Chapter 3

Student assessment

Student assessment in the Netherlands is largely the responsibility of schools and classroom teachers, supported by well-developed standardised assessment tools. The Dutch education system has rich resources and expertise in assessment design, development and administration. There has been strong attention to reaching high quality standards and investing in continuing improvements to central tests. The assessment framework relies on a balance between teacher-based and central assessments, with a recent focus on clarifying reference levels for student learning and strengthening "results-oriented work" in schools. There is an extensive system of formative and summative assessments as well as pockets of promising innovations. To exploit and scale up these innovations, it seems essential to engage a broad-based consultative process to build consensus on the education goals for future generations. It is likely that such a consensus will involve a rethinking of traditional learning goals, as well as the adoption of some of the 21st century skills as important curricular goals. As a consensus on student learning goals begins to emerge, work on a corresponding assessment strategy can begin, including further development of the assessment infrastructure, efforts to strengthen teacher professionalism in assessment and support for innovative assessment practice at the local level. In the short term, there are opportunities to leverage assessment data that is currently being generated and to critically examine current practices that may impede innovation and improvement, with a view to ensuring a balanced use of assessment as, for and of learning.

This chapter focuses on approaches to student assessment within the Dutch evaluation and assessment framework. Student assessment refers to processes in which evidence of learning is collected in a planned and systematic way in order to make a judgment about student learning (EPPI, 2002). This chapter looks at both summative assessment (assessment *of* learning) and formative assessment (assessment *for* learning) of students.

Context and features

References for student assessment

As described in Chapters 1 and 2, there is no national curriculum, but the Ministry of Education, Culture and Science sets core learning objectives that students are expected to achieve by the end of both primary and lower secondary education. For the upper cycle of secondary education, the Ministry of Education, Culture and Science has formulated learning targets, which have been translated into centrally set examination syllabi. These strongly influence the curricula taught in upper secondary education. In addition, more detailed reference levels were set for Dutch language and mathematics. The reference levels are defined for the end of primary education and the end of each of the educational programmes in secondary education. Schools have been required to implement the reference levels since 2010 and they will have to report on their students' performance in relation to these from the 2015/16 school year onwards. To facilitate this, all primary schools will participate in a pilot study during the 2014/15 school year to evaluate the coverage of the reference levels in existing end-of-primary tests.

Approaches to student assessment

Student assessment, in its different guises, plays many important roles in the Dutch education system. These roles can be organised under three main headings, which will be explored further below: monitoring/improvement, decision-making/certification, and school reporting/accountability. However, it should be noted that the purposes for assessments are not always clearly stated and that results from the same assessment are sometimes used for several purposes. New legislation developed in 2013 is likely to strongly influence student assessment policy and practice in the coming years (Box 3.1).

Box 3.1 New Dutch laws on student assessment developed in 2013

At the time of the OECD review visit in June 2013, important changes to the legislation on student assessment were being prepared. In December 2013, the First Chamber of Parliament accepted a law proposal making it mandatory for primary schools to administrate regular student monitoring systems as well as a final summative test at the end of Year 8. In implementing this policy, schools will be allowed to choose between different tests developed by the Central Institute for Test Development (Centraal Instituut voor Toetsontwikkeling, Cito) or other companies, provided that they meet central quality requirements. These laws will be implemented from the 2014/15 school year.

Source: www.eerstekamer.nl/behandeling/20140116/publicatie_wet/document3/f=/vjggcjjiv0zs.pdf.

Assessment for monitoring and improving student learning

Assessment strategies that aim to monitor and improve student learning are broadly referred to as formative assessment. The essence of formative assessment is the collection of evidence regarding the present status of each student in a class and the use of that evidence to adjust the content and/or method of instruction so as to optimise the learning trajectories of all students (Earl, 2003). Formative assessments in Dutch classrooms take many forms, ranging from the informal to the formal. In interviews with the OECD review team, teachers referred to a variety of instruments, some "home-made" and others adopted from instructional packages or purchased directly from vendors. Teachers also use other contextual information (typically only available at the classroom level) in their pedagogical decision-making.

Although formative assessment in the Netherlands is not inscribed in central education laws or regulations (OECD, 2013), it is promoted through other documents and initiatives. Since 2010, there has been considerable policy focus in the Dutch education system on using assessment results for the improvement of student learning. Under the heading "results-oriented work", the Ministry of Education, Culture and Science and the Inspectorate of Education are encouraging schools to develop systematic, goal-oriented processes to maximise student learning. The policy on results-oriented work emphasises different elements of formative assessment, including working with the national reference levels, using student monitoring systems to measure learning growth, providing feedback to students and differentiating instruction to achieve improvements in learning outcomes (Dutch Inspectorate of Education, 2010).

Several initiatives have been developed since 2010 to facilitate the implementation of results-oriented work in schools. The Dutch Inspectorate of Education (2010) published a report on results-oriented work and the General Institute for Curriculum Development has been developing further guidance regarding the implementation of a results-oriented approach. Results-oriented work also features among the priority domains of the *Schools have the Initiative* programme (Chapter 1). As part of this programme, self-evaluation instruments were developed for schools to monitor their own progress in implementing results-oriented work. In addition, two university research projects (the *Focus* project run by the University of Twente[1] and the *STREEF* project run by the Rijks Universiteit Gronigen[2]) were developed to train teachers in using assessment results to improve teaching strategies for different sub-groups of students. Since 2010, the *Focus* project has provided training to 45-65 schools each year.

The new laws on student assessment (Box 3.1), which will take effect in the 2014/15 school year, further strengthen the focus on results-oriented work. They mandate that all primary schools implement a student monitoring system to regularly assess their students' progress in a broad range of subjects (Box 3.1). While primary schools will be required to use a student monitoring system, they will retain the freedom to choose the provider and the frequency of test administration (Scheerens et al., 2012). There are three comprehensive student monitoring systems available to schools for this purpose: LVS (*Leerling Volg Systeem*)[3], ParnaSys[4], and ESIS (*Elektronisch School Informatie Systeem*)[5]. In the primary sector, virtually all schools participate in the LVS developed by the Central Institute for Test Development (*Centraal Instituut voor Toetsontwikkeling, Cito*) (Box 3.2). A range of student monitoring instruments are also offered by Cito for the first two years of secondary education (Box 3.2). In addition, the government commissioned Cito to develop an adaptive test for students with special educational needs (SEN) students.

In line with the focus on results-oriented work, there has been increased interest in measuring students' learning growth. With the introduction of a unique student number (as part of the *Basic Register for Education, BRON*), it is now possible to track students' learning trajectories wherever they attend school. At the time of the OECD review visit, a pilot study regarding student learning growth and value-added measurement was being conducted, in cooperation between the Ministry of Education, Culture and Science, the Inspectorate, educational researchers and practitioners. The final report of this study was published in January 2014 (Janssens et al., 2014). The report argues that learning growth models can be used by schools to analyse the progress of their students, as part of results-based work.

Box 3.2 Dutch student monitoring systems developed by Cito

The LVS (*Leerling Volg Systeem*) is a longitudinal student monitoring system developed by the Central Institute for Test Development (Centraal Instituut voor Toetsontwikkeling, Cito) and offered for Years 1-8 (Scheerens et al., 2012). The tests are taken once or twice a year and are completed by hand or, for some subjects (since 2003), by using computer-based modes. Tests in ordering, language and orientation in space and time are given in Years 1 and 2 only. For Years 3-5, tests are given in several aspects of Dutch language, arithmetic/mathematics, and social and emotional development. These tests are also given in Years 6-8, along with world orientation (geography, history, biology), science and technology, and English (Years 7 and 8 only). The formative/diagnostic function is accomplished through provision of interpretive materials, as well as suggestions for relevant pedagogical strategies.

In a given subject, the tests are vertically linked so that a student's progress across grades can be represented as a series of steps along a single scale. That trajectory is depicted graphically in a student report, along with normative comparisons against five different reference groups. Families of students enrolled in special education schools receive an alternative student report that indicates the level of achievement of the student and provides assistance in comparing the results with children of the same age who are attending mainstream schools. There is also a group survey report that displays the results (trajectories and level scores) for a specific group of students over a number of years.

For the first two years of secondary school, several computer-based, multiple-choice, monitoring assessments are offered by Cito for different tracks, subjects and levels. The tests are administered at the beginning of secondary school and at the end of the first and second years. Tests are offered at three levels of difficulty, corresponding to the three tracks in lower secondary education. For each track, four subjects are tested: Dutch reading comprehension, English reading comprehension, mathematics and study skills. The results are used for didactic purposes and to assist in deciding the appropriate track in upper secondary. Thus, for students in lower secondary education, the monitoring and evaluation system has a dual role of providing pedagogically useful advice, and of directly offering information relevant to the decision of which track to pursue in upper secondary.

Source: Scheerens, J., et al. (2012), OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes: Country Background Report for the Netherlands, University of Twente, Netherlands, www.oecd.org/edu/evaluationpolicy.

Assessment for decision-making and certification purposes

Assessment for summative (decision-making or certification) purposes typically occurs at the end of a learning unit, school year or educational level. Throughout primary and lower secondary education, summative assessment in the Netherlands is largely based

on teachers' professional judgement and supported by the availability of standardised assessment tools. Assessment tools chosen by teachers vary and students are given various opportunities to demonstrate their ability across a range of assessment contexts. In school visits, educators mentioned developing tests on their own or with their colleagues, using tests provided by the particular "method" they were employing, and purchasing off-the-shelf tests offered by different vendors. Some schools have introduced portfolio assessments to provide a more comprehensive view of student work. Others are employing various systems to aggregate, organise and display student data in order to make it more usable. Student results are reported to students and parents three times a year. They are represented on a scale of one-ten, with six being the pass mark.

In the final phase of primary education, schools are required to report on learning results and clarify the extent to which students have reached the core learning objectives for primary education. While schools are free to use different assessment instruments for this purpose, the vast majority of schools (85%) purchase and use the end-of-primary test developed by Cito. The results from this test provide information on the school type that would be most suitable for each student in the next phase of education. This is a key indicator employed by secondary schools, as well as by students and their parents, in the selection of a secondary school and an appropriate track within the school. At this transition point, the end-of-primary test functions as an external reference for achievement for students, parents and teachers, a validation of the teacher-based school advice and a link between learning targets for primary and secondary education.

The end-of-primary tests are in multiple-choice format, with tested subjects comprising Dutch language, arithmetic/mathematics, and study skills. World orientation is an optional subject. Cito has developed tables that, for almost all secondary tracks, indicate the advice associated with different score bands. There are three different bands: scores in the highest band lead to automatic acceptance, those in the middle band call for consultation with the primary school, and those in the lowest band require more extensive research. The outcome of this test, as well as the recommendation of the teacher (in consultation with the parents) combine to advise parents on the type of secondary school their child should attend.

New laws, that will take effect for the 2014/15 school year (Box 3.1), mandate that all primary schools administer a standardised end-of-primary test. However, schools will retain the right to choose among different examination providers (i.e. alternatives to the Cito test[6]). The draft law also stipulates that the timing of administration will be moved to later in the school year, from February to April. As a result, students will start applying for admission to secondary schools before having received the results of the end-of-primary test. This means that test results will be less useful to the secondary schools in the selection of students and that the recommendation of the school will gain more weight.

In the secondary sector, there is a collection of school-leaving examinations for each track. Typically, in each subject/track, there is a central examination developed by Cito that is aligned with the core learning objectives for that subject and follows the blueprint prepared by the College for Examinations (*College voor Examen, CVE*). Cito constructs the examination based on input from both subject matter experts and teachers. In addition, schools develop their own examination in each subject in general accordance with the syllabus for the subject. School examinations are reviewed and approved by the Inspectorate (Scheerens et al., 2012). Typically, examinations comprise both multiple-choice and open-ended questions.

All students must sit for examinations in Dutch language, a foreign language (usually English) and social sciences. In the pre-vocational streams at the secondary level, students also take examinations in practical and vocational studies. In addition, depending on the track and student interests, they may choose from among mathematics, science, modern foreign language, or other subjects. An examination is first graded by the student's own teacher following guidelines prepared by the CVE. These guidelines include detailed prescriptions including a set of rules for scoring students' responses on each individual item of an examination. A second grading is conducted by a teacher from another school. If there is a disagreement, then there is an attempt to achieve consensus. If that fails, the Inspectorate appoints a third grader whose grade is final (OECD, 2013).

School examination performance is given approximately equal weight to the central examination performance, although some interviewees indicated a shift in importance toward the central examination. Overall performance on the examinations determines the nature of the certificate earned by students and the kind of training or further education they can pursue. Overall performance is reported on a scale of one-ten, with 5 and a half (without rounding) being a passing mark. In the HAVO and VWO tracks, one mark of five in either Dutch language, mathematics or foreign language is allowed for students to pass. The CVE makes some adjustments to the scores in order to take account of variations in difficulty from year to year. Certificates awarded to students at the end of each track in secondary information provide pass/fail information and marks in each subject (OECD, 2013).

Assessment for school self-evaluation, accountability and reporting purposes

With respect to school-level accountability/monitoring, many of the assessments mentioned above are also used to provide information at the school, school board, regional and national levels. For school self-evaluation, results from the LVS by grade/subject within a school can be combined over a number of years (for greater stability) and then compared with the results at the regional or national levels. LVS results can also be aggregated for different segments of the population (e.g. students with special education needs, or students from particular ethnic groups) to enable comparisons: *(i)* between the group and the national results; *(ii)* among groups; *(iii)* within the group over time. These comparisons can be used by schools for school self-evaluation, early identification of risks and reporting to stakeholders.

The Inspectorate emphasises that the LVS is owned by the school and not part of public accountability. The LVS results are not used in the Inspectorate's annual risk analysis of schools. Only in the evaluation of schools considered at risk, inspectors will ask schools to share this data to feed into the inspectors' preparation for their school visit. In schools not at risk, inspectors on site will look into parts of the LVS results to understand learning trends in the school.

Aggregate results from the Cito end-of-primary test are currently used by the Inspectorate (and others) as an indicator of school quality. In addition, the central authorities are proposing to set national targets for performance based on these tests, as a means of encouraging schools to achieve at higher levels. The assessments administered at the end of secondary school are also aggregated to the level of the school (and higher) for monitoring and accountability purposes[7].

In addition, sample-based national assessments have also been developed specifically for education system evaluation. For further details regarding the use of assessment information for school evaluation and system evaluation, see Chapters 5 and 6.

Responsibilities for student assessment

As with other aspects of education in the Netherlands, responsibility for student assessment involves a balance between central authorities, such as the Ministry of Education and the College for Examinations, and local actors such as governing boards, schools and individual teachers. The nature of the balance depends on the particular function: formative assessments are decided and conducted primarily at the local level, even though schools and teachers are influenced by the results of standardised student monitoring systems; decisions about students' academic tracks or their attainment of a diploma at the end of secondary education are based on information drawn both from the school level and from standardised assessments; and monitoring and accountability is largely conducted by central authorities using evidence provided by both schools and national databases.

The central examinations in secondary education are prepared under the auspices of the College for Examinations, in collaboration with various stakeholders. The test development and administration is carried out by specialised organisations such as Cito. In secondary education the syllabus for a subject allows flexibility for what is tested as part of the school-based component. The extensive involvement of teachers in secondary examinations, both through the school-based component and through the grading of the open-ended questions in the central component, is likely to be a factor in existing support for the system among educators.

Strengths

In the area of student assessment, the Netherlands possesses many strengths. It has rich resources in assessment design, development and administration, with Cito being the best known testing organisation. There has been strong attention to reaching high quality standards and investing in continuing improvements to central tests. The assessment framework relies on a good balance between teacher-based and central assessments, with a recent focus on clarifying reference levels for student learning and strengthening results-oriented work in schools.

Reliable measures of student learning outcomes are available

A clear priority in assessment frameworks is the development of reliable measures of student learning outcomes. The Netherlands stands out internationally with regards to the development of standardised assessments at key stages of education. Major advantages of external standardised assessment include its high reliability and low cost of administration. Standardised central assessment also helps to clarify learning expectations for all schools and motivate teachers and students to work towards high standards and steer their teaching and learning strategies in that direction (OECD, 2013).

Standardised assessment ensures that all students are assessed on the same tasks and that their results are measured by the same standards. The results are made as objective as possible so that they are, within a year, comparable among students, regardless where they go to school. For example, the high-stakes tests developed by Cito for the end of primary and secondary school have excellent psychometric properties as they are highly reliable and carefully equated from year-to-year to preserve scale stability and interpretability over time. They make data on student learning outcomes available, providing a picture of the extent to which student learning objectives are being achieved,

and they grant the opportunity to compare individual student achievements in the tested areas.

High level of expertise in developing standardised assessments

As indicated above, Cito produces tests for summative assessments, as well as tests for monitoring and evaluation. In general, the technical characteristics of the assessments appear to be very strong and throughout the review team's interviews with stakeholder groups, there was little or no criticism of tests on that account. Cito has a large staff with expertise in educational measurement, test design and construction, and test administration. In particular, they employ individuals with specialised expertise in areas such as psychometrics, computer-adaptive testing, and the design and maintenance of the technology infrastructure for large-scale assessment. Cito supplements its in-house staff with consultants drawn from schools and tertiary institutions. Teachers are also heavily involved in the design and quality evaluation of Cito tests employed in the Netherlands. Cito's scientists are well-known contributors to the research literature on assessment and they provide support to the Ministry for the development of innovations in school assessment. Cito's international reputation is attested to by its many partnerships with Ministries of Education in other countries, other testing firms, and its successes in obtaining international contracts.

There are many other entities and vendors providing either stand-alone assessments or tests that complement other educational materials, such as textbooks. In addition, university departments and research institutes undertake studies regarding assessment practices, many of which are commissioned by the Ministry of Education to inform policy decisions.

Continuous development and innovation in student assessment

There is considerable innovation in assessment at the different levels of the system. For example, for some subjects in the vocational education track, the Ministry has introduced a framework for competency-based assessments that is aligned with the nature of proficiency in those domains.

With regard to new modes of test delivery, the monitoring tests for lower secondary education are computer based, and some components of the LVS are administered as computer adaptive tests. Adaptive testing involves tailoring the sequence of items (or sets of items) presented to the student's response pattern. Thus, a student who answers the first few questions correctly is administered more difficult questions, while the student who answers incorrectly is administered easier questions. The result is that the scores for most students are estimated more accurately than would be the case with the same, fixed examination administered to all students. This is especially the case for students at the low and high ends of the score range (Wainer, 2000).[8] Consequently, if and when the Ministry of Education decides to employ computer administration for the end-of-sector examinations, it can draw on substantial experience with this approach and can rely on some familiarity on the part of schools.

Cito invests in continuing improvements to the secondary school examinations. In the interviews with the OECD review team, Cito representatives mentioned moving to a matrix assessment design in order to broaden coverage of the syllabus, and experimenting with "self-auditing examinations" to detect inappropriate test preparation (Koretz and Beguin, 2010). In a typical matrix design, all students are administered a common block of items, and then there are one or more blocks of items that are administered only to

random sub-samples of students. These latter blocks can be used to: *(i)* obtain estimates (at a population level) of performance on a larger group of items than would be possible in a single examination, given time constraints; *(ii)* equate tests from year-to-year; *(iii)* construct or maintain a vertical scale; *(iv)* obtain information on the psychometric properties of new items; *(v)* introduce items from previous examinations (usually several years earlier) to compare the performance of different cohorts and determine if there has been "scale drift", and/or if students have become adept at answering specific kinds of questions, but have not truly developed the essential competencies the items are intended to test.[9]

At the request of some schools, Cito has also developed a digital portfolio to support more innovative approaches to documenting and evaluating student work. Although this is not currently a high priority, it may become sought after by more schools as greater emphasis is placed on developing and assessing complex competencies, such as problem-solving, creativity, critical thinking and teamwork.

The Ministry has established a unique student number that tracks students as they change schools. This facilitates longitudinal studies of student achievement, and means that fewer students are "lost" during monitoring studies. Analysis is therefore less biased and more reliable, as transient students often perform differently to more stable students.

The newly introduced reference levels provide greater clarity on expectations for student learning

The Ministry has recently published reference levels for Dutch language and mathematics to provide more clarity about instructional objectives at the primary and secondary level, which will result in some modifications to central tests and examinations. This is particularly important in view of the principle of freedom of education that is so prominent in the Netherlands. Greater clarity should help teachers across the country create syllabi at each grade level that better represent national learning goals, as well as develop assessments with improved coverage of those learning goals. Ideally, implementation of the reference levels will result not only in more equality in students' opportunity to learn, but also in better alignment of instruction across different year levels and sectors of education.

The introduction of more detailed national expectations of what should be taught and assessed in schools has, to varying degrees, been debated and tested in many countries over the last twenty five years. In all student assessment systems, there is a need for expected standards of student performance at different levels of education to have clear external reference points. While it is important to leave sufficient room for teachers' professional judgements in the classroom, it is necessary to provide clear and visible guidance concerning valued learning outcomes. Such benchmarks are intended to provide consistency and coherence, especially where there is a high degree of local autonomy in the development of curricula, teaching programmes and assessments. They can help bring about equality and fairness in educational opportunities for students across the country.

Recent policy emphasises "results-oriented work" at the classroom and school level

Recent educational policy-making in the Netherlands has had a strong focus on stimulating formative assessment and differentiated instruction through "results-oriented work" at the school level. New laws, to be implemented in the 2014/15 school year, will

require schools to use student monitoring systems for results-oriented work in schools (Box 3.1). Results-oriented work involves helping schools to more fully exploit student monitoring systems and, by analysing the information generated, to design appropriate teaching and learning strategies. Teachers are expected to explicitly define learning targets, regularly assess student performance, adapt teaching and learning to student needs, and intervene rapidly to help those who are falling behind in relation to set targets (Visscher and Ehren, 2011; Scheerens et al., 2012). Such approaches are to be stimulated and monitored by the Inspectorate.

Student monitoring systems, such as the LVS, are important tools that support results-oriented work at the school level. Schools are free to choose in which subjects and at what frequency they use the tests with their students. Such externally designed assessments can provide important signposts for teachers and students by indicating the learning goals that are expected nationally and, by producing timely data that may inform teaching strategies they can offer interesting pedagogical tools for teachers. Frequent use of high-quality monitoring systems can also stimulate teachers' own assessment expertise by providing examples of adequate test items to measure particular learning goals. Positive effects of using the results from such assessments to inform teaching may include: greater differentiation of instruction, greater collaboration among colleagues, an increased sense of efficacy and improved identification of students' learning needs (van Barneveld, 2008).

Because the tests are equated across years and vertically scaled across grades, individual student growth trajectories can be calculated, which provide a longer term perspective on student progress. A recent pilot study (Janssens et al., 2014), conducted in cooperation between the Ministry of Education, Culture and Science, the Inspectorate, educational researchers and practitioners, emphasised the importance of focusing on student progress rather than absolute performance. The report found that measuring student learning growth can make valuable contributions to the further development of results-oriented approaches in schools (Janssens et al., 2014).

The assessment framework relies on a good balance between school-based and standardised assessments

Overall, there is a reasonable balance at key decision points between the use of school-based results and central examination scores. In the transition from primary to secondary education, for example, the school's recommendation is as important as the Cito test results in determining the school and the track most suitable for the student. Although some secondary schools prefer to admit students to certain tracks only if they have sufficiently high Cito scores, a strong argument from the primary school can sway the decision, even if the threshold scores are not achieved. The importance of the school's recommendation may even increase with the proposed later administration time of the Cito test.

As noted earlier, at the end of secondary school, a subject assessment consists of both a centrally prepared examination and a school-developed examination. The fact that schools have some flexibility in deciding the content of their examinations is a mark of educators' professional autonomy, and can lead to improvements in the coverage of the syllabus. For example, in modern languages, the central examination at the end of secondary education focuses on reading skills, which allows school-developed examinations to focus on other skills such as writing, speaking and listening. Moreover, the school tests can be administered over a longer period of time, which can reduce the

pressure on students. A similar division of labour can take place in other subjects, although the actual extent of coverage and the quality of the school-based components is not generally monitored.

This flexibility, together with the integral role that teachers play in the construction and scoring of the central examinations, is likely to account for the general credibility that the secondary examination system has among teachers. This is an important strength of the system as there is less of a disjuncture between what the syllabus mandates and what teachers feel they must do to prepare their students for the examinations. This stands in contrast to some other countries where teachers can be demoralised by having to "teach to the test" in high stakes settings (Madaus et al., 2009).

Although the involvement of teachers in this aspect of the final examinations is both commendable and cost effective, it does raise questions of score comparability across schools. There is also some concern about teachers designing the school-based part of the examination to resemble the central part, which may reduce the scope of material covered in examinations, with potential adverse consequences on teaching and learning. These issues will be examined further under 'Challenges'.

Challenges

The OECD review team formed the impression that a national consensus on a set of education priorities is some time off (see Chapter 2). Given the structure and traditions of the Dutch system, it is likely to emerge slowly, informed both by the many pockets of innovation arising around the country and by political developments at the national level. The principal challenge for student assessment is how it should evolve in order to best support innovative practices and progress toward new priorities, rather than being an obstacle to their attainment. The difficulty is that assessment development tends to lag behind curricular innovations, so that student results on current assessments may not reflect new or deeper learning. Such a disjuncture can lead to premature and misleading inferences about the utility and efficacy of the innovation.

Developing assessments for the mid-21st century

Across OECD countries, there is a shared ambition to move education systems beyond traditional pedagogical approaches focussing on knowledge transmission and acquisition of basic skills and to promote a broader model of learning that includes competencies to synthesise, transform and apply learning in real-world situations. Such key competencies, or 21st century skills, typically include dimensions such as critical thinking creativity, problem-solving, communication, information and communications technology (ICT) literacy, as well as collaborative, social and citizenship skills (OECD, 2013).

Although some of these skills are already incorporated in national learning goals in the Netherlands, it is likely that a broader set of these skills and competencies will become part of the goals that are set at the national level (Chapter 2). Current paper and pencil tests with their limited item formats will not be able to appropriately assess these skills, neither for formative nor for summative purposes. Thus, there will be a need to develop the expertise and technical capacity to design, develop, deliver and evaluate more complex assessments. Many of these assessments will be integrated with instruction and take place in settings using ICT, with access to the internet, in closed micro-worlds, or in game-like environments (OECD, 2013).

Such developments towards more innovative assessment strategies are likely to raise capacity challenges at the school level. As discussed further below, a significant proportion of teachers and school leaders need to further develop their assessment expertise to fully exploit information from formative assessments or to properly interpret aggregate results (Visscher and Ehren, 2011). This gap will only be exacerbated with the introduction of more challenging learning goals and the corresponding need for new forms of assessment. Efficiently remedying this situation in light of the tradition of school autonomy and budgetary constraints will be a challenge.

Differing views on future directions for assessment

As has been noted earlier in the chapter, the Netherlands has deep resources in many facets of educational measurement that support an extensive system of formative and summative assessments, as well as a range of monitoring assessments. Moreover, there are pockets of promising innovations. Exploiting and scaling these innovations requires a national strategy that is "under development" but has not yet fully crystallised.

As might be expected, various stakeholders have different views on whether the current balance in the assessment framework is a reasonable one and what, if anything, should be done to shift that balance in a specific direction. This was illustrated during interviews as part of the OECD review visit. On the one hand, leaders from the Ministry of Education, Culture and Science asserted the need for a greater role for assessment. The introduction of laws making the use of standardised formative and summative assessment mandatory for schools is one manifestation of this view. They also noted that greater use could be made of the assessment results that are now generated on a regular schedule. On the other hand, some interviewees expressed worries regarding the apparent increase in the importance and influence of standardised examinations, particularly those administered at the end of secondary school. Some interviewees referred to this as a "testing frenzy", and argued that overreliance on test results was a symptom of a lack of trust in the professionalism of teachers. For example, the syllabi governing the central examinations, and the examinations themselves, are playing an increasingly important role in the lives of students and educators, which leads to tensions with some stakeholder groups (more on this below).

Freedom of education enables schools to undertake their own experiments with new assessments (or other aspects of education) and the Inspectorate can contribute to spreading innovations by drawing attention to schools that have developed innovative projects. But collaboration among schools tends to be local, mostly within the school board. All six schools interviewed by the OECD review team indicated that they participated to some degree in a local network, but only one school mentioned participation in a broader network[10]. During OECD interviews, there was no mention of the existence of national networks, or of plans to initiate such networks. The challenge, then, lies in how to effectively support, leverage and disseminate local innovations and best practices. Without a strategy to accomplish this, many promising ideas will remain localised or even fade away for lack of external support.

Building capacity to interpret and use assessment results for improvement

One of the key priorities of the government's policy on "results-oriented work" is to ensure that student assessment results lead to durable improvements in student learning. While student monitoring systems are widely used in Dutch classrooms, there is evidence that many teachers have difficulties in interpreting and effectively using the information

generated by such assessments to improve teaching and learning (Dutch Inspectorate of Education, 2010; 2013). A study by Visscher and Ehren (2011) found that data from the LVS was generally underutilised. Specifically, they noted the following problems: *(i)* technical snags in the storage and retrieval of data; *(ii)* a focus on underperforming students, rather than all students who need assistance; *(iii)* errors in interpreting the data; *(iv)* failure to take advantage of the longitudinal nature of the data and to use the data as a basis for collective discussion around student work.

The latest annual report of the Inspectorate notes that many teachers have difficulties in providing effective feedback to students and differentiating instruction to respond to individual student needs. The Inspectorate also finds that, despite improvements in this area, there are still major differences among schools in the degree to which teachers and schools appropriately adapt instruction to students with special education needs, a sub-population that appears to be growing (Dutch Inspectorate of Education, 2013). Shewbridge et al. (2010) noted similar challenges with respect to students who are not native Dutch speakers. It is likely that these challenges are shared by school leadership teams and that the information provided by student monitoring systems could be better utilised for pedagogical decision-making at the classroom and school level (Scheerens et al., 2012). Thus, both teachers and school leaders need more training and experience in using assessment data to inform instruction and resource allocation within the school. This opinion was expressed by nearly all the stakeholders interviewed.

Since 2010, two university research projects were developed to train teachers in using formative assessment results at the University of Twente (Project *Focus*) and the University of Groningen (Project *STREEF*). An analysis of the *STREEF* project (Doolaard, 2013) found that the training led to improved capacity of teachers regarding results-oriented work (compared to non-trained teachers) and that students benefitted from results-oriented approaches. However, the students' learning gains were not as large as expected. One of the reasons identified for the modest improvements in student learning was that many teachers had difficulties in moving from the analysis of student results to the differentiation of teaching practices. This suggests that further training should focus specifically on the interpretation and use of assessment results to adapt teaching strategies.

Strengthening formative assessment in daily classroom practice

Much important evidence on student learning needs can be obtained by teachers through ongoing formative assessment strategies in daily classroom interactions. Classroom-based formative assessment is essentially a pedagogical approach consisting of frequent, interactive checks of student understanding to identify learning needs, provide feedback to students and adapt teaching strategies (OECD, 2005). While medium- and long-term formative assessments (such as the regular use of LVS) are important for identifying areas of need and developing broad teaching strategies, it is short-cycle formative assessment happening in daily classroom interactions, which has the most direct and measurable impact on student achievement (Wiliam, 2006; Looney, 2011).

Although there seemed to be a general recognition of the efficacy of formative assessment in the Netherlands, classroom-based assessments are generally taken "off-the-shelf", from instructional packages or purchased directly from test providers. The previous section described the strengths of student monitoring systems such as the LVS and their important role in results-oriented work in schools. At the same time, research

from different countries indicates that externally developed tests often do not fully address the information sought by the teacher to identify the learning difficulties of students. They typically can only cover a sub-set of the intended curriculum goals. Also, the data obtained from such assessments is typically not at the level of detail needed to diagnose individual student needs and identify strategies for future instruction (McGehee and Griffith, 2001; Rupp and Leseaux, 2006).

While regular monitoring tests can provide important clues about learning areas that need attention, other assessment strategies may be necessary to identify the causes of poor performance and develop an appropriate instructional intervention (OECD, 2013). In the Dutch context, an analysis of the *Focus* project (see above) found that information from the LVS provided useful feedback to teachers and students, but that additional diagnostic information was necessary to respond to individual student learning needs (Faber et al., 2013). This suggests that teachers could benefit from professional development helping them to select and design adequate diagnostic assessments to respond to the specific learning needs of their students. Getting involved in creating targeted assessments, and evaluating students' responses, can also be a powerful tool for teachers' professional development (Wiliam et al., 2004).

Concerns about the reliability of examination results

In secondary education, the flexibility in formats for the school-based examination component can result in better curriculum coverage, but at the same time, guidelines issued by the Inspectorate introduce other issues. In particular, the Inspectorate requires that school-based examination results be calibrated so that there is, at most, a small average difference (in the aggregate) between those scores and the scores on the central examinations. A large gap is noted as a problem by the Inspectorate. At least one function of this policy is to maintain a rough comparability in school examination results across the country. This is valuable given the flexibility schools and teachers have in developing and scoring these examinations. However, there may be genuine differences in the "objective" quality of students' performances on the two types of examinations that are virtually eliminated by the calibration.

A striking example of this problem can occur with modern languages where students' average productive language skills may differ systematically (higher or lower) from their average reading skills. Thus, schools with similar distributions of reading scores will be forced to have similar total scores, irrespective of meaningful differences in the distributions of scores on other language competencies. In the case of Dutch language, the introduction of reference levels is presumably meant to lead to criteria-referenced scoring, which may clash with the pseudo-norm-referenced scoring imposed by the calibration requirement. Specifically, in evaluating student work in the school-based examination, teachers should be guided by the reference levels that describe the expectations corresponding to different score levels. Thus, the average quality of, for example, the written work produced by a class may correspond to a particular reference level, based on a reasonable interpretation by the teacher of the meaning of the reference levels. However, if that average differs substantially from the class average on the central examination, then the class grades will need to be adjusted to meet the guidelines of the Inspectorate. One obvious consequence is that the adjusted grades will no longer accurately reflect the quality of the work submitted by the class and so interpretations based on these grades will be misleading to some degree.

A more general concern is related to the reliability of teachers' marks. Reliability refers to the consistency of the marks provided by independent graders for a set of papers. The literature is clear that achieving high levels of reliability when evaluating complex student work (e.g. essays) requires substantial training of raters, as well as ongoing monitoring and moderation of the scoring process (Gwet, 2012). In the Netherlands, although scoring guidelines for the secondary school examinations are centrally provided, training of raters (teachers) is done at the school level. Thus, it is important to continually review the reliability of teachers' grading in settings such as end-of-secondary examinations. Since these assessments carry high stakes for students, low reliability of teacher grading would pose challenges to the fairness of assessment.

Standardised assessments strongly influence the teaching programme

Reaching some degree of a consensus on national education goals will not be achieved in the short term. In the interim, education goals are being set with reference to existing assessments. One example is the drive to increase the proportions of students who reach the highest levels in examinations of Dutch language and mathematics. Although this approach certainly has some merit, it should be borne in mind that, in general, the examinations do not fully represent the target content, so that excessive pressure can lead to unintended consequences, such as an unwanted narrowing of the curriculum.

The previous section described the strengths of standardised assessments in the Netherlands, which provide reliable information on student learning in relation to key national learning objectives. At the same time, there is a risk that the high visibility of standardised assessment might lead to distortions in the education process. In the primary sector, the Cito tests use only multiple choice items, with all the accompanying constraints on what skills can be adequately measured. Because of the role of the test results in determining placement in the secondary sector, there is pressure on both students and teachers to concentrate on the specific content and format of the tests, which leads to a narrowing of the delivered curriculum. Excessive time spent on test preparation reduces the amount of content to which students are exposed.

Similar issues pertain to the assessments administered at the end of secondary education where, arguably, the pressures on students and teachers are even greater than in primary education. Although schools have substantial autonomy in developing the school-based part of the examination, teachers interviewed by the OECD review team indicated that they tended to align their own assessments to the formats used in the central examination. Ironically, this phenomenon may be an unintended consequence of the longstanding principle of freedom of education. As there is no national curriculum, the syllabi of the central examinations administered at the end of the primary and secondary education constitute a *de facto* national curriculum and, thus, ensure some degree of comparability across schools. The syllabi, along with features of the central examinations, therefore play a strong role in shaping instruction; particularly in the years when examinations are administered.

Concerns exist that assessment results are increasingly being put to multiple uses. For example, student scores on the tests given at the end of primary and secondary school were originally intended to identify the level of proficiency attained by the student to enter secondary education. However, these scores, aggregated to the school level, are also used as a key indicator of school quality, which increases the stakes associated with these scores for schools. This heightens the pressure on teachers to improve test performance

through a focus on the specific item formats employed by the test, and other strategies that may lack educational value. For example, some schools may adopt strategic approaches to boost the schools' test results, such as making large groups of students repeat the year before the examination year (Dutch Inspectorate of Education, 2013). Moreover, to the extent that these tests suffer from construct-irrelevant variance and/or construct underrepresentation[11] (Messick, 1989), their value as monitors of school quality is compromised. That is, there is a risk that the high stakes associated with the test lead to a narrowing of the delivered curriculum and, possibly, a shift of emphasis from other valued outcomes of schooling. The new policy of making the end-of-primary assessments mandatory makes their use for school accountability even more attractive and reinforces the risks of the test being put to multiple uses (more on this in Chapter 5).

Policy recommendations

Student assessment is a critical component of the Dutch evaluation and assessment framework. The previous sections have outlined the many strengths of the Dutch approach and discussed a number of challenges related to student assessment in the Netherlands. Building on the identified strengths, this section suggests potential directions for further policy development in building a coherent framework for student assessment:

- Build consensus on key learning goals for the 21st century.

- Consider developing learning progressions to complement curriculum goals.

- Develop an assessment strategy corresponding to agreed education goals.

- Further develop the assessment infrastructure.

- Strengthen teacher professionalism for effective development and use of assessment.

- Support innovative assessment practices at the local level.

- Balance the use of assessment for improvement and accountability.

- Critically examine unintended consequences of assessment.

Build consensus on key learning goals for the 21st century

A simple model for education can be represented as a triangle with the vertices representing curriculum, instruction and assessment. At the centre of the triangle is "student learning", with links to each vertex. A healthy system is one in which there is a both a good balance among the vertices, and meaningful coherence among the activities undertaken under the auspices of each vertex (Figure 3.1). However, when assessment results are used for accountability purposes, they can assume greater importance than originally intended. This is especially the case if curriculum and instruction are not properly grounded and well supported.

Figure 3.1 The "education triangle"
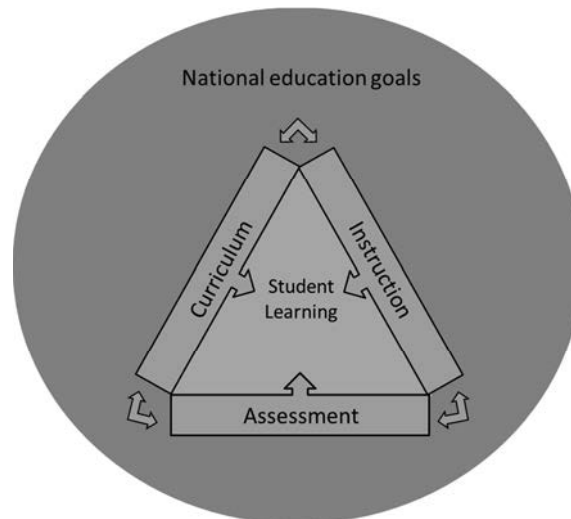
Source: Authors' own work

If student assessment is to be a tool for improving learning, rather than the driver of education in the Netherlands, it is critical that efforts are made to achieve a national consensus on the education goals for future generations. Such a consensus can then inform the work to be done at each vertex of the "education triangle". It is likely that such a consensus will involve a rethinking of the content and attainment standards of the traditional academic competencies, as well as the adoption of some of the 21st century skills as important curricular goals.

There are a number of different documented approaches to describing 21st century skills (e.g. Rychen and Salganik, 2001; 2003; Partnership for 21st Century Skills, [n.d.]; National Research Council, 2012; Kyllonen, 2012; Kyllonen, forthcoming), for a summary see Annex 3.1 and OECD (2013). What they share is an appreciation of how important it is for students to develop the skills and dispositions that will enable them to both lead economically productive lives in a global economy, and to function as engaged citizens in a democracy. In the United States, for example, new standards for mathematics and English/language arts (Common Core State Standards, n.d.) have been adopted by almost all states, and new standards for science have just been published (Next Generation Science Standards, n.d.). One aspect of all three sets of standards is that they have a strong developmental focus; that is, they are informed by, and clearly reflect, decades of research on how students learn and how to appropriately characterise expertise at different age levels. This developmental focus will have important and, hopefully, beneficial implications for assessment.[12]

In the Netherlands, despite the challenges outlined in Chapter 2, there is continuing interest in the Ministry of Education, Culture and Science to encourage schools to focus on 21st century skills. This broader view of student development is also promoted by the Education Council. With this level of interest, it may be a good time to revisit this domain in a systematic fashion, with due regard to the perspectives of the different stakeholders. A policy proposal to move forward could: *(i)* argue for deeper learning in traditional academic subjects; *(ii)* highlight the importance of key skills and dispositions that transcend those subjects; *(iii)* call for all educators to have sufficient opportunity to

develop the professional competencies needed to support student learning; *(iv)* celebrate and support innovations in pedagogy and authentic assessment; *(v)* recognise the importance of engaging all students and having students take greater responsibility for their own learning.

Such a proposal, drawing on the emerging international consensus on 21st century skills, could serve as a starting point for a conversation that would lead the Netherlands to a next-generation education system. It also recognises not only the key role of educators, but also the necessity of providing the sustained support they require to appropriately implement an expanded curriculum and to help students attain higher levels of competence. At the same time, the innovations discussed above hold the promise of increasing student engagement and effort, increasing pedagogical effectiveness and the satisfaction that comes with teaching well.

Consider developing learning progressions to complement curriculum goals

The OECD Reviews of Evaluation and Assessment in Education found that specific learning progressions, describing the way in which students typically move through learning in each subject area, can further help to clarify national curriculum goals (OECD, 2013). Research-based learning progressions can provide a picture from beginning learning to expertise and help provide teachers, parents and other stakeholders with concrete images of what to expect in student learning, with direct links to the final learning objectives and reference levels. Such learning progressions can provide a clear conceptual basis for a coherent assessment framework, along with assessment tools that are aligned to different stages in the progressions.

Teachers can use these learning progressions as roadmaps to identify the set of skills and bodies of enabling knowledge that students must master en route to becoming competent in the more complex and multifaceted learning objectives defined for the end of primary and secondary education. The Ministry of Education, Culture and Science together with stakeholder groups could also facilitate the development of assessment criteria for rating different aspects of performance and exemplars illustrating student performance at different levels of proficiency. These can be used to define what constitutes adequate, good and excellent work and support professionals in clarifying quality definitions and making accurate judgements about student performance and progress in different curriculum areas.

Such materials can be promoted as voluntary resources that teachers use as signposts in their assessment. They can help raise aspirations and communicate a focus on excellence and continuous improvement. In line with the government's focus on teacher professionalism, such guidance could help teachers design their instructional plans and classroom assessment strategies in alignment with national objectives and progressions. Teachers should also be encouraged to share and co-construct intermediate learning goals and assessment criteria with students so that they understand different levels of work quality. Such common work on goals and criteria can promote both student learning and reflective teaching practice (Andrade, 2005; Jonsson and Svingby, 2007).

Develop an assessment strategy corresponding to agreed education goals

As a consensus on learning goals begins to emerge, work on a corresponding assessment strategy can begin. Given the novelty of 21st century skills for most teachers, formative assessment should be the primary focus, as it can contribute directly to improved learning. Ideally, a coherent set of formative assessments (across grades, within

a sector), along with the corresponding scoring rubrics and exemplars of student work, will help to provide illustrations of both the learning goals and the expectations for student performance. As noted by the OECD (2013), such support is essential to the constructive inclusion of broader competencies into the curriculum. Wherever possible, these skills should be incorporated into the existing curriculum so that they enhance the instruction and assessment of the core academic competencies and are not seen simply as an "add-on" to an already crowded schedule[13]. To help this, and where it is feasible and appropriate, teachers should use available technology for both teaching and assessment.

As curriculum changes are introduced, particularly those involving new and/or more advanced skills, they should be accompanied by an assessment strategy that is appropriately matched to the target competencies. An example is offered by the Technasium[14], a grass-roots effort to strengthen the teaching of science through a more engaging, contextualised curriculum that incorporates problem-based learning. Allocating resources to provide assessment support would not only benefit the initiative, but also offer experience in how to develop an assessment system that incorporates more complex challenges. More generally, novel formats for student work may require new psychometric models and analysis strategies.

Exploratory projects can build the expertise needed when versions of these assessments are eventually brought to scale. In the Dutch context, one possibility is to begin by building measures of the "non-cognitive skills" that are relevant to tracking the progress of students with special educational needs (SEN) for whom the development of social skills, for example, may be as important as the development of traditional academic skills. Since most teachers will be unfamiliar with the frameworks for 21st century skills and the ways in which specific skills are defined and operationalised at different levels, they will need extensive professional development with respect to didactics (more on this below).

As mentioned in Chapter 2, the development of a broader set of learning goals, and the accompanying assessments, will have to involve multiple stakeholders. Given the concerns of many stakeholders regarding what they view as the increasing importance of central examinations, too early introduction of these skills into mandated summative assessments could trigger a backlash that would limit progress. Expanding the end-of-sector assessment frameworks to include more of the 21st century skills is likely to be problematic given the time and cost constraints under which they operate, and encouraging their adoption in the school examinations would run counter to the traditional autonomy accorded to schools. Thus, it seems more feasible for the Ministry to provide funding (perhaps through a revised *Schools have the Initiative* programme) for those schools and networks of schools already focussing on the teaching and assessment of 21st century skills to accelerate their work and, equally importantly, to provide regional and national forums where these schools can showcase their efforts to a broader audience. This, in turn, can lead to a national conversation on how traditional learning goals should be augmented to meet the challenges of the 21st century, and what systemic changes must be made to support teaching and learning of this broader set of competencies.

As suggested in Chapter 2, the Ministry of Education could start by commissioning a review of research and of strategies adopted by other education systems. A number of countries (e.g. Singapore, Costa Rica) are moving towards the introduction of problem-solving (in different forms) into national goals and assessment strategies in a systematic fashion, with strong central ministry support. The experiences of Finland and Australia

would also be especially valuable. As with the Netherlands, Finland's education system is highly decentralised and schools have substantial autonomy with respect to pedagogy and student assessment. Australia has a federal system that maintains somewhat greater central control, but has invested considerable resources in the assessment of these skills. Both countries have made considerable progress in raising the focus on 21st century skills (Adamson and Darling-Hammond, 2013).

Further develop the assessment infrastructure

Despite the absence of a new set of education goals, the Netherlands can be proactive in preparing for an uncertain future by building an assessment-related infrastructure that would prove useful in a variety of scenarios. Experimentation with the assembly, delivery and data management of new forms of technology-based assessment will provide the hands-on experience necessary if and when such assessments are introduced at the national level. One option to enhance the assessment technology infrastructure would be to further work on developing data information systems for schools, taking full advantage of the introduction of the unique student number and the requirement for all schools to implement student monitoring systems. Beyond providing real-time access to student data on attendance, enrolment, marks and schedules, such tools could serve as instructional management systems and provide planning tools, instructional materials and national-level performance standards. For example, some education jurisdictions in Australia have developed sophisticated data information systems, such as the School Measurement, Assessment and Reporting Toolkit (SMART) developed by New South Wales (Box 3.3).

Box 3.3 New South Wales, Australia: The School Measurement, Assessment and Reporting Toolkit

SMART provides information on national standardised assessment and state examinations. This information, together with information from school-based assessment activities provides a wealth of objective diagnostic information to which teachers can respond. The SMART package allows educators to identify areas for improvement as well as strengths in student performance. SMART also provides support through specific teaching strategies designed to improve student outcomes. This initiative has the potential to assist teachers in the instruction of their students, provide quick feedback to school agents, serve as a platform to post relevant instructional material to support teachers and improve knowledge management, operate as a network to connect teachers and schools with similar concerns, and create a better data infrastructure for educational research.

Source: Santiago, P., et al. (2011), *OECD Reviews of Evaluation and Assessment in Education: Australia 2011*, OECD Reviews of Evaluation and Assessment in Education, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264116672-en.

It would also be helpful to initiate work on building expert systems to automatically evaluate some types of student work. Research literature shows that expert systems have been developed to grade essays, short answer questions in various disciplines, mathematical expressions, problem-solving strategies and graphical responses of various types (Williamson et al., 2006).[15] As technology-based assessment becomes more ubiquitous, the capacity to score student work automatically reduces overall costs and shortens the time lag between administration and score reporting. On a deeper level,

designing assessments to facilitate automated scoring can exert a useful discipline that increases comparability and validity (Bejar and Braun, 1994).

In many ways, the Netherlands is in a good position regarding technology-based assessment infrastructure, as most schools already have the capacity to administer tests by computer. Going further, the next steps could involve reflecting about a national strategy to expand and upgrade the infrastructure to support high-speed internet access (see Box 3.4 for an example from the United States). The strategy will have to take into account the rapid development of the internet backbone, as well as the proliferation of devices such as tablet computers. In addition, it will have to establish priorities (e.g. between the primary and secondary sectors) and determine the funding mechanism (e.g. cost-sharing between the national government and local boards). Such a strategy should be led by the Ministry in conjunction with key stakeholders, such as the Primary and Secondary Education Councils. Input from the Ministry of Finance might also be useful. Because of the technical and financial complexities, it would be advisable to have a set of white papers prepared in advance. These papers should describe the different purposes that the next-generation infrastructure would serve, the current state of the infrastructure, the various options for development with their advantages and disadvantages (with respect to cost and timing).

Box 3.4 United States: The National Education Technology Plan

The United States Department of Education recently released a document titled National Education Technology Plan (NETP) that describes goals in five areas: Learning, teaching, assessment, infrastructure and productivity. The plan focuses on how the thoughtful use of technology can contribute to advances in all areas. It builds on earlier initiatives including: *(i)* The E-Rate programme (www2.ed.gov/about/offices/list/oii/nonpublic/erate.html) that provides eligible schools and libraries with discounted telecommunications charges that enhance internet access and *(ii)* the State Educational Technology Grants programme (www.ed.gov/edblogs/technology/grants/) that provides funding for various initiatives related to the development and implementation of education technology. Currently, one of the main goals is the provision of technology-related services to support students with disabilities. Another grant programme (http://nces.ed.gov/programs/slds/) provides funding for statewide longitudinal data systems that helps states build and expand their data systems to collect, organize and analyse longitudinal student records for policy-related analyses and to facilitate the construction of more sophisticated, test-based indicators for use in school and teacher evaluations.

Source: U.S. Department of Education (2010), *National Education Technology Plan 2010,* available at: www.ed.gov/technology/netp-2010.

Strengthen teacher professionalism for effective development and use of assessment

In addition to further developing the assessment infrastructure, it is equally important to continue to build assessment expertise, including the capacity to use results for improvement, among both teachers and school leaders. The OECD Reviews of Evaluation and Assessment in Education found that teacher professionalism in assessment is key to developing balanced and effective assessment frameworks (OECD, 2013).

Combining teacher-based and external assessments can help ensure maximum validity and reliability in assessment. While learning outcomes that can be readily assessed in external examinations should be covered this way, continuous teacher-based assessment can cover a broader range of complex learning outcomes (Crooks, 2004). Due to its continuous nature, teacher-based assessment allows for important achievements to be measured that are more difficult to capture in an external examination, such as extended projects, practical assignments or oral work. However, in order to reach the full potential of teacher-based assessment, it is important for policy makers and stakeholders to adopt a strategic approach to teacher learning in assessment and invest in professional development opportunities. To be able to assess students' progress in developing complex competencies, it is important that teachers learn to select and/or develop a variety of assessment approaches and understand different aspects of validity, including what different assessments can and cannot reveal about student learning (OECD, 2013).

To this end, assessment capacity, including the capacity to use results for improvement, should be reflected in teacher standards and be addressed in a coherent way across teacher preparation programmes and publicly funded professional development programmes. Eventually, assessment-related competencies should become part of the teacher registration system and teacher appraisal approaches (Chapter 4). The human capital development agenda sketched out below will also require professional development of teacher educators and of providers of in-service teacher training. Fortunately, teacher education programmes can draw on the rapidly expanding resources available internationally. Again, inducements and support from the Ministry will be essential in this regard. The above analysis points to three priority areas for further capacity development: *(i)* classroom-based formative assessment; *(ii)* interpretation and use of assessment results for improvement; *(iii)* reliable summative assessment and marking of examinations.

Capacity for classroom-based formative assessment

The current focus in the Netherlands on using regular student monitoring systems for results-oriented work is commendable. Such medium- and long-term formative uses of results are important for identifying areas for further improvement, developing broad teaching strategies to address needs identified within the student cohort, planning, allocation of resources, and so on. It can also feed into the school-wide coordination of pedagogical support and remediation for students facing learning difficulties.

While medium- and long-term formative assessment strategies are important to ensure consistency of support throughout a student's learning trajectory, research indicates that short-cycle formative assessment – the daily interactions between and among students and teachers – has the most direct and measurable impact on student achievement (Looney, 2011). In short-cycle interactions, formative assessment is part of the classroom culture, and is seen as an integrated part of the teaching and learning process. Teachers systematically incorporate such formative assessment methods in their course planning – for example, in how they intend to develop classroom discussions and design activities to reveal student knowledge and understanding. These interactions encompass effective questioning to uncover student misconceptions and identify patterns in student responses, feedback on student performance and guidance on how to close learning gaps, and student engagement in self- and peer-assessment (OECD, 2013).

The active participation of students in such formative assessment processes has given rise to the term *assessment as learning*, which focuses on students reflecting on and

monitoring their own progress to inform future learning (Earl, 2003). While feedback by teachers and others provides information that can help students improve, it is the students themselves who must make sense of that information, relate it to prior knowledge and take action to close gaps in their own learning. But developing skills for self-assessment and self-regulation takes time and requires structured support by teachers in the classroom. Teachers can use classroom assessment to provide opportunities for students to engage in reflection and critical analysis of their own learning, for example by guiding students in setting learning goals and monitoring their progress towards them; working with them to develop criteria to judge progress; using exemplars and models of good practice and questioning of their own thinking and learning processes. Policy makers can support such practices by developing requirements, guidelines and support regarding learner-centred teaching and assessment (OECD, 2013).

Capacity for interpreting and using assessment results for improvement

If student assessment is to improve future learning, then assessment evidence must be acted upon in subsequent classroom practice, to provide the right levels of support and challenge to each student. In order to successfully implement differentiated instruction, teachers require relevant professional development both to make the best use of the evidence collected, and to better manage a classroom in which multiple learning activities are taking place. The investment appears to be worthwhile as the research literature documents important learning gains of students exposed to high quality differentiated instruction informed by relevant formative assessment (Black et al., 2003; Wiliam et al., 2004).

Competence in designing or selecting an assessment, interpreting the results and using the results effectively should become a key goal of teacher preparation. This can be accomplished, for example, by employing the inquiry cycle[16] as a fundamental didactic method. Undoubtedly, this will require changes in the curriculum of the teacher training programmes (White and Fredericksen, 1998). Currently, according to representatives of teacher education institutions interviewed by the OECD review team, teacher preparation does not allow sufficient time for assessment-related competencies, but there is an effort in some programmes to improve in this area, despite the severe time constraints under which they operate. Practicing teachers should be encouraged to use some of their professional development time to develop a range of assessment skills, appropriate to the subject and grade. Resources can be provided locally, regionally, or online and supported by various networks.

Capacity for reliable summative assessment and marking of examinations

Teachers play an important role in summative assessments. At the end of secondary schooling, not only do teacher-made examinations carry roughly equal weight to the central examinations, but teachers also grade the open-ended responses in the central examinations. Thus, it is important that teachers carrying out these tasks receive sufficient training so that the quality of the teacher-made examinations is high, and that the scores assigned to responses in the central examinations are reliable and valid. Ideally such training should begin during the period of teacher preparation. Exemplars of student work at different levels of proficiency and in different year levels can be used both to provide an example of the expectations at those levels/grades, and to provide opportunities for would-be teachers to develop their own expertise. However, such a redesign of teacher preparation necessarily involve systemic changes that are challenging to implement because of political and bureaucratic obstacles, as well as faculty resistance (Mandinach

and Gummer, 2013). Similar activities can be conducted for in-service teachers through specialised workshops delivered in-person or online.

During the grading of end-of-course or end-of-sector examinations, in addition to the CVE's open telephone line for teachers, consideration should be given to offering online training, and strengthening moderation processes to carry out real-time auditing of teachers' marking. Moderation involves strategies for quality assurance of assessment judgements, such as teachers cross-marking each other's assessments within a school or across schools, teachers discussing samples of student work in groups or in collaboration with experienced graders, or a competent authority or examination board externally checking school-based assessments. The objective is to reduce variations in the ways teachers assess students and set marks, in order to achieve fairness in student assessment and reporting. The Education Cooperative has made some progress in this direction and their work should be supported.

While moderation is a key strategy in validating the consistency of teacher judgement, it also involves professional discussions between teachers about the quality of authentic exemplars of student work, and as such it has the potential to provide a powerful professional learning opportunity for teachers that they can apply directly in their classroom practices. This contributes to improving teachers' professional judgements about student work and developing a shared understanding of marking criteria or standards within schools and between schools (Timperley et al., 2008). Research suggests that collaborative discussions centered on student work can lead to improved pedagogy and greater gains in student achievement (Black et al., 2002, 2003; Hargreaves and Fullan, 2012). Teachers are provided with a chance to reflect on assessment in their subject, with respect to both topics and performance criteria. From a strategic perspective, as teachers develop this expertise they will have a stronger base on which to build the more sophisticated competencies that will be required for next generation assessments. Box 3.5 offers examples on moderation procedures in place in Queensland (Australia) and New Zealand.

Box 3.5 Moderation of examinations in upper secondary education in Queensland, Australia and New Zealand

In **Queensland, Australia** the upper secondary examination system is school-determined and based, but achievement standards and scoring are externally moderated. Moderation processes for the Senior Certificate (Year 12) involve subject-based panels of expert teachers providing advice to schools on the quality of their assessment programme and their judgements of quality of student performance based on sample portfolios. The system involves follow-up with schools where panels identify issues regarding assessment and standards. There is negotiation of the final results to be recorded on the Senior Certificate (Sebba and Maxwell, 2005 in Santiago et al., 2011). Similarly, procedures adopted by educational jurisdictions and particular schools for moderating internal summative teacher judgements (so-called A-E ratings) also facilitate common understanding of year level proficiency standards and foster the development of professional learning communities that can provide crucial support for improving opportunities for student learning and building teacher capacity.

Box 3.5 Moderation of examinations in upper secondary education in Queensland, Australia and New Zealand *(continued)*

In **New Zealand**, an external moderation system is also in place to ensure the dependability of internal assessments in Years 11-13. The New Zealand Qualifications Authority (NZQA) directly checks the quality of internal assessment through a sampling approach. Schools are required to submit ten percentage points of internally assessed student work for NZQA moderation to make sure the assessment is appropriately aligned with standards. The moderation process does not affect the marks assigned to assessment samples by teachers, but is intended to provide feedback to teachers and to inform future assessment policy development at the system level.

Sources: Santiago, P., et al. (2011), OECD Reviews of Evaluation and Assessment in Education: Australia 2011, OECD Reviews of Evaluation and Assessment in Education, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264116672-en.

Nusche, D., et al. (2012),*OECD Reviews of Evaluation and Assessment in Education: New Zealand 2011*, OECD Reviews of Evaluation and Assessment in Education, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264116917-en.

Support innovative assessment practices at the local level

As noted in the "Strengths" section, the Netherlands is rich in assessment expertise. That expertise is concentrated on the design, development and analysis of current assessments. However, it is necessary to expand that expertise in directions that will facilitate the introduction of the new forms of assessment demanded by new curricular goals. One approach is to fund small-scale innovation projects (perhaps on a competitive basis) that will lead to pilot administrations of these new assessments.[17] These projects should involve partnerships between testing organisations and groups of schools where the pilots would take place. Ideally, some of the educators in these schools would participate in the research and development effort, and not simply serve as proctors during administration.

Initially, these assessments would generate data for formative purposes and could even be embedded in instruction. For example, in science classes, pedagogy centred on micro-worlds, and other kinds of simulations, could incorporate tasks that require higher order critical thinking, collaboration and effective communication, which are all 21st century skills. Such tasks would challenge both students and teachers, and yield new types of data that could serve as a focus for professional development. The design of these more complex assessments requires more sophisticated design strategies, of which Evidence Centered Design (ECD)[18] is the best known (Mislevy et al., 2006).

In addition, the Inspectorate should consider strategies for accommodating and even rewarding schools' innovative assessment practices, especially if there is reasonable evidence that they are supporting student learning. Central examinations act as an anchor in the accountability system, which is valuable in a system where educators are accorded so much freedom and flexibility. At the same time, anchors impede movement, so understanding when and how to "raise the anchor" will be important in not stifling innovation. Consequently, the Inspectorate should experiment with different ways to balance central and local results, especially in schools that meet the basic requirements.

More generally, the Inspectorate could favourably recognise schools instituting programmes of professional development that highlight improved assessment practice as a key target. Assessment practice comprises developing tests, grading student work and using the results to inform instructional practice. The Inspectorate has a particular interest in the accuracy of the grading of the end-of-sector examinations, so by recognising those schools that institute some sort of quality control mechanism in the grading process, the Inspectorate can encourage all schools to move in this direction.

The success of the initiatives described above depends, in part, on the existence of networks of educators that can facilitate collaboration and dissemination of new approaches and best practices in assessment (as well as other aspects of pedagogy and didactics). The Ministry should encourage, support and sustain such networks, working through existing organisations (e.g. the Education Council, the College for Examinations, Primary and Secondary Education Councils, the Education Co-operative). Some support could be channelled through the *Schools have the Initiative* programme. Participation would be voluntary, but incentives could be used at the start with the reasonable expectation that the value of being involved in one or more such networks would quickly become evident. Participation by some faculty in each teacher training institution should also be encouraged. Box 3.6 provides an example of a learning network in Norway.

Box 3.6 Norway: Centrally-supported networks on assessment

In Norway, the Ministry of Education and Research and the Directorate for Education and Training identified formative assessment as a priority area for education policy and professional development and launched a range of support programmes and learning networks at the regional, local and school level. For example, the *Assessment for Learning* programme (2010-14) is organised in learning networks at the local and regional level, where practitioners can exchange experiences and create spaces for common reflection on effective practice. Participating municipalities and counties employ a formative assessment contact person who assists in running the project locally. These contact persons attend *Assessment for Learning* workshops run by the national Directorate for Education and Training. The programme also provides online resources including tools and videos on how to enact effective formative assessment in the classroom.

Source: Nusche, D., et al. (2011),OECD Reviews of Evaluation and Assessment in Education: Norway 2011, OECD Reviews of Evaluation and Assessment in Education, OECD Publishing, http://dx.doi.org/10.1787/9789264117006-en.

Balance the use of assessment for improvement and accountability

Eventually, a coherent system of assessments as, for and of learning would emerge. Some assessment results will be used directly for improving teaching and learning, while other assessment data would contribute indirectly through informing accountability. In principle, these paths should complement one another, although in practice that may not be the case. As Hargreaves and Braun (2013) point out,

Although educational accountability is meant to contribute to improvement, there are often tensions and sometimes direct conflicts between the twin purposes of improvement and accountability. These are most likely to be resolved when there is collaborative involvement in data collection and analysis, collective

responsibility for improvement, and a consensus that the indicators and metrics involved in [data-driven improvement and accountability] are accurate, meaningful, fair, broad and balanced. When these conditions are absent, improvement efforts and outcomes-based accountability can work at cross-purposes, resulting in distraction from core purposes, gaming of the system and even outright corruption and cheating.

In the Netherlands, the Inspectorate must play a critical and, perhaps even a leading, role in harmonising improvement and accountability. In collaboration with the Ministry, the Primary and Secondary Education Councils and the College for Examinations, it should conduct a thorough and critical review of the indicators now employed for school evaluation. The goal should be to determine whether and how current evaluation practices may be impeding innovation, and to consider how to broaden the set of indicators to better reflect school practices that contribute to student development but are not now recognised. For example, schools should be recognised for incorporating 21st century skills into the curriculum, for the effective use of formative assessments, and for accelerating student progress towards competence standards.

Although didactics may lie outside the direct responsibility of the Inspectorate, evaluating the culture of pedagogy and collective responsibility within a school and, where appropriate, within larger school boards, would provide an impetus for school leaders to move in this direction. Assessment vendors such as Cito and others, under the leadership of the College for Examinations, could develop a framework for a coherent system of formative and summative assessments that more fully address the broader range of learning goals. With sufficient foresight and planning it should be possible to devise an assessment strategy that will both support continual improvement in education processes and assure the public that its investments in education are well spent. Box 3.7 provides examples of communication strategies in Canada and New Zealand.

Box 3.7 Defining and communicating balanced assessment strategies in Canada and New Zealand

In **Canada**, the *Principles for Fair Student Assessment Practices for Education in Canada* outline key elements for assessment practice that have served as the foundation for teacher handbooks, board polices and departments of education policy documents on assessment and test development in all Canadian jurisdictions. The Principles were developed in response to assessment practices that were judged as inappropriate for Canadian students. These principles and guidelines, intended for both assessment practitioners and policy makers, identify the issues to be taken into account in order that assessment exercises to be deemed fair and equitable. The text acts both as a set of parameters and a handbook for assessment. The first part deals with developing and choosing methods for assessment, collecting assessment information, judging and scoring student performance, summarising and interpreting results, and reporting assessment findings is directed towards practising teachers and the application of assessment modes in the classroom setting. The second part is aimed at developers of external assessments such as jurisdictional ministry/department personnel, school boards/districts, and commercial test developers. It includes sections dealing with developing and/or selecting methods for assessment, collecting and interpreting assessment information, informing students being assessed, and implementing mandated assessment programmes (for more information, see: *www2.education.ualberta.ca/educ/psych/crame/files/eng_prin.pdf*).

Box 3.7 Defining and communicating balanced assessment strategies in Canada and New Zealand *(continued)*

The **New Zealand** Ministry of Education *Position Paper on Assessment* (2010) provides a formal statement of its vision for assessment. It describes what the assessment landscape should look like if assessment is to be used effectively to promote system-wide improvement within, and across, all layers of the schooling system. The paper places assessment firmly at the heart of effective teaching and learning. The key principles highlighted and explained in the paper are: the student is at the centre; the curriculum underpins assessment; building assessment capability is crucial to achieving improvement; an assessment capable system is an accountable system; a range of evidence drawn from multiple sources potentially enables a more accurate response; effective assessment is reliant on quality interactions and relationships. To support effective assessment practice at the school level, the Ministry of Education is also currently conducting an exercise which maps existing student assessment tools. The purpose is to align some of the assessment tools to the National Standards and provide an *Assessment Resource Map* to help school professionals select the appropriate assessment tool to fit their purpose.

Sources: Fournier, G. and D. Mildon (forthcoming), OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes: Country Background Report for Canada, prepared for The Council of Ministers of Education, Canada (CMEC).

New Zealand Ministry of Education (2010), Ministry of Education Position Paper: Assessment [Schooling Sector], Wellington.

Nusche, D., et al. (2012),*OECD Reviews of Evaluation and Assessment in Education: New Zealand 2011*, OECD Reviews of Evaluation and Assessment in Education, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264116917-en.

Critically examine unintended consequences of assessment

The preceding subsections focused on preparing for the medium-term and long-term future. However, there are opportunities to leverage assessment data that is currently being generated and to critically examine current practices that may impede innovation and improvement. For example, greater effort should be devoted to examining the dynamic responses of the education system to different forms and uses of assessment; that is, "What are the intended and unintended consequences of assessment in the different sectors"? A better understanding of these consequences, and particularly how they vary with context, can help to shape assessment policy for the future (De Wolf and Janssens, 2007).

This brings to the fore the problem of "evidential asymmetry" (Hargreaves and Braun, 2013). This term refers to the fact that the intended consequences of a policy are likely to be related to indicators that are generally collected as a normal output of the system (the obvious examples are test scores and graduation rates). On the other hand, unintended (usually negative) consequences are likely to be related to indicators that are not generally collected (examples are changes to the curriculum, excessive test preparation, and manipulation of student enrolment).

Documenting these consequences, therefore, involves conducting focused studies that require special funding and often take considerable time. Thus, there is a substantial time lag between the reporting of the intended and unintended consequences. Again, conducting such studies during the present regime will help to inform education policy with respect to the proposed new curricula and the accompanying assessment and accountability systems.

Notes

1 http://project-focus.gw.utwente.nl

2 http://www.project-streef.nl/

3 http://www.cito.nl/nl/onderwijs/primair%20onderwijs/cito_volgsysteem_po.aspx

4 http://www.parnassys.nl/

5 http://www.rovict.nl/?main=esiswebbasedinfo

6 These alternatives include: Onderwijsadvies.nl, Bureau ICE, A-Vision and Boom test uitgevers.

7 The results from the end-of-sector examinations constitute one of four indicators for Quality Aspect 1 ("The outcomes are at the level that may be expected on the basis of the characteristics of the student population") of the inspection framework for secondary education (Dutch Inspectorate of Education, 2009). The four indicators for Quality Aspect 1 are: (1.1) In the lower years, students reach the educational level that may be expected; (1.2) Students need little extra time to complete the second stage of their programme; (1.3) In the national examination, students attain the marks that may be expected; (1.4) The differences between marks obtained in the school examination and those obtained in the national examination are at an acceptable level.

8 With computer-adaptive testing employing a large pool of items, nearly every student is administered a unique sequence of items. Nonetheless, using item response theory, the results can be put on a common scale just as they would if all students had taken the same test.

9 Tests incorporating such blocks of items are termed "self-auditing" since they provide evidence as to whether scores are affected by sources of construct-irrelevant variance.

10 It is noteworthy that this network, the Technasium programme, is based on a grass-roots initiative.

11 Construct-irrelevant variance refers to sources of differences in student test performance that are unrelated to what the test is attempting to measure. Construct underrepresentation refers to the gap between the intended curriculum (content standards) and what the test actually measures. Test validity is greater as construct-irrelevant variance and construct underrepresentation are reduced.

12 Two multi-state consortia have received funding from the United States Department of Education to build assessment batteries in mathematics and English/Language arts tightly aligned to the Common Core State Standards. The consortia are the Partnership for the Assessment of Readiness for College and Career (PARCC) and the Smarter Balanced Assessment Consortium (SBAC). Two other consortia are charged with the task of developing assessment batteries with especially severe or multiple disabilities.

13 One approach to such integration has been proposed by Braun (1996).

14 This is a nationally recognised programme created by a private agency and adopted in an increasing number of Dutch secondary schools. See the website of the programme here: http://www.technasium.nl.

15   This is an area of ongoing research. For essay scoring see for example http://www.measurementinc.com/sites/default/files/ASAP-PH1-PressRelease-12April2012.pdf.

16 There are many versions of an inquiry cycle: They all begin with posing a problem or issue to the student that engages her interest. The student alone, or in a group, explores the background to the problem, investigates possible solutions or explanations and presents a draft of the work. The teacher or teachers examine the work and offer feedback that students use to refine their work. This can be repeated as necessary until a final submission is made. The power of the inquiry cycle lies in the balance of responsibility between student and teacher, the opportunity for teachers to intervene constructively in the process of learning through repeated examination of student work, and how reflecting on these experiences provides teachers with a basis for improving their practice – particularly if this is done in the context of a collaborative community of professionals.

17 Presumably this could be accomplished in the same manner as the current pilot projects on the use of value-added models.

18 ECD is based on the simple premise that the best way to obtain the desired information from a test is to design it from the outset with that focus. Specifically, it begins with a clear articulation of the claims and inferences (about student proficiencies) that are to be made. Those are derived both from the nature of the subject matter and the uses to be made of the results. One then asks, "What sort of evidence is required to make those claims and inferences in a defensible manner"? The succeeding question is, "What sorts of items, tasks, or challenges are needed to generate that evidence"? With the (tentative) answers to those questions in hand, it is possible to draft a preliminary test design that also takes into account the constraints of time and cost. In almost all cases, various compromises are made to achieve a feasible design. Nonetheless, with a clear idea of what evidence may be lost as a result of certain decisions, the final design is usually superior to one built in the traditional manner.

References

Adamson, F. and L. Darling-Hammond, (2013), *Policy Pathways for 21ˢᵗ Century Skills,* http://atc21s.org/wp-content/uploads/2013/02/ATC21S_Policy_Paper-20120730.pdf.

Andrade, H. (2005), "Teaching with Rubrics: The Good, the Bad, and the Ugly", *College Teaching,* Vol. 53, No. 1, pp. 27-30.

Bejar, I.I. and H.I. Braun, (1994), "On the synergy between assessment and instruction: Early lessons from computer-   based simulations", *Machine Mediated Learning: An International Journal*, *4*(1), 5-25.

Black, P.J., et al. (2002), *Working inside the black box: Assessment for learning in the classroom,* King's College London School of Education, London.

Black, P.J., et al. (2003), *Assessment for learning: Putting it into practice*, Open University Press, Buckingham, United Kingdom.

Braun, H.I. (1996), "A school-based strategy for achieving and assessing work-readiness skills", in L. B. Resnick and J. G. Wirt (eds.), *Linking School and Work - Roles for Standards and Assessment*, pp. 223-241, Jossey-Bass, San Francisco, California.

Common Core State Standards Initiative (CCSS) (n.d.), www.corestandards.org.

Crooks, T.J. (2004), "Tensions between assessment for learning and assessment for qualifications", paper presented at the Third Conference of the Association of Commonwealth Examinations and Accreditation Bodies (ACEAB), Fiji, March 2004.

De Wolf, I.F. and F.J.G. Janssens (2007), "Effects and side-effects of inspections and accountability in education: An overview of empirical studies", *Oxford Review of Education*, pp. 1-18, Routledge, London.

Doolard, S. (2013), *Effecten van het trainings- en begeleidingstraject 'Streef: Gebruik maken van opbrengsten', Rapportage BOPO-programmalijn Onderwijskwaliteit PO*, GION, Gronings Instituut voor Onderzoek van het Onderwijs, Groningen.

Dutch Inspectorate of Education (2009), *Risk-Based Inspections as of 2009: Primary and Secondary Education*, Dutch Inspectorate of Education, Utrecht.

Dutch Inspectorate of Education (2010), *Opbrengstgericht werken in het basisonderwijs: Een onderzoek naar opbrengstgericht werken bij rekenen-wiskunde in het basisonderwijs,* Dutch Inspectorate of Education, Utrecht.

Dutch Inspectorate of Education (2013), The State of Education in the Netherlands: Highlights of the 2011/2012 Education Report, Dutch Inspectorate of Education, Utrecht.

Earl, L. (2003). Assessments as learning: Using classroom assessment to maximize student learning. Thousand Oaks, Corwin Press, California.

EPPI (2002), *A Systematic Review of the Impact of Summative Assessment and Tests on Students' Motivation for Learning*, Review conducted by the Assessment and Learning Synthesis Group, EPPI, London.

Faber, M., M. van Geel and A. Visscher (2013), *Digitale Leerlingvolgsystemen als basis voor Opbrengstgericht werken in het Primair Onderwijs: Een analyse van de wijze waarop scholen en besturen de mogelijkheden van digitale leerlingvolgsystemen kunnen benutten*, Universiteit Twente, Twente.

Gwet, K. L. (2012) *Handbook of Inter-rater Reliability* (3rd ed.), Advanced Analytics, LLC, Gaithersburg, Maryland.

Hargreaves, A. and H.I. Braun (2013), *Data-driven improvement and accountability*, Policy brief, National Education Policy Center, Boulder, Colorado.

Hargreaves, A. and M. Fullan (2012), *Professional capital: Transforming teaching in every school,* Teachers College Press, New York.

Janssens, F.J.G., L. Rekers-Mombarg, E. Lacor (2014), Leerwinst en toegevoegde waarde in het primair onderwijs: Eindrapportage, available at: http://www.rijksoverheid.nl/ministeries/ocw/documenten-en_publicaties/kamerstukken/2014/03/07/aanbieding-eindrapport-pilot-leerwinst-en-toegevoegde-waarde-primair-onderwijs.html.

Jonsson, A. and G. Svingby (2007), "The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences", *Educational Research Review,* 2, pp. 130-144.

Koretz, D. and A. Beguin (2010), "Self-monitoring assessments for educational accountability systems", *Measurement: Interdisciplinary Research and Perspectives*, 8(2-3), 92-109, Routledge, London.

Kyllonen, P.C. (2012), *Measurement of 21st Century Skills within the Common Core State Standards, Educational Testing Service, Austin, Texas,* www.k12center.org/rsc/pdf/session5-kyllonen-paper-tea2012.pdf.

Kyllonen, P.C. (forthcoming), "Designing tests to measure personal attributes and noncognitive skills", in S. Lane, M. Raymond and T. Haladyna, (eds.), *Handbook of Test Development, 2nd edition*. London: Routledge.

Looney, J. W. (2011), "Integrating Formative and Summative Assessment: Progress Toward a Seamless System?", *OECD Education Working Papers*, No. 58, OECD Publishing, Paris, http://dx.doi.org/10.1787/5kghx3kbl734-en.

Madaus, G., M. Russell and J. Higgins (2009), *The Paradoxes of High Stakes Testing*, Information Age Publishing, Charlotte, North Carolina.

Mandinach, E.B. and E.S. Gummer (2013), "A Systemic View of Implementing Data Literacy in Educator Preparation", Educational Researcher, Vol. 42 No. 1, pp. 30–37, available at: http://edr.sagepub.com/content/42/1/30.full.pdf+html

McGehee, J.J. and L.K. Griffith (2001), "Large-scale assessments combined with curriculum alignment: Agents of change", *Theory into Practice*, Vol. 40, pp. 137-144.

Messick, S. (1989), "Validity", in R. L. Linn (ed.), *Educational Measurement (third edition)*, American Council on Education and Macmillan Publishing, New York.

Mislevy, R. J., et al. (2006), "Concepts, terminology, and basic models of evidence-centered design" in D. M. Williamson, I.I. Bejar and R. J. Mislevy (eds.), *Automated*

scoring of complex tasks in computer-based testing (pp. 15-47), Erlbaum, Mahwah, New Jersey.

National Research Council (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. J.W. Pellegrino and M.L. Hilton (eds.), The National Academies Press, Washington DC.

Next Generation Science Standards (NGSS) (n.d.), Washington DC, www.nextgenscience.org.

Nusche, D., et al. (2011), *OECD Reviews of Evaluation and Assessment in Education: Norway 2011*, OECD Reviews of Evaluation and Assessment in Education, OECD Publishing, http://dx.doi.org/10.1787/9789264117006-en.

Nusche, D., et al. (2012), *OECD Reviews of Evaluation and Assessment in Education: New Zealand 2011*, OECD Reviews of Evaluation and Assessment in Education, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264116917-en.

OECD (2013), *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*, OECD Reviews of Evaluation and Assessment in Education, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264190658-en.

OECD (2005), *Formative Assessment: Improving Learning in Secondary Classrooms*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264007413-en.

Partnership for Assessment of Readiness for College and Careers (PARCC) (n.d.) www.parcconline.org.

Partnership for 21st Century Skills (n.d), *Framework for 21st century learning*, www.p21.org/our-work/p21-framework.

Rupp, A.A. and N.K. Lesaux (2006), "Meeting expectations?: An empirical investigation of a standards-based assessment of reading comprehension", *Educational Evaluation and Policy Analysis*, Vol. 28, No. 4, pp. 315-333.

Rychen, D.S. and L.H. Salganik (eds.) (2001), *Defining and selecting key competencies,* Hogrefe and Huber, Gottingen, Germany.

Rychen, D.S. and L.H. Salganik (eds.) (2003), *Key competencies for a successful life and a well-functioning society*, Hogrefe and Huber, Gottingen, Germany.

Santiago, P., et al. (2011), *OECD Reviews of Evaluation and Assessment in Education: Australia 2011*, OECD Reviews of Evaluation and Assessment in Education, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264116672-en.

Scheerens, J., et al. (2012), *OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes: Country Background Report for the Netherlands*, University of Twente, Netherlands, www.oecd.org/edu/evaluationpolicy.

Shewbridge, C., et al. (2010), *OECD Reviews of Migrant Education: Netherlands 2010*, OECD Reviews of Migrant Education, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264086210-en.

Smarter Balanced Assessment Consortium (SBAC) (n.d.) www.smarterbalanced.org, accessed 29 January 2014.

Timperley, H., et al. (2008), Teacher Professional Learning and Development: Best Evidence Synthesis Iteration (BES), New Zealand Ministry of Education, Wellington.

U.S. Department of Education (2010), National Education Technology Plan 2010, available at: www.ed.gov/technology/netp-2010.

Van Barneveld, C. (2008), "Using data to improve student achievement", What Works? Research into Practice, Research Monograph No. 15, The Literacy and Numeracy Secretariat, Ontario, www.edu.gov.on.ca/eng/literacynumeracy/inspire/research/ Using_Data.pdf.

Visscher, A.J. and M. Ehren (2011), De eenvoud en complexiteit van opbrengstgericht werken, Vakgroep Onderwijsorganisatie en –management, Universiteit Twente, Enschede, Netherlands.

Wainer, H. (ed) (2000), Computer Adaptive Testing: A Primer (2nd edition), Lawrence Erlbaum Associates, Mahwah, New Jersey.

White, B.Y. and J. Frederiksen (1998), "Inquiry, modeling and metacognition: Making science accessible to all students", Cognition and Instruction, 16(1), 3-118, Routledge, London.

Wiliam, D., et al. (2004). Teachers developing assessment for learning: Impact on student achievement. Assessment in Education: Principles, Policy and Practice, 11(1), 49-65, Routledge, London.

Wiliam, D. (2006), "Formative assessment: Getting the focus right", Educational Assessment, Vol. 11, pp. 283-289.

Williamson, D.M., I.I. Beja and R.J. Mislevy (eds.) (2006), Automated scoring of complex tasks in computer-based testing (pp. 15-47), Erlbaum, Mahwah, New Jersey.

Annex 3.1

Research on 21st century skills

The volumes edited by Rychen and Salganik (2001; 2003) describe DeSeCo, a multi-year project organised under the auspices of the OECD, that "was initiated to provide solid theoretical and conceptual foundations for the broad range of competencies that are needed to face the challenges of the present and the future" (Rychen and Salganik, 2003, p.vii-viii). Calling on an international group of experts from many different disciplines and perspectives, they developed an overarching tri-partite framework: interacting in socially heterogeneous groups, acting autonomously and using tools interactively. Equally important, the report by Rychen and Salganik (2003) includes discussion of the implications of this framework for building the corresponding assessments and national indicators, as well as "the challenges associated with the development of a coherent, long-term assessment strategy based on the theoretical and conceptual foundations provided by DeSeCo" (Rychen and Salganik, 2003, p. 12).

The Partnership for 21st Century Skills (P21) was founded in 2002 in the U.S. and constitutes a joint effort by the Federal government, many state governments and NGOs, and leading corporations to develop and promote the skills necessary for individual and national success in a global economy. The framework augments the core academic skills with four others: critical thinking and problem solving; communication, collaboration, creativity and innovation. P21 works with educators to incorporate the development of these skills into the curricula of all disciplines. Subsequently, the National Research Council of the National Academy of Sciences (U.S.) convened a task force to clarify and organize the vast literature on 21st century skills, as well as to offer recommendations on how these competencies can be developed in school settings and how student progress can be assessed for both formative and summative purposes. The resulting report (National Research Council, 2012) constitutes a valuable foundational document for any country wishing to explore policy alternatives in this area.

Building on recent progress in defining key competencies, academics and testing organizations have been developing assessments for certain 21st century skills, and refining them to achieve acceptable levels of reliability and validity. The work of Kyllonen and his associates at the Educational Testing Service (Kyllonen, 2012; Kyllonen, in press) exemplifies the high quality of this work and the nature of the information obtained from the administration of these assessments. The crucial point is that there has been substantial progress in this arena and, coupled with the introduction of computer-based assessments, acceleration of this progress can be expected.