19794		76.4	83.2	
2002	32389	19703		73.1	74.6	
2003	32667	18423		76.0	77.5	
2004	29975	18842		76.0	88.5	
2005	30836	18840		74.4	75.5	
2006*	30520	14592		71.2	71.9	
2007*	34855	14892		78.2	79.1	
2008*	33566	15125		79.3	83.4	
2009*	34242	14289		77.0	79.5	

* Data for Revised Science Syllabus only

Appendix E, National Assessments, Final Examinations and International Assessments in Lower-Secondary Level Schooling in Comparison Countries

Country	Census National Assessment	Sample National Assessment	Final Examination	International Assessments
Denmark	De nationale test (National Test – full implementation in 2010)	None	Folkeskolens afgangsprove (Leaving Examination of Folkeskole) (end of Year 9)	PISA and TIMSS
Finland	None	Opimistulosten kansallinen arvointi (National evaluation of learning outcomes)	None	PISA
France	Évaluations diagnostiques (System of Diagnostic Assessment)	National Tests 1 - Cycle des évaluations bilans en fine d'école et en fine de college (Cycle of monitoring and assessment at end of lower primary and lower secondary schooling) and National Tests 2 – Évaluations bilans des competences en françcais et en mathematiques en fin d'école et en fin de college (Assessment of basic competences in French and Mathematics)	Brevet des collèges (end lower secondary, ages 14-15)	PISA
Netherlands	None*	None		PISA and TIMSS
New Zealand	None**	None	The National Certificate of Educational Achievement (linked to national qualifications framework)	PISA and TIMSS
Norway	Nasjonale prover (National Tests)	None	Eksamen (Examinations)	PISA and TIMSS
Scotland	National 5-14 Assessment Bank	Scottish Survey of Achievement (SSA)	National Qualifications (NQs), including Standard grades, National Courses and National Units.	PISA and TIMSS

Sources: EA&EA/Eurydice (2009); Questionnaire to National Education Departments (September, 2009); INCA International Review of Curriculum and Assessment (http://www.inca.org.uk/)

*School-based CITO standardised tests are administered in first and second years of lower secondary, and taken by almost all students.

**School-based assessment (Diagnostic Assessment of Individual Students' Progress) is available to schools.

Appendix F, National Assessments at Lower-Secondary Level Schooling in Comparison Countries

Country	Name of National Assessment	Objective	Uses	Subjects Tested
Denmark	De nationale test (Nation Test – full implementation in 2010) Years 2-8 (compulsory, all students)	• To monitor achievement and provide teachers with diagnostic information related to individual students • To provide feedback to schools, students and parents		Danish /reading in years 6 and 8; Mathematics in year 6; English in year 7; biology, physics/chemistry and geography in year 8; voluntary test of Danish as a second language in year 7.
Finland	Opimistulosten kansallinen arvointi (National evaluation of learning outcomes) Taken in year 9 (end of lower secondary)	• To follow up at national level how well the objectives set in the core curricula have been reached; • To monitor implementation of equality and equity policies in schools (gender, regional, social, language)	• Schools are informed of the outcomes for their own development purposes • National results are used for national development and as a basis for political decisions; • For meta-analyses e.g., on learning outcomes and their relation to different perspectives of promoting equality and equity.	Mother tongue or mathematics. In 2008-09, Swedish as a second foreign language, and Swedish as a mother tongue were assessed. Cross-curricular abilities, problem solving ability, learning strategies/skills, and ability to work in groups are also assessed.
France (1)	Évaluations diagnostiques (System of Diagnostic Assessment) Compulsory at entry to lower secondary (age 11)	• To identify levels of attainment of schools and classes (strengths and weaknesses)	• Teachers take necessary actions to help students in their learning process, taking into account the heterogeneity of classes and diversity of students' pace of learning.	French and mathematics

Appendix F (cont.), National Assessments at Lower-Secondary Level Schooling in Comparison Countries

Country	Name of National Assessment	Objective	Uses	Subjects Tested
France (2 & 3)	National Test 1 - Cycle des évaluations bilans en fine d'école et en fine de college (Cycle of monitoring and assessment at end of lower primary and lower secondary schooling) and National Test 2 – Évaluations bilans des competences en françcais et en mathematiques en fin d'école et en fin de college (Assessment of basic competences in French and Mathematics) Test 1: Representative sample of schools and students at end of compulsory education (14-15 years) Test 2: Representative sample of schools, classes and students at end of compulsory schooling.	• To monitor the education system at national level • To compile an objective report on basic competencies in French and mathematics	For regulating educational policy at national level and for acting on curricular content, the definitions of socles de competences (competence thresholds), the organisation of academic courses, the pedagogical organisations, and certain school populations	Test 1: Rotation of all subjects taught at ISCED 2 (except art and sport) on a five-year cycle; Year 1: French; Year 2: Foreign languages (English, Spanish, German); Year 3: Civic behaviour and life in society; Year 4: Life and earth sciences; physics and chemistry; Year 5: mathematics. Test 2: French and mathematics
Netherlands	None*			
New Zealand	None**			
*Netherlands	School-based CITO standardised tests	• To evaluate whether students have achieved the attainment targets of the compulsory core curriculum for lower secondary education.	Help determine the best learning pathway for students to follow, particularly those in vocational education tracks.	
**New Zealand	School-based Assessment (Diagnostic Assessment of Individual Student's Progress) Teachers use Assessment Resource Banks (ARBs), Assessment Tools for Teaching and Learning (asTTle) and National Exemplars	• To improve teaching and learning by diagnosing learning strengths and weaknesses, measuring students' progress against the defined achievement objectives, and reviewing the effectiveness of teaching programmes	Informs teachers about student learning, and development, and provides the basis of feedback for students and parents.	

Appendix F (cont.), National Assessments at Lower-Secondary Level Schooling in Comparison Countries

Country	Name of National Assessment	Objective	Uses	Subjects Tested
Norway	Nasjonale prover (National Tests) (Compulsory for all students in year 8 (age 13)	• To provide diagnostic information on students' basic skills- To provide a basis for improvement and development at school level	Intended as an instrument for improvement and development activities locally and centrally	Literacy (reading in Norwegian), mathematical literacy, and reading in English.
Scotland (1)	National 5-14 Assessment BankStudents ages 5-14; Optional, but almost all schools in the public sector use the test; very few independent schools do so. The assessments are delivered to schools via the National Assessment 5-14 website. Schools go online, select the appropriate curriculum area and level and download the assessment package. There is no choice of assessment package beyond the choice of curriculum area and level	• To confirm teachers' judgements against national standards	Provides information to parents, schools, local authorities	Mother tongue (English, Gaelic) and mathematics
Scotland (2)	Scottish Survey of Achievement (SSA) Compulsory for selected sample of schools (public and independent) and students, at end of second year of post-primary schooling.	• To provide a national overview of achievement levels		Subject varies by year; In 2009 the focus was on the literacy skills of reading and writing. SSA also gathers evidence of students' performance in core skills such as numeracy, communications, using ICT, problem solving and working with others.

Sources: EA&EA (2009); Questionnaire to National Education Departments (September, 2009); INCA International Review of Curriculum and Assessment; (http://www.inca.org.uk/)

Appendix G, Examinations at End of Lower-Secondary Level Schooling in Comparison Countries

Country	Examination at End of Lower-Secondary	Objective	Uses	Subjects Tested
Denmark	Folkeskolens afgangsprove (Leaving Examination of Folkeskole) (end of Year 9)	Document degree to which students satisfy requirements stipulated in course regulations	For certification; so significance for entry into upper secondary.	Compulsory: Danish (written and oral); mathematics (written); English (oral); physics/chemistry (oral); and one subject in humanities and one in sciences. Students can also be tested on optional subjects (i.e., German, French, needlecraft, woodwork or home economics) (the latter three tests can be taken at end of year 8 at the discretion of school head)
Finland				
France	Brevet des collèges (end lower secondary, ages 14-15)		Award of national certificate. No significance for entry into upper secondary level.	French, Mathematics and History/Geography/Citizenship, foreign language, basic use of computer and Internet.
New Zealand	The National Certificate of Educational Achievement	• Recognise the results of written examinations, along with internally assessed unit standards, in one comprehensive qualification. • Provide a wide range of learning pathways and subject choices for students, all leading to one qualification. • Deliver useful, accurate and meaningful information about student achievement to whomever needs that information.	Certifies students who leave at end of compulsory schooling.	

Appendix G, Examinations at End of Lower-Secondary Level Schooling in Comparison Countries

Country	Examination at End of Lower-Secondary	Objective	Uses	Subjects Tested
Norway	Eksamen (Examinations) (Compulsory for all students at end of Year 10)	• To assess students at end of lower secondary schooling	• For certification at end of lower secondary schooling	Either mathematics, Norwegian or Sami, or English.
Scotland	National Qualifications (NQs), including Standard grades, National Courses and National Units.	• To certify students' attainment at Secondary 3 and 4 (ages 14-16); not compulsory, but almost all students take it. • Schools use results for self-evaluation and to improve practice.	National Qualifications are intended to be attainable by all students, and are gained by external examination, together with an element of assessment carried out by the school itself, and moderated by the Scottish Qualifications Authority (SQA).	Students working towards Standard grades, for example, often take seven or eight subjects including mathematics and English. There are three levels of study for Standard grade: Credit, General and Foundation. Students usually take examinations at two levels – Credit and General or General and Foundation.

Sources: EA&EA (2009); Questionnaire to National Education Departments (September, 2009); INCA International Review of Curriculum and Assessment; (http://www.inca.org.uk/)

Notes: France: Although the Brevet is a written examination with content standardised at national level, and there are centralised procedures for administering the marking the exam, it cannot be regarded as a standarised test, given the wide variety of practices in marking and interpreting its results (EA&EA, Eurydice, 2009).

Boyd Freeman Design
www.boydfreeman.ie
boyd@boydfreeman.ie
+ 353 (0)87 2442506

ISSN 1649-3362

© NCCA 2010
National Council for Curriculum and Assessment
24, Merrion Square, Dublin 2.