KeyCoNet
Key Competence Network on School Education

KeyCoNet 2013 Literature Review:

Assessment for key competences

Author: David Pepper

Table of Contents

KEYCONET
Key Competence Network on School Education

1. Introduction

Background and scope

This is one of two literature reviews written in the first year of the KeyCoNet project – a European policy network on the implementation of key competences in school education. In a previous European project, the assessment of learners' key competences was identified as essential for this implementation (Gordon et al., 2009). This particular literature review therefore focuses on assessment issues and responses, and complements the more general literature review on approaches to key competence development across Europe.

This work was originally initiated through the European Reference Framework of Key Competences for Lifelong Learning (OJEU, 2006). The Reference Framework identified eight key competences as necessary for personal fulfilment, active citizenship, social inclusion and employment:

- Communication in the mother tongue
- Communication in foreign languages
- Mathematical competence and basic competences in science and technology
- Digital competence
- Learning to learn
- Social and civic competences
- Sense of initiative and entrepreneurship
- Cultural awareness and expression.

The key competences all emphasize: critical thinking, creativity, initiative, problem solving, risk assessment, decision taking and constructive management of feelings. More fundamentally, the Reference Framework defined competence as knowledge, skills and attitudes applied appropriately to a given context.

The scope of this literature review is therefore the assessment of learners' key competences, or similar constructs that emphasise not only the knowledge but also the skills and attitudes needed for lifelong learning, as an aspect of the implementation of the European Reference Framework of Key Competences for lifelong learning. With reference to Gipps (1994), Mislevy (1994) and CEDEFOP (2008b), this literature review is based on a working definition of assessment as:

> The process of making inferences about an individuals' knowledge, skills, attitudes or other constructs using information from one or more methods such as tests, observations, interviews, projects or portfolios with reference to pre-defined criteria.

CEDEFOP's (2008b) *Terminology of European Education and Training Policy* comments that, in the English-language literature:

> …'assessment' generally refers to appraisal of individuals whereas 'evaluation' is more frequently used to describe appraisal of education and training methods or providers.

Evaluation is therefore defined as distinct from assessment and beyond the scope of this literature review. However, assessment can contribute to evaluation, and assessment policies and practices require evaluation. This is a subtle distinction that recurs in the course of this review.

The literature review was conducted in July 2012 and built on an earlier review of the assessment of key competences (Pepper, 2012b) and also incorporates research on 'computer-based assessment' or 'e-assessment'. Following Busuttil-Reynaud and Winkley (2006), this is defined as assessment using information and communication technology to present information or record, analyse, report or feedback on responses. This aspect of the literature review draws on an earlier review of the e-assessment of key competences (Redecker, 2013).

The literature reviews published as Pepper (2012b) and Redecker (2013) contributed to the European Commission policy guidance on the assessment of key competences (European Commission, 2012).[1] This policy guidance accompanied the European Commission's 'Rethinking Education' strategy, which emphasised the need for individuals who can contribute to innovation and entrepreneurship, particularly at a time of economic difficulty.

The present literature review was updated in July 2013 on the basis of a literature search covering the previous 12 months. As a result, 10 new sources were added and four forthcoming sources were amended to reflect published versions.

Curriculum and assessment

There is evidence of a growing trend towards curricula based on key competences or similar broad conceptions of teaching and learning encompassing not only knowledge but also the skills and attitudes needed in a wide range of real-life contexts (Gordon, et al., 2009). Although curricula generally continue to be organised according to subjects or areas, the aim is for learning not just within these subjects and areas but also across them and sometimes beyond them altogether (Gordon, et al., 2009; Pepper, 2011; Schneider & Stern, 2010).

There is, however, evidence that changes to curricula have not been fully reflected in changes to assessment. With reference to the 27 EU Member States, a joint progress report of the European Council and the Commission (2009, p. 3) found that:

> A large number of countries are introducing reforms that explicitly use the Key Competences framework as a reference point. Good progress has been made in adapting school curricula. But there is still much to be done to support teachers' competence development, to update assessment methods, and to introduce new ways of organising learning.

Eurydice (2009) found that of the eight key competences, only communication in the mother tongue, communication in foreign languages, mathematical competences and basic competences in science and technology were commonly assessed by national tests. Yet there is evidence of a wider range of key competences assessed by teachers and learners using a wider range of methods. However, these practices appear less widespread or systematic and require more support through policy mechanisms such as teacher education and evaluation (Gordon, et al., 2009; Pepper, 2011). Since assessment has the potential either to support or undermine the

[1] The policy guidance was published at: http://ec.europa.eu/education/news/rethinking/sw371_en.pdf

conception of teaching and learning that underpins the curriculum, it clearly needs more attention in policy and practice.

Assessment issues are central to the education research literature, where there is widespread recognition that assessment strongly influences teaching and learning (P. Black, 1998; Koretz, 2005; Stobart, 2008b). Key competences arguably represent a valuable but complex view of learning. There is a particular risk is that if only a few competences are assessed, assessment will distort the curriculum, leading to the neglect of other competences. Furthermore, if only limited aspects of these competences are assessed, they will be distorted too. Thus if only knowledge is assessed, the development of skills and attitudes will be, at best, incidental.

The potential of assessment is that, rather than only assessing the learning that is easy to assess, it will tell us about the learning that is, by consensus, important. Crucially, assessment will then result in increased time and effort spent on this learning. Assessment will therefore support effective changes not only in what is taught but also how it is taught, and consequently what is learnt and how it is learnt. In other words, assessing learners' key competence not only documents learners' key competences but is also essential to the development of learners' key competences. It is therefore doubly important to have some basis for evaluating assessments of learners' key competences.

Validity, reliability and equity

In educational assessment, validity is a central concept because it provides an overarching criterion for evaluating assessments. It is therefore the foremost technical consideration for any assessment, including the assessment of key competences. A broad definition of validity has been gradually accepted as unifying the various earlier definitions (Brennan, 2006). This describes validity as:

> …an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (Messick, 1989, p. 13).

A general methodology for the validation of assessment therefore begins with an explicit statement of the proposed *inferences* and *actions* that will be based on assessment results (Kane, 2006). The proposed inference could be the extent to which each learner has developed the key competences. The issue, simply stated, is the extent to which the assessment assesses what it is intended to assess (Gipps, 1994; Wiliam & Black, 1996) – in this case, each learner's key competences. The proposed actions could serve either formative or summative purposes (Newton, 2007).

Formative assessment is often called 'assessment for learning' because it is concerned with using assessment information to promote an individual's learning during a period of instruction. This is distinguished from summative assessment or 'assessment of learning', which reports an individual's learning at the end of a period of instruction. Since assessment purposes refer to the use of assessment information rather than the assessment itself, the assessment method is independent of the assessment purpose (P. Black & Wiliam, 2003). Written tests or teachers' observation could each either be used for formative or summative purposes. However, with reference to validity, it is important for the design of an assessment and the use of information from assessment to be consistent with one another. Assessment systems need to be designed in order to serve formative and summative purposes, so that assessments can contribute to the development and reporting of learners' key competences.

The broad definition of validity as a criterion for evaluating assessments subsumes other important but narrower criteria, such as reliability and equity (see, for example, Morris, 2011). Reliability is '…often defined as, and measured by, the extent to which the assessment, if repeated, would give the same result' (Harlen, 2007, p. 18). Although validity and reliability are seen as being in tension with one other, reliability is an aspect of the broad conception of validity. Thus whilst an assessment can be reliable without being valid, it cannot be valid without being reliable. For example, a test can be made more reliable by limiting its question types and response formats, making it straightforward to interpret and highly reliable. However, such a test would provide a narrow picture of the key competences needed for lifelong learning, undermining its overall validity. On the other hand, teachers' day-to-day observations may provide a broader picture of learners' competences. However, different teachers may interpret and weight these observations differently, compromising reliability and therefore validity. In practice, however, assessments can find a balance between reliability and overall validity according to the assessment purpose. Assessments designed for summative purposes therefore emphasise reliability, assessing a limited number of performances and range of the curriculum. Assessment designed for formative purposes then emphasise overall validity, assessing more performances in a wider range of contexts. This literature review explores some of the potential for assessment to move beyond this dichotomy using technological or professional innovations.

Equity is sometimes seen as distinct from validity and reliability because inferences and actions could be repeatedly inequitable. Equity emphasises the social nature of assessment and highlights the need to consider differences that are not the focus of an assessment but could influence the assessment. For example, assessments can be developed or modified to ensure that when a learners' disability is not relevant, it is not assessed (Pepper, 2007). Since key competences emphasise real-life contexts, learners should have access to special arrangements such as Braille, large print or screen readers – just as they would in their everyday lives. There is, however, a more general need to accommodate the different backgrounds and circumstances of all students. Since it is impossible for assessments to be acultural, it is important: firstly, for the learning outcomes that will be assessed to be clearly articulated; secondly, for assessment methods to be justified; and, thirdly, for the assessment process to be transparent (Stobart, 2008b). The following section focuses on the articulation of learning outcomes as a fundamental basis for developing assessments for key competences.

2. Learning outcomes

The rationale for learning outcomes

The European Reference Framework identifies and defines eight key competences that learners should develop in a wide range of everyday contexts by the end of initial education. The Reference Framework is intended to be interpreted within the particular circumstances of each Member State's education system. It therefore stops short of prescribing the contexts in which learners should develop key competences, the knowledge, skills and attitudes they require for particular contexts or the levels of competence they should develop as they progress through initial education.

In a major study on key competence for the European Commission, Gordon et al (2009) found several examples of EU Member States interpreting the Reference Framework in the circumstances of their own education system, sectors and levels. Pepper (2011) reviewed and updated this finding and confirmed a trend for Member States to specify key competences in learning outcomes relating to knowledge, skills and attitudes for specific contexts within individual subjects, across two or more subjects or beyond traditional subjects boundaries altogether. These ‘learning outcomes’ are statements of what a learner should be able to do or be and contrast with learning inputs such as time, location and method (CEDEFOP, 2008a; Leney, Gordon, & Adam, 2008). The process of developing the learning outcomes begins by defining the competence and proceeds with identifying its sub-domains. For example, the OECD’s PISA 2012 survey assessed the following problem-solving processes: exploring and understanding; representing and formulating; planning and executing; and, monitoring and reflecting (PISA Consortium, 2010). Learning outcomes can be then be specified within sub-domains and can be set out in curriculum, programme, standards, syllabi or assessment documents.

Several theoretical and policy perspectives, supported by empirical research, identify a need to specify learning outcomes to provide a basis for teaching, learning, assessment and evaluation.

Though interrelated, it is possible to discern three theoretical perspectives and three policy perspectives:

Psychometrics: Specifying the scope of the assessed domain, relevant constructs and proposed interpretations provides a basis for developing instruments that collect the necessary information. (Brennan, 2006). The emphasis is on promoting validity.

Competence-based assessment: The need to identify and assess the learning outcomes necessary for particular sectors or occupations originates from the vocational education literature of the 1970s (Wolf, 2001). The emphasis is on focused learning and predictive validity.

Assessment for learning: A shared understanding of the learning outcomes provides a basis for identifying where learners should be, where they are now and how to close the gap between the two (P. Black & Wiliam, 1998b; Sadler, 1987). The emphasis is on dialogue and learning.

These theoretical perspectives have, to a greater or lesser extent, influenced three overlapping policy perspectives:

Standards: There has been national and international interest in specifying learning outcomes in ‘standards’ that provide a focus for assessment and a basis for steering education systems through evaluation since the 1980s and this continues to be relevant today (Looney, 2011; Mays, 1995).

Equivalence: The need to compare the competences of a mobile workforce across national boundaries has led to important development such as the European Qualifications Framework, which is predicated on a shift to using learning outcomes (CEDEFOP, 2011b; Leney, et al., 2008).

Lifelong learning: International agencies such as UNESCO, OECD and EU and countries in and beyond Europe have sought to identify the learning outcomes needed by individuals and society as a whole, such as those denoted by the key competences (Delors, 1996; OJEU, 2006; Pellegrino & Hilton, 2012; Rychen & Salganik, 2003).

The need to specify key competences as learning outcomes that provide a focus for assessment practices is therefore supported by a range of empirically-derived theoretical and policy perspectives. However, a number of sources suggest that although key competences are widely recognised as important, they are generally not specified in learning outcomes.

An OECD survey of ‘transversal skills or competencies’, including creativity, innovation, critical thinking, problem solving, decision-making and communication in 17 countries, including 10 EU Member States, found, that few of the countries had either defined these terms or developed clear assessment policies for them. The authors conclude that these omissions were closely related because: ‘Rigorous assessment methods cannot of course be developed without clear definitions of the skills and competencies in question’ (Ananiadou & Claro, 2009, p. 16). In their research on the key competences across the 27 EU Member States, Gordon et al similarly found that the ‘transversal’ key competences and cross-cutting themes were rarely specified in learning outcomes. Whilst the ‘traditional’ key competences were generally specified in learning outcomes, these were often limited to contexts that related only to their most closely-related subjects rather than the wider curriculum.

Knowledge, skills, attitudes and contexts

The literature suggests that the specified learning outcomes should consist of not only knowledge and skills but also attitudes. The OECD’s DeSeCo project defined competence as ‘the ability to successfully meet complex demands in a particular context… the mobilization of knowledge, cognitive and practical skills, as well as social and behavior components such as attitudes, emotions, and values and motivations’ (Rychen & Salganik, 2003, p. 2). Competence was therefore a ‘holistic notion’ and ‘therefore not reducible to its cognitive dimension’ (Ibid.). The specification of attitudes that support the development and application of knowledge and skills is therefore essential.

In keeping with the European key competences, DeSeCo also asserted that the ‘constellations’ of key competencies would vary according to the context. It follows that learning outcomes relating to knowledge, skills and attitudes should be derived from an analysis of the needs of particular contexts. In their own terms, Pellegrino and Hilton (2012, p. 128) suggest that:

> ...a clear delineation of the learning goals and a well-defined model of how learning is expected to develop...[which] may be hypothesized or established by research – provides a solid foundation for the coordinated design of instruction and assessment aimed at supporting students’ acquisition and transfer of targeted competencies.

The authors therefore give a sense of the need to sequence contexts so that their demands are matched to learners’ developing competences, and that these demands can be identified through prior theoretical or empirical research. Furthermore, when selecting contexts for the demonstration of competences, it is important to consider the extent to which the competences required transfer to other contexts, and therefore have wider significance.

With reference to earlier research, Haggerty Haggerty, Elgin, and Woolley (2012) used five interrelated social and emotional competencies as the basis of their review of social and emotional learning assessments for use with middle school students:

- **Self-awareness:** Accurately assessing one’s feelings, interests, values, and strengths; maintaining a well-grounded sense of self-confidence.
- **Self-Management:** Regulating one’s emotions to handle stress, controlling impulses, and persevering in addressing challenges; expressing emotions appropriately; and setting and monitoring progress toward personal and academic goals.
- **Social awareness:** Being able to take the perspective of and empathize with others; recognizing and appreciating individual and group similarities and

differences; and recognizing and making the best use of family, school, and community resources.

- **Relationship skills:** Establishing and maintaining healthy and rewarding relationships based on cooperation; resisting inappropriate social pressure; preventing, managing, and resolving interpersonal conflict, and seeking help when needed.
- **Responsible Decision Making:** Making decisions based on consideration of ethical standards, safety concerns, appropriate social norms, respect for others, and likely consequences of various actions; applying decision-making skills to academic and social situations; and contributing to the well-being of one's school and community.

Each of these social and emotional competencies is therefore defined, indicating potential sub-domains and bringing them closer to being operationalised for the assessment instruments that the authors identify – though not yet specific learning outcomes *per se*. In the context of the European competences, self-awareness and self-management could be seen as sub-domains of learning to learn. Responsible decision making – a cross-cutting theme in the Reference Framework – could be particularly applicable to the initiative and entrepreneurship competence. Social awareness and relationship skills, however, could be seen as distinctive sub-domains for social and civic competences. In other research, *perceived emotional intelligence* has been defined as comprising three factors, namely: attention to your feelings; clarity of perception of your feelings; and, terminating your negative emotions and prolonging your positive emotions (Sanchez-Nunez, Fernandez-Berrocal, & Latorre, 2013). All three of these factors could be seen as comprising the constructive management of feelings, one of the themes that cut across the key competences.

In research in Chile that drew upon national evaluations in the UK, USA and Australia, ICT literacy was defined as solving problems of information, communication and knowledge in digital environments (Claro et al., 2012). Three dimensions of ICT literacy, each with two sub-dimensions, were identified:

1. **Information fluency**
   - ICT fluency in sourcing information
   - ICT skills in processing information

2. **Effective communication**
   - ICT skills in effective communication
   - ICT skills in collaborative and virtual environments

3. **Ethics and social impact**
   - Evaluation of responsible ICT use
   - Evaluation of ICT social impact

Particularly notable in the operationalisation of this definition of ICT literacy, which recalls the EU definition of digital competence, is the explicit emphasis on the ethics and social impact of ICT. Here, each of the dimensions and sub-dimensions were elaborated so that, for example, the ICT fluency in sourcing information sub-dimension meant to: 'search, select, evaluate, organize and manage digital information'. The assessment was then designed to gather information relating to these and other requirements.

Critical thinking is one of the cross-cutting themes in the EU framework of key competences. In a programme of research supporting Cambridge Assessment's development of qualifications relating to *critical thinking*, an expert panel was convened to consider the validity of these assessments (B. Black, 2012). The panel began by developing a definition of critical thinking, then a taxonomy of critical thinking and finally a glossary of critical thinking. Critical thinking was defined as:

> The analytical thinking which underlies all rational discourse and enquiry. It is characterised by a meticulous and rigorous approach (p.125).

The definition therefore places a particular premium on analysis. However, it also incorporated 'the processes involved in being rational', which were elaborated as five skills/processes and many sub skills/processes in the taxonomy of critical thinking:

Skill/process	**Sub skills/processes**
1. Analysis	1. eg Identifying unstated assumptions
2. Evaluation	2. eg Assessing analogies
3. Inference	3. eg Considering the implications of claims
4. Synthesis/construction	4. eg Selecting material relevant to an argument
5. Self-reflection/self-correction	5. eg Questioning own's own preconceptions

To provide further precision for the development of assessments, the expert panel then developed the glossary of critical thinking. Naturally, this included common terms used in relation to critical thinking. However, it also included terms on the fringes or even outside of the conception of critical thinking. This was a basis for not only assessment developers but also students and teachers to gain a shared understanding of critical thinking and to draw fine distinctions with closely-related concepts such as problem-solving. This suggests that the definition, taxonomy and glossary of critical thinking could therefore inform the development of learning outcomes for use in formative and summative assessments.

Creativity is another of the cross-cutting themes in the EU framework of key competences. Spencer, Lucas, and Claxton (2012) developed a conception of creativity that was based on a review of different bodies of literature. The researchers identified five 'habits' of creativity across this literature. These were being: inquisitive, persistent, imaginative, collaborative and disciplined. Five habits was thought a small enough number for practical but precise assessments. Each of the habits comprised a further three sub-habits (thus 15 sub-habits in total). In a field trial with six primary/secondary schools in England, teachers found the sub-habits too onerous to be practical for formative assessment. In a second field trial with 11 primary/secondary schools (five of which had participated in the first field trial), the sub-habits were therefore consolidated. However, this approach was not directive enough, particularly for learners to self-assess themselves, and 'gaps' were less obvious. The researchers therefore recommended separating back out the three sub-habits of each habit but developing training materials for teachers.

In relation to curricula and training programmes, CEDEFOP (2011b, p. 24) offers the following advice on constructing learning outcomes:

> Learning outcomes in curricula [and training programmes] usually begin with the phrase:
>
> *…The learner is (or will be) able to…*
>
> This phrase is followed by an action verb so that students are able to demonstrate what they have learned. Words such as 'know' or 'understand' do not help with this demonstration of learning and are therefore usually avoided because it is not clear to the learner the level of understanding or amount of knowledge required.

> Different verbs can be used to demonstrate different levels of learning… At a basic level the learning outcomes may require learners to be able to define, recall, list, describe, explain or discuss. For a more advanced programme the learners may be expected to be able to formulate, appraise, evaluate, estimate or construct. The verb will usually be followed by words indicating on what or with what the learner is acting and the nature or context of the performance required as evidence that the learning was achieved. These additional words also indicate the level of learning achieved.

This source also offers examples where 'demonstrate' is used in learning outcomes and the extract itself also uses this verb. This verb has the benefit of indicating that the assessment method should seek evidence relating to the learning outcome, providing the learner with the opportunity to demonstrate their competences through the information that is gathered and interpreted for assessment purposes.

In summary, the specification of learning outcomes can therefore provide a basis for focusing teaching and learning, including assessment, on creating opportunities for learners to develop and demonstrate their key competences. However, for assessment validity it is important to ensure that:

> …the assessment concerns all aspects – and only those aspects - of students' achievement relevant to a particular purpose. Including irrelevant aspects is as much a threat to validity as omitting relevant aspects. Thus a clear definition of the domain being assessed is required, as is adherence to it (Harlen, 2007, p.18).

This quote highlights not only the need for learning outcomes to be clearly specified but also for these learning outcomes to be the sole focus of the methods used to assess learners' key competences. If the range of specified learning outcomes is not assessed, *construct under-representation* may result. This means users of assessments won't know enough about the different aspects of each learner's key competences. If learning outcomes other than the ones that were specified are assessed, construct-irrelevant variance may result. This means what users of assessments might think they know about learners' key competences is actually affected by 'noise' from irrelevant information. Avoiding these problems is linked to a more general need to accommodate the different backgrounds and circumstances of all students. It is therefore important for the learning outcomes on which assessments are based to be clearly articulated (Stobart, 2008b).

Balancing specification and judgement

Although the literature makes a strong case for the specification of key competences in learning outcomes, it also emphasises the need to balance the amount specification that assessors are required to work with. Thus the European report by Gordon et al (2009, p.146) found that:

> There were a number of examples of Member States adopting this type of approach to key competence assessment. The challenge for them may be to make this assessment manageable without reducing learning to a series of narrow targets that militate against key competence acquisition.

Recent literature reviewing implementation in different countries suggests that highly specified learning outcomes should be avoided. Over-specification of learning outcomes in the South African national qualifications framework has become a case in point (Allais, 2007). There are two main issues. Firstly, when learning outcomes are highly specified, holistic competences are reduced to atomised tasks. Teaching, learning and assessment is then characterised by the following of scripts provided by long check lists of actions and behaviours (Kerka, 1998; Wolf, 2001). Rather, when competences are specified, it should be the case that ‘the whole is greater than the sum of the parts’ (Council on Education for Public Health, 2011).

Secondly, the need for assessment to be relevant to complex contexts, including occupational contexts and social contexts more generally, means that assessors need to be able to exercise their judgement in any given set of circumstances. Wolf (2001) argued that:

> The inherent variability of the contexts in which competence is tested and displayed means that assessors have to make constant, major decisions about how to take account of that context when judging whether an observed piece of evidence "fits" a defined criterion. In other words, they operate with a complex, internalised, and holistic model-not a simple set of descriptors lifted from a printed set of performance indicators.

The exercise of assessors’ judgement is therefore unavoidable and, in fact, desirable and makes it practicable for competence to be demonstrated in different ways in different contexts (Kerka, 1998). This is also consistent with the literature on assessment for learning, which emphasises that all those involved in assessment should have a shared understanding of the learning outcomes so that they can be applied consistently.

Lastly, the precise balance between specification of learning outcomes and the judgement of assessors will depend on the assessment purpose. CEDEFOP (2011b, p. 7) argues ‘that the way in which learning outcomes are expected to be used, affects the way in which they are formulated’ and that ‘the key attribute of a learning outcome is that it is expressed in a level of detail that makes it fit for purpose’. Thus the learning outcomes that provide the basis for the development of a summative assessment for a qualification will be more tightly specified than the learning outcomes used in connection with formative assessment in the school curriculum. Indeed, the literature emphasises the value of using learning outcomes to outline the expected progression of learners but, where possible, leaving some scope for teachers to adapt the curriculum and their pedagogy to their local contexts and the needs of their students (Stanley, MacCann, Gardner, Reynolds, & Wild, 2009).

3. Assessment methods

This section considers assessment methods with the potential to assess the scope and range of learners’ key competences for summative or formative purposes, including methods such as standardised tests, attitudinal questionnaires, performance-based assessment, portfolio assessment, and teacher, peer and self-assessment practices.

Standardised tests

Standardised tests are tests that are developed, administered, scored and graded according to uniform procedures designed to ensure consistent outcomes that can be meaningfully compared across a population (Morris, 2011). Eurydice (2009) found that, of the eight key competences in the European Reference Framework:

> …only three, namely communication in the mother tongue, communication in foreign languages, and mathematical competences and basic competences in science and technology, can be directly linked to individual subjects… these three competences are the ones most commonly assessed in national [standardised] tests. By contrast, in many European countries the remaining key competences such as ‘learning to

> learn' or social and civic competences, which usually relate to more than one subject, are not at present generally assessed in national tests.

A few Member States reported recently developing standardised tests at national/system level for social and civic competences. However, there were none for the remaining key competences: learning to learn, sense of initiative and entrepreneurship or cultural awareness and expression. It is possible that these competences are implicitly assessed through some national standardised tests, or even explicitly assessed through methods other than these tests. However, national tests tend to reflect the priorities of education systems. The evidence suggests that, although highly valued, these four key competences are much less widely assessed (Eurydice, 2012) .

Although Popham (2001) expressed concern about 'item-teaching', now commonly known as 'teaching to the test', others argue that what is really needed is a test worth teaching to (P. Black et al., 2011). Indeed, the literature reviewed in this section suggests that standardised tests can contribute to the assessment of key competences if they include items with:

- Structure and content that reproduce real-life contexts authentically
- Multiple steps requiring a chain of reasoning and a range of competences
- A range of formats allowing responses that require different competences.

This section provides some examples for the assessment of a wider range of competences using standardised tests, sometimes with reference to combinations with other methods.

The OECD's PISA surveys focus on the competencies of students aged 15 in reading literacy, mathematical literacy and scientific literacy. Other assessed domains include financial literacy and problem-solving. Pellegrino and Hilton (2012) note that there is an emerging consensus that problem-solving needs to be developed and assessed in the contexts provided by specific content domains. However, these domains can introduce other constructs such as literacy and numeracy, reducing the focus on problem-solving. The PISA 2012 assessment of problem-solving attempted to minimise the literacy demand by developing several items for each problem context. In other situations, learners' combinations of different competences might, however, be of interest.

In fact, the frameworks for all of the PISA domains emphasise problem-solving in real-world contexts and tests are employed to assess students in each domain. The test items have a range of formats including open-constructed response (requiring details or explanation), closed-constructed response (often numerical) and selected-response (multiple choice) items. Some selected-response items are complex multiple choice items, where more than one response may be correct, which potentially more closely resembles real-life conditions in some contexts. As a whole, the items present students with different types of context. In mathematics, for example, these are individual, societal, occupational or scientific contexts (OECD, 2010). This provides some insight into how mathematics, for one, might be assessed as a competence across the curriculum (or even beyond the formal curriculum). Items in each of the domains present students with varying degrees of complexity, sometimes requiring multiple steps, as might be the case in real contexts. Similarly, students are frequently asked to make sense of a significant amount of information presented as text or in graphics. This reading demand introduces another sense in which the items require cross-curricular key competences. However, such combined demands need to be varied in order to gain a clear picture of each learner's specific key competences. Saxton, Belanger, and Becker (2012) show how this can be implemented with acceptable reliability using an 'analytical rubric' to score students' responses to test items requiring short or extended responses designated as assessing individual or combined critical thinking 'sub-skills'.[2]

Researchers at the University of Amsterdam developed a cross-curricular skills test to assess the 'competence' of students aged 15-16. The test was designed to assess 'cross-curricular skills', defined as general skills that could be taught and practiced in different disciplines. The test consisted of 56 multiple-choice items, which the researchers acknowledged is 'an item-format that is not a customary one for measuring general skills' (Meijer, Elshout-Mohr, & Wolters, 2001, p. 79). The test was administered to 465 students in a pilot study and 9,000 students in the main study. The researchers concluded that it was a valid and reliable test for cross-curricular skills. However, the authors accepted that, whilst a multiple-choice test is practicable for large-scale surveys, alternative formats such as portfolios and authentic performance tests may be preferred in 'classroom settings'. They suggested that:

> Ideally, users should have a complete set of assessment instruments at their disposal with different values on at least such factors as 'content' (covered skill area), 'format' (multiple-choice, performance measure, self report), and 'practicability' (Meijer, et al., 2001, p. 104).

In a review of the literature, Morris (2011) found wide agreement that standardised tests can only provide a limited picture of student performance. This is because tests can only:

- Assess performance infrequently (without seriously reducing instruction time)
- Sample part of a domain at any one time (without becoming a test of endurance)
- Reproduce a limited range of contexts authentically and only require certain response types.

However, e-assessment offers the potential to meet these challenges. Bunderson, Inouye, and Olsen (1989) anticipated four generations of e-assessment:

1. Conventional 'linear' standardised tests administered using computers
2. As Generation 1 but 'adaptive' tests adjust their difficulty to learners' ability
3. Continuous assessment and reporting integrated into pedagogic activities
4. As per Generation 3 but the assessment also makes inferences about students' competences.

Whilst first and second generation e-assessments offer efficiency gains in test administration, the move from second to third generation e-assessment represents a more significant change. It introduces the possibility of accessing learning outcomes from processes that are too complex for conventional assessment to capture in context, reproduce authentically or interpret meaningfully. Thus whilst first and second generation e-assessments require learners to develop basic digital competences even simply to access tests in relatively limited contexts, third and fourth generation e-assessments would create sophisticated environments with contexts that can be designed to assess various levels and combinations of competences. Negotiating the significant shift from Generation 2 to 3 therefore presents an opportunity for the development of assessments for key competences. Bennett (2010) refers to this step change as heralding 'Generation Reinvention' with simulation of complex tasks, the sampling of student performance repeatedly over time, the integration of assessment and pedagogy and collection of information about complex learning outcomes in ever more sophisticated ways.

[2] Specifically, these sub-skills were: interpretation, analysis, evaluation, inference, explanation and disposition. These 'sub-skills' are similar to the 'processes' in the Cambridge Assessment taxonomy of critical thinking (B. Black, 2012), detailed in the section in this review on specifying learning outcomes.

None of this, however, is to say that the development of Generation 1 linear tests and Generation 2 adaptive tests has reached an endpoint. In Generation 2 computerised adaptive testing (CAT), for example, learners first respond to some items so that computers can make an initial estimate of their 'ability' (in the language of psychometrics). The computer then adapts subsequent item difficulty to this ability estimate. The computer then makes interim estimates of ability, making further adjustments of item difficulty until the estimates of ability converge. Modern computers apply complex algorithms make the assessment decisions in real time within the testing environment (van der Linden & Pashley, 2000). CAT is in widespread use, with several examples in the USA and in European countries such as the Netherlands and Denmark (Redecker, 2013). These tests require a large bank of items firstly to accommodate different levels of ability and secondly to avoid public exposure of the best (ie most discriminating) items and practice effects. However, the existing examples appear to be relatively conventional, assessing subject-based knowledge and skills.

Researchers from universities in Chile developed a test of ICT literacy, which represents a set of cross-curricular aims for primary education in Chile. The development of this test was informed by national evaluations of ICT tests in the UK, USA and Australia. The section of this review on specifying learning outcomes detailed how ICT literacy was operationalised with three dimensions: information fluency; effective communication; and, ethical and social impact. These dimensions were assessed in a virtual environment intended to mirror actual environments in which ICT literacy is required. Notably, the environment incorporated interaction with *virtual* peers, suggesting potential for developing both social and digital competences. The assessment comprised three major tasks, each relating to one of the three dimensions. The 40 test items retained for these tasks had either a multiple-choice format requiring students to work through the tasks before responding or an open format requiring students to produce an artefact such as an email, a document or a post in a forum. When Chilean students were assessed (N=1185), the researchers found that their performance on the information dimension and the communication dimension were closely-related. However, their performance on the ethics and social impact dimension was distinct from these two dimensions. This indicates that, if learning in relation to the ethics and social impact of ICT is a valued outcome, then it may need to be an explicit focus for specific test items. However, the researchers also found that the extensive definition of ICT literacy had resulted in a test that was too long (2.5 hours) and concluded that a better balance between priorities for assessment was required.

Third and fourth generation e-assessment offers particular hope for context simulation and dynamic interaction. By simulating real-life contexts and repeatedly sampling performance, these e-assessments open up new possibilities for eliciting information about the range and scope of learners' key competences. Furthermore, rather than explicitly presenting learners with contextual information, simulations can require learners to make sense of the context for themselves. Simulations can also create dynamic contexts which interact with learners. This means that every action leads to a reaction, requiring repeated re-evaluation of the task conditions, much as in real life. For example, online games can transpose the social and political context of a real-world problem into a virtual model.

McLarin's Adventures is a Massive Multiplayer Online Game (MMOG) designed to develop the problem-solving skills of 8th and 9th grade students in the USA. In this MMOG, students play the role of researchers exploring an uninhabited Earth-like planet in order to ensure the survival of humans in this new environment. The game involves a complex and ill-structured task scenario, requiring students to apply mathematics, science, geography, geology, social studies and literacy to solve problems. This involves two processes: representing the problem and generating a solution. These are expressed diagrammatically in ‘causal representations’. Directions and hints are embedded in the game. Eseryel, Ifenthaler, and Ge (2013) compared the validity of assessing students’ complex and ill-structured problem-solving in this MMOG using either manual or automated methods. The former was an established method where raters used a rubric (containing descriptors and criteria) to compare students’ and experts’ causal representations. The latter was a new method where a computer algorithm similarly compared these causal representations. The two methods achieved similar results, suggesting that the efficient automated method could be developed into a single, unified means of assessing complex problem-solving. Pre- and post-tests using the methods indicated that students’ causal representations became more complex during an extended period of playing the MMOG. However, these representations were founded on students’ development of new misconceptions. This highlights the need for developmentally appropriate problems and feedback through formative assessment.

Quest Atlantis is a multiplayer game that immerses children aged 9-15 online and offline learning activities where they role play and make responsible decisions in fictional circumstances. In one set of circumstances, children play park rangers trying to identify the reasons for declining numbers of fish in their national park and develop a solution to the problem. The activity is based on scientific enquiry, such as taking water samples and conducting interviews, and assessment is embedded in the activity. To solve the problem, the children need to develop knowledge, skills and attitudes associated principally with scientific competence but also with civic competence, social competence and digital competence. There is potential to assess learners use of combinations of these competences (Redecker, 2013).

Unless or until e-assessment is ready to provide the basis for a quantum leap in our conception of tests and assessment instruments more generally, then multiple sources of information about learners’ competences are likely to be needed. Since well-designed tests can provide part of the picture of student performance, they can usefully be combined with other methods of assessment. Employing multiple measures of students’ learning outcomes, potentially including tests, therefore ‘reduces the risk of making incorrect decisions… improves the validity of the system, and reduces the likelihood of excessive narrowing of curriculum’ (Morris, 2011, p.44). In particular, whilst tests may only implicitly assess attitudes, other instruments such as questionnaires, observation, dialogue and performance-based assessments such as presentations, projects or portfolios may be more explicit in their assessment of attitudes. However, Redecker and Johannessen (2013) argue that education policies and practices need to do more to exploit the potential of existing technologies through research, development and evaluation.

Attitudinal questionnaires

In educational psychology and education research more generally, learners’ attitudes (or ‘affect’) are frequently treated as explanatory variables for their academic performance (or 'cognition'; Alexander & Winne, 2006; Stobart, 2008a). Thus, for example, PISA surveys employ questionnaires to survey students’ attitudes to learning to help explain their individual performance. In contrast, the European framework for learning to learn identified attitudes as learning outcomes in their own right (Fredriksson & Hoskins, 2008). This framework combined tests assessing the cognitive sub-domain of learning to learn with questionnaires assessing the affective and metacognitive sub-domains.[3] These instruments were piloted with students aged 14 in 8 European countries in 2008. The results were reported as indicating that all aspects of

[3] Kupiainen, Hautamäki and Rantanen (2008) report that some of the affective sub-scales were assessed using two questionnaire items, which was too few to draw valid inferences about students’ attitudes to learning.

the test and questionnaire instruments required further development but one specific finding is instructive. The most complex items on the test appeared to assess not only the cognitive aspects but also, indirectly, the affective aspects of learning to learn. Many students opted not to attempt these items or to persevere with them. This could also be interpreted as metacognition and strategic test-taking behaviour, with learners focusing on questions that appear more likely to lead to success. The example highlights the difficulty of trying to assess cognitive and affective aspects of learning in isolation from one another.

In their review of assessments of the social and emotional competencies of middle school students, Haggerty, Elgin and Woolley (2012) identified and evaluated 73 assessment instruments. 10 of these instruments met their criteria relating to: the target population (middle school students in general), the assessment of changes over time, the measurement properties (reliability and validity) and the practicality of administering the assessments. The authors identified five social and emotional competencies. These were detailed in the section of this literature review on specifying learning outcomes, which identified two of these competencies, namely relationship skills and social awareness, as distinctive aspects of social competence. The authors found that all 10 of the instruments assessed relationship skills and 9 of the 10 assessed social awareness.[4] All of the instruments were questionnaires with one or more type of rating. These were self-report ratings (used in 8 instruments), teachers/staff ratings (used in 7) or parent/guardian ratings (used in 6) of learners competencies. The authors expressed a preference for self-report ratings, emphasising their lower administrative burden. Five instruments included all three types of rating, thus incorporating three different perspectives on learners' competencies. In these cases, guidance for assessors on how to draw inferences from these three perspectives may be appropriate. Similarly, self-report questionnaire items have been used to assess the different aspects of learners' own perceived emotional intelligence (Sanchez-Nunez, et al., 2013). These aspects were detailed in the section on specifying learning outcomes.

These assessment instruments for social and emotional competencies, generally developed in the USA and commercially available, are mainly orientated towards identifying or predicting problems and in some cases this means assessing risk factors rather than learning outcomes (including not only at individual level but also at school, family and community level). They would therefore require some re-orientation to assess social competences envisaged as learning outcomes by the European Reference Framework. However, the Devereux Student Strengths Assessment (DESSA) was one of the 10 instruments with an explicit strengths focus and reported good levels of validity and reliability. There are 72 DESSA items, each with a five-point frequency scale (eg During the past four week, how often did the child... cooperate with peers or siblings? Never/Occasionally/ Frequently/ Very Frequently). The authors of the review also highlighted this assessment, noting that it is a teacher rating assessment that could be developed into a self-report assessment.

Although most of the other instruments in this review included self-report ratings and a large number of items (in some cases, well in excess of 100). This raises two issues. Firstly, with so many items, respondents, whether children or adults, may reflect less reflection on the issues raised by each item. Although questionnaire instruments may be practical, a more fundamental issue is whether the questionnaire items and their agreement or frequency rating scales can capture varied social contexts and complex emotional experiences. The social and emotional instruments generally did not make references to specific contexts and therefore related only to

[4] The authors also found that 9 of the 10 instruments also assessed self-management and responsible decision making but only 4 of the 10 assessed self-awareness, suggesting this specific competence was not seen as part of the social and emotional domain by the developers of six of the instruments.

learning processes. They therefore lacked information about learning outcomes associated with the demands of specific contexts.

The PISA student questionnaires are relevant here because they have included items relating to students' motivation to learn, their beliefs about themselves as learners and their use of self-regulatory learning strategies. Wolters (2010) concluded that self-regulated learning:

> ...encompasses many important skills, abilities and attitudes that substantially overlap with those viewed as core competencies for the 21st century. The level of conceptual similarity makes some of the core competencies appear nearly synonymous with dimensions of SRL. This conceptual congruity lends support to the critical importance of competencies such as self-direction, adaptability, flexibility, and collaboration (p.18).

The practical implication is that, as Wolters suggested, the research evidence on self-regulated learning can be applied to 21st Century competences. Since self-regulated learning means learners monitoring and controlling their learning practices and outcomes, it is arguably central to learning to the learn competence and to lifelong learning more generally (Dignath & Büttner, 2008). The self-regulated learning literature has recently fed into new research on the co-regulation of learning, which occurs between peers and their teachers, and may provide insights into the assessment of social competences. However, it will be important to interpret the co- and self-regulated learning research bases with an eye to the type of learning that is being regulated – whether narrower cognitive outcomes alone or broader affective outcomes too.

The PISA student questionnaire uses self-report items reliant on their accurate recall and reporting of their thoughts and actions. These items therefore do not provide direct measures and students' responses may differ from what they actually think and do. Furthermore, the items may not support the interpretation developers and researchers attribute to them, particularly between linguistic and cultural contexts (Pepper, 2012a). These issues could help to explain the low correlation between students' reported use of self-regulated learning strategies in the PISA questionnaires (eg 'I try to figure out which concepts I have not understood properly') and their competence as assessed in the PISA tests (OECD, 2004). The problem is that direct measures such as interviews and observations have not been practical for a large-scale survey like PISA (OECD, 2004). However, after three decades of research on self-regulated learning, a review of the academic literature found that a combination of these direct measures are the very ones necessary for valid measurement of students' use of these learning strategies (Boekaerts & Corno, 2005). Following this finding, Panadero, Tapia, and Huertas (2012) sought to self-regulated learning by asking students not only to think aloud during tasks (for a full treatment of this methodology, see Ericsson & Simon, 1993) but also to respond to questionnaire items after the task. They concluded that the questionnaire items assessed students' self-regulation awareness rather than, arguably more importantly, their actual use of self-regulation.

For research or assessment purposes, the self-regulated learning literature therefore suggests a higher profile for classroom or workplace observation and dialogue than for questionnaires and tests (although these instruments may nonetheless provide a basis for observation and dialogue). Furthermore, if self-regulated learning implies self-control informed by accurate self-monitoring, then an important role for self-assessment is also implied. Moreover, this need not be limited to formative assessment. Comparison of self-assessments and expert assessments yields useful information about the apparent accuracy of students' self-monitoring of their learning outcomes (Winne, 1996). Students have little to gain from inflating their self-assessments when they know these will be compared with expert assessments and judged accordingly. As a result, it may be possible to combine the use of questionnaires and tests to assess this potential aspect of the learning to learn competence. Comparison of student performance on self-efficacy questionnaire items and competence-based test items (both already available from PISA instruments) therefore appears to be a particularly promising path for exploration (Greene & Azevedo, 2007). This is one promising way in which self-assessment via questionnaires and assessment via tests could be combined in formative or summative assessments. The Cascade e-assessment is one existing example of a similar comparison:

Cascade is an e-assessment developed through cooperation between researchers in France and Luxembourg that give secondary school students feedback on their use of self-assessment in a computer-based environment. It is therefore a multi-lingual platform designed to provide formative e-assessment that promotes learners' effective use of self-assessment. The assessment comprises different phases. In the first phase, learners respond to multiple choice items each containing a statement and true/false response categories. Importantly, they also rate how certain they are that their true/false response is correct. Learners are then asked to use multimedia information to check their initial true/false response and, based on their findings, give a new true/false response and certainty rating. The computer-based environment makes it possible to capture learners' initial responses, track their multimedia search strategies, compare their subsequent responses and assess their use of multimedia information. Although the first phase could be characterised as Generation 1, the tracking of learners' use of multimedia is consistent with the integrated e-assessment of Generation 3 (although it seems to bypass the flexibility made possible by the adaptive assessment associated with Generation 2). Integrating feedback on learners' responses, perhaps with a third phase enabling them to refine their search strategies, would be consistent with an e-assessment in Generation 4 (Binkley et al., 2010; Jadoul, Merche, Martin, & Latour, Undated; Redecker, 2013).

Performance-based assessment

Performance-based assessment[5] refers to 'authentic' tasks such as exhibitions, experiments, group work, interviews, plays, presentations, projects and role plays in real or realistic contexts. The type of assessment may involve the use of listening and observation or portfolios and diaries to collate information about performances. In contrast with standardised tests, teachers usually have a central role in performance-based assessment, whether for formative or summative purposes. This can increase their workload but also enhance their sense of professionalism and improve outcomes for their students (Stanley, et al., 2009). One particularly benefit of performance-based assessment is that it can be very effective at encouraging and capturing learning processes and outcomes relating to complex constructs such as key competences (Darling-Hammond & Snyder, 2000; Firestone, Mayrowetz, & Fairman, 1998; Looney, 2011). This is valuable because, as Pellegrino and Hilton (2012) argue:

> ...assessment in problem-solving and metacognition should use modeling and feedback techniques that highlight the processes of thinking rather than focusing exclusively on the products of thinking (p.21).

Performance-based assessment could therefore be valid for assessing key competences for formative or summative purposes. Since teachers can observe their students over a period of time and use a range of performance tasks, the reliability of their judgements can be comparable to that of standardised tests – though this comparison also reflects some unreliability in tests (Harlen, 2005).

[5] This term is also occasionally used to contrast constructed-response items with multiple choice items, such as in Claro et al (2012) or Saxton et al (2012) in the section of this review on standardised tests.

Variation in teachers' judgements within and between schools is nonetheless a risk, particularly when under pressure in high stakes assessments such as school-leaving examinations. However, there is evidence that this risk can be managed with training (Looney, 2011) and moderation (Stanley, et al., 2009). Stanley et al identified two forms of moderation. The first was statistical moderation requiring the outcome of an additional assessment to be compared with the performance-based assessment judgement. In practice this usually means a standardised test provides the point of comparison. The problem is that the test may be more limited than the judgement in what it assesses and the comparison may therefore be flawed. The second form of moderation was social moderation based on samples of assessed work, often held in portfolios, within and between schools. The authors reported that a form of social moderation using portfolios in Australia (Queensland) achieved very high reliability. P. Black (2010, p. 10) identified moderation meetings with 'blind marking' to compare, discuss and resolve judgements based on samples of pupils' work, conveniently stored in portfolios, as 'the key to securing intra- and inter-school comparability' of assessment judgements. This could create a positive feedback loop, enhancing teachers' assessment judgements and resulting in acceptable levels of reliability for summative assessment.

In their review of different literatures on creativity, Spencer, et al. (2012) identified five habits of creativity (inquisitive, persistent, imaginative, collaborative and disciplined). Each of these habits comprised three sub-habits. These habits and sub-habits varied according to three dimensions: strength (ie independence), breadth (across contexts) and depth (sophistication and appropriateness). This model of creativity was field trialled with small samples of primary and secondary schools in England by means of a tool for formative assessment in lessons or special activities. The tool was a circle whose circumference was labelled with the habits and the sub-habits. Along the radius, teachers marked off the strength, breadth and depth of learners' creativity against each of the sub-habits. In some cases, learners' assessed themselves with their teachers' simply checking the evidence for their judgements. The evidence itself was stored in recording sheets with a grid of sub-habits on one axis and strength, breadth and depth on the other axis. This was facilitated with exemplar statements. The researchers found that although the teachers thought the 15 sub-habits too onerous to assess, they were necessary for precision. The three dimensions were also too onerous but the researchers found that combining strength and depth in a single dimension, as some teachers suggested and some teachers implemented, had varying degrees of success. The researchers concluded that their study demonstrated a proof of concept for the tool but that further development was required, particularly of assessment criteria and moderation methods – if the tool were to be used across the curriculum and from one year group to the next.

Group work is one type of performance-based assessment currently receiving international attention. Pellegrino and Hilton (2012) report that large-scale performance-based assessment of problem-solving in the 1990s '...revealed an essential tension between the nature of group work and the need to assign valid scores to individual students' (p.131). However, PISA is currently targeting collaborative problem-solving and the Assessment and Teaching of 21st Century Skills project includes work on social competences. These developments rely on '...technology to engage students in interaction, to simulate others with whom students can interact, to track students' ongoing responses, and to draw inferences from those responses' (Ibid., p.132).

Portfolio assessment of key competences and their cross-cutting themes is also receiving sustained international interest in and beyond Europe (Pepper, 2011). A portfolio is a place to store a series of entries compiled over a period of time which are intended to be representative of a learner's progress or to showcase work identified as their best in relation to a set of learning outcomes (Busuttil-Reynaud & Winkley, 2006; Simon & Forgette-Giroux, 2000). Portfolio entries can contain information about learners' performances on tasks in real-life contexts or contexts that are authentic representations of real-life. E-portfolios expand the range of possible formats for entries, so that audio-visual files and internet links can be included. They also provide a flexible format for sharing entries and gaining feedback from teachers and peers. In essence, this is formative assessment. Internet or intranet social networks can facilitate dialogue about entries. In some cases, portfolio or e-portfolio entries may be accompanied by learners' narratives, including explanations and reflections. Teachers may then gain further insights into their students' attitudes to learning and their progress towards learning outcomes (Simon & Forgette-Goroux, 2000).

The use of e-portfolio assessment has been documented in schools in several countries including Austria, Belgium, Bulgaria, France, Greece, Iceland, Portugal, Romania, Turkey, the UK and the USA. Developing and reviewing e-portfolios can help learners to develop digital competence, social competence, learning to learn competence and problem-solving skills. There is some evidence that e-portfolios, like conventional portfolios, can encourage learners to reflect on their peers' work and their own work (Redecker, 2013). This peer and self-assessment, as aspects of a wider pedagogy for formative assessment, is a focus for the next section.

Although portfolios and e-portfolios can be valuable assessment tools, their reliability for summative assessment needs to be addressed. In order to ensure an acceptable level of reliability, Herman et al (1992) had argued that portfolio assessment should be based on the identification of a clearly-stated assessment purpose, guidelines for the selection of entries and criteria for the assessment of portfolio contents. Focusing on the selection of entries, Simon and Forgette-Giroux (2000) developed a content selection framework for the portfolio assessment of problem-solving competency. The authors divided the competency into different aspects and offered examples of potential entries for each aspect. These aspects and entries included:

- **Cognitive** – tests requiring not just recall but analysis, synthesis or evaluation; personal summaries of problem-solving strategies; reporting problem-solving solutions
- **Affective** – a biography in mathematics, inventories of a student's personal reactions to a specific problem, excerpts from a log book, or attitudinal scales and various types of self-reports
- **Metacognitive** – short written or tape-recorded verbal justifications accompanying each entry, comparisons of entries collected at various stages of development, or a personal overview of all the various entries and how their integration reflects their competency
- **Developmental** – a four- to six-level descriptive scale (from limited to full development) could be used holistically at regular intervals, filled out by the student and/or teacher, and included in the portfolio.

Since portfolios can contain a rich variety of information, assessing their content can be a complex and time-consuming process. The technology for 'data mining' e-portfolios is under-development and currently lacks the sophistication for assessing complex learning outcomes such as key competences. Estimates put the technology required for data mining and interpretation of assessments more generally at about five years away. In the meantime, a low-tech strategy for reducing the assessment burden on teachers and other assessors is to mark 'comparative pairs' of portfolios, where assessors simply have to decide which portfolio is better. Although this is predicated on assessors' understanding of the intended learning outcomes, no conventional analytical judgement is made. The research evidence suggests this approach nonetheless results in scoring with relatively high levels of reliability (Redecker, 2013). There is also promising new research that is underway in England using 'adaptive comparative judgements' to assess lower secondary students' mathematical competence[6].

In the eSCAPE project, a six-hour collaborative design workshop replaced school examinations in design and technology (comprising aspects of the key competence – cultural awareness and artistic expression) for students aged 16 in 11 participating schools across England. Students worked individually, but within a group context, to build their design solution. Students were given stage-by-stage assessment instructions and information via a PDA (Personal Digital Assistant). The handheld device also acted as a tool to capture assessment evidence – via video, camera, voice, sketchpad and keyboard. During the six hours, each student developed their design prototype and the PDA provided a record of their progress, interactions and self-reflections. At the end of the assessment, evidence was collated in a short multimedia portfolio loaded onto a secure website. The project resulted in 250 e-portfolios and the reliability of the assessment method was reported as very high (Binkley, et al., 2010).

Pellegrino and Hilton (2012) recently expressed various on-going concerns with using portfolios for summative purposes. These include:

> ...differences in the nature of the interactions reflected in the portfolios for different students or at different times; differences in raters' application of the scoring rubric; differences in the groups with whom individual students have interacted, and other differences. This lack of uniformity in the sample of interpersonal skills included in the portfolio poses a threat to both validity and reliability (p.132).

If, as may be the case across a range of 'real-life' contexts, the performance reflected in a portfolio entry is completed without supervision, a further potential issue is plagiarism. Computer-based assessment may be helpful in identifying potential plagiarism but a judgement will need to be made as to whether a case represents fraud or carelessness (Stanley, et al., 2009).

In summary, like any assessment used for summative purposes, portfolios need standardisation but there are specific issues requiring responses. Perhaps Black (2010) is therefore right to take a more prescriptive stance than Simon and Forgette-Goroux, arguing that portfolio frameworks should specify not only the aims and assessment criteria in advance but also the number and timing of tasks. Stanley, et al. (2009), concerned not to increase the burden on teachers, instead suggest a bare minimum of common assessment tasks, as seen in the USA (New York State) and Australia (Queensland), which could be included in the portfolios. Both Black and Stanley et al identify variation in the assistance students receive from teachers and other as a potential issue. Stanley et al again refer to the example of Australia (Queensland) where clarification of the guidance on what assistance is considered legitimate was issued and Black suggests recording what assistance is given and taking it into account in overall assessment judgements. Black concludes that portfolio frameworks should be trialled to ensure they allow all learners to demonstrate their competences.

Teacher, peer and self-assessment

Crooks (1988) made a distinction between surface-learning and deep-learning. Whereas surface-learning involved passive acceptance and reproduction of ideas, deep-learning involved active interaction, linking ideas and relating new and previous experiences. Such deep-learning

[6] http://www.nuffieldfoundation.org/improving-quality-gcse-mathematics-examinations

appears consistent with the aims of personal fulfilment, active citizenship, social cohesion and employability espoused by the European Reference Framework of key competences. Harlen and James (1997) argued that deep-learning requires a flexible approach to teaching and learning which, strongly informed by formative assessment, responds to learners' different rates of progress and interests. This section focuses on formative assessment, with its concern 'to help learning and foster deeper engagement with it: essentially a pedagogical approach rather than a separate activity added to teaching' (Harlen, 2007, p. 16).

Since the early work on formative assessment (Ramaprasad, 1983) teachers' feedback to their students has been a central concern in the English-language literature. Pellegrino and Hilton (2012) found growing evidence that feedback explaining why something is incorrect is more effective than feedback that identifies errors. Some computer-based assessment software is consistent with this research, offering not only corrections but also explanations (Redecker, 2013). This software could promote not just procedural learning but also conceptual understanding, and is therefore consistent with the aims of the key competences (Pepper, 2011).

The timing of feedback is, however, crucial. For example, if feedback is given before learners have an opportunity for reflection, the opportunity for independent learning is negated. Embedding formative assessment into computer simulations, virtual laboratories and immersive games with sophisticated analysis of and feedback on learners' day-to-day progress would be a major achievement. However, e-assessment technology is still some way from judging and adjusting the timing of feedback or offering the questioning and dialogue emphasised in recent formative assessment literature (P. Black & Wiliam, 2009). A particularly subtle challenge is providing the emotional support that learners may need when faced with challenging activities. To this end, there is currently some developmental work on an intelligent tutoring system that attempts to detect and produce emotions by monitoring learners' verbal and physical communication cues (Redecker, 2013).

Rather than feedback alone, Sadler (1989) emphasised three steps towards formative assessment involving learner's understanding of: the intended learning outcomes; their present position in relation to those outcomes; and, most importantly, how they could close the gap between the two. This influenced P. Black and Wiliam (1998b), who proposed that everyone involved in assessment needed this understanding. In the context of the present review, teachers and learners should therefore develop a shared understanding of learning outcomes relating to key competences. This should include what might count as evidence of the development of a competence in different contexts within, across or beyond subjects. Black and Wiliam's research on 'Assessment for Learning' (AfL) was influential in the UK and, subsequently, several other countries (Third International Conference on AfL, 2009). According to the Assessment Reform Group in the UK, AfL refers to the extensive, systematic use of formative assessment. Although a range of assessment methods can contribute to formative assessment, including even summative tests, assessment by teachers is seen as the most important method, providing detailed insights over time (Assessment Reform Group, 2002). However, peer and self-assessment has an important contribution to make and this will be addressed later in this section.

P. Black and Wiliam (1998a) had found a substantial evidence base for the positive impact of formative assessment on learners' attainment and motivation. Indeed, formative assessment can produce greater learning gains than reductions in class sizes or increases in teachers' content knowledge, and it is also more cost effective to implement (Wiliam & Thomson, 2007). It should be noted, however, that teachers' content knowledge is likely to be an important prerequisite for effective formative assessment. Pellegrino and Hilton (2012) note that although a recent meta-analysis (Kingston and Nash, 2011) identified a smaller effect for formative assessment, this increased with professional development for teachers.

In order to extend the research base beyond the English-language literature, the OECD had commissioned two reviews of formative assessment, one from the literature in French and one from the literature in German. Allal and Lopez (2005) found that, whereas the English-language literature emphasised feedback to correct learners' errors, the French-language literature emphasised feedback for regulating learning. This greater emphasis on the learning process arguably brings formative assessment conceptually closer to the development of learning to learn competence. Indeed, the recent literature in French had focused on peer and self-assessment, including teacher and learner co-construction of assessments. However, the authors conceded that, though theoretically insightful, the literature required more empirical research. Köller (2005) focused on the empirical literature in German but also found a lack of research that had evaluated formative assessment interventions. However, one area of research had evaluated the classroom use of reference norms. This research had found that when teachers compared learners' current performance with their previous performance (an individual reference norm) they learnt more than when teachers compared their current performance with other learners' performance (a social reference norm). This is consistent with the English-language literature on students' use of reference norms (eg Dweck, 1999) and suggests that the findings are generalisable across these different language contexts.

Black and Wiliam (1998b) identified self-assessment as an inevitable aspect of successful formative assessment. To internalise teachers' feedback, learners need to reflect on their learning. However, an explicit focus on self-assessment and peer assessment is a feature of successful formative assessment for children as young as five years old. Although there is evidence that the reliability of learners' assessment judgements greatly varies and that it is therefore not suitable for summative assessment (Redecker, 2013), the value of peer and self-assessment resides in promoting learning through formative assessment. Although peer and self-assessment for the formative purpose is generally honest and trustworthy, it is a competence that needs to be developed over time (perhaps actually as sub-domains of social competence and learning to learn). The result is, however, more committed, effective and independent learners (Black & Wiliam, 1998b).

James, Black, McCormick, and Pedder (2007) identified peer and self-assessment practices as important features of assessment for learning as a strategy to promote 'learning (how) to learn'. The authors also reaffirmed that the underlying competences for these learning practices need to be developed. Since the practices involve learners working alone and with others, they imply developing both learning to learn and social competences. Several studies emphasise the need to gradually train or prepare learners for peer and self-assessment, particularly so that they understand the learning outcomes and assessment criteria (Mills and Glover, 2006). Indeed, following Boud (1995), self-assessment can be defined as an individual identifying learning outcomes and making judgements about the extent to which they have fulfilled those outcomes. In peer assessment, one or more of an individual's peers takes on this role. Peer and self-assessment therefore has the potential to help learners develop their understanding of learning outcomes, particularly those that may be unfamiliar, such as those relating to key competences. Although both peer and self-assessment require gradual preparation, there is a natural synergy. For example, P. Black (2010, p. 5) considers that: 'Peer discussion of specific examples of work should help pupils to understand the criteria by which to assess the quality of their own work'.

One valuable way of providing a structure to encourage learners to identify and act upon assessment information about themselves is simply to encourage them to reflect upon examples of their work. The section on performance-based assessment showed how a framework could provide a basis for selecting entries for portfolios. Such a framework could help learners, with the support of their teachers, to reflect on the learning outcomes, their progress towards them and their next steps. A process of peer and self-assessment for the selection of entries for e-portfolios or conventional portfolios could help learners to develop several key competences including learning to learn, initiative, social competence, communication and digital competence (Pepper, 2011).

Redecker (2013) reports several findings relating to computer-based peer and self-assessment. Learners' peer and self-assessment judgements can be automatically graded but teacher feedback on these judgements may go further, helping learnt to understand what the learning outcomes for key competences mean in practice. The current evidence base is mixed but there is potential for peer and self-assessment through school intranet or internet forums involving social networking, blogs or wikis. These virtual meeting points can enable learners to communicate with one another, carry out collaborative activities together and share content in a range of multimedia formats. Educational games can also provide opportunities for cooperation between learners. Through timely peer or teacher feedback, wikis can promote effective collaboration, peer learning and confidence in peer and self-assessment. These collective and individual outcomes link, most obviously, to communication, social and learning to learn competences.

4. Related policies

Evaluation of policies and practices

The scope of this literature review was the formative and summative assessment of learner's key competences. This is distinct from evaluations of the contribution of education programmes, institutions or systems to learners' development of key competences (Newton, 2007). However, the analysis of assessment results aggregated across a sample or population is an important feature of evaluation in education (Harlen, 2007). Evaluations can inform the initial development, piloting and subsequent refinement of assessment policies and practices. These evaluations need to establish certain success criteria, such as the validity, reliability and equity of assessments (Morris, 2011) prior to implementation so that 'before and after' can be compared.

One broad approach to evaluation concerns the social validity of policies and practices. Social validity can refer to the degree to which a policy or practice has 'social importance or is valued by consumers' such as teachers, parents and learners (Hurley, 2012, p. 164). This approach can be found in the literature on social *competences*, as distinct from the social *validity* of policies and practices associated with the assessment of these competences. The evaluation of social validity can focus on the extent to which the goals, processes or effects of policies and practices satisfy these 'consumers'. There is frequently a consultative approach to such evaluation and a wide range of assessment methods can contribute data to it.

Early evaluations may indicate an 'implementation dip' as teachers adjust to new assessment practices (Fullan, 2001), particularly those associated with constructs that may be unfamiliar, such as key competences. This finding suggests that the management of professional and public expectations is critical for sustaining changes in practices. Indeed, public confidence in assessments is important for the credibility of education systems and, in particular, the currency of qualifications. Furthermore, to promote realistic expectations, government agencies need to raise awareness of unavoidable 'measurement inaccuracy' and 'human error' in educational assessment (Newton, 2005). Although this assessment error is likely to be more pronounced in assessments of complex constructs such as key competences, it could be addressed through further assessment research and assessment training.

To avoid compromising their integrity, it is important to ensure that formative and summative assessments, and the evaluations to which they may contribute, are aligned in overall systems (Black & Wiliam, 2003). Indeed, this is the focus of the current OECD review of assessment and evaluation frameworks that was launched in 2009, which includes a working paper emphasising the importance of coherence between curriculum, assessment and evaluation frameworks (Looney, 2011) and, arguably, teacher education frameworks.

Training and development

The research literature reviewed here indicated that all those involved in educational assessment should develop a shared understanding of the learning outcomes relating to key competences. The depth of this understanding will naturally vary between groups, such as teachers, examiners, students, parents, employers or higher education. However, this understanding is necessary for practices that support the valid use of assessment, whether the technical development of assessment instruments or the lay use of assessment information. Communication with stakeholders about the key competences and their assessment is therefore crucial but formal training and informal development will be necessary for those developing, using or reviewing assessments. The research literature also indicated that, although computer-based assessment may increasingly provide them with support, teachers will have a central role in the assessment of key competences. This is certainly the case for formative assessment and potentially also the case for summative assessment. In general, changes in teachers' practices will be required and these will need to be codified in the teacher education frameworks and supported through teacher education, recognition and progression.

Teacher education frameworks

Teacher education frameworks will need to mirror the key competences in order to ensure that teachers can support their students' development of the key competences. Since the key competences need to be formulated in learning outcomes for learners, it follows that they also need to be formulated in learning outcomes for teachers. However, these learning outcomes will clearly also need to include pedagogic competences such as assessment competence. More specifically, with reference to the possibilities raised by this literature review, this assessment competence could include learning outcomes such as: formulating learning outcomes for students; using a range of techniques for formative and summative assessment; facilitating peer and self-assessment; using assessment information effectively and responsibly; and, attitudes that support these assessment practices.

A unified and comprehensive teacher education framework can support teachers in different roles throughout initial, induction and in-service teacher education (European Commission, 2010). The learning outcomes relating to assessment competence in this framework will vary according to teachers' career stage and role. CEDEFOP (2011a, p. 4) highlighted, that 'systematic upskilling for teachers in new pedagogy and assessment methods, can extend beyond those directly responsible for assessments'. Fullan (2001) had argued that the attendance of head teachers, principals or managers in training sessions can signal the importance of changes in policy and that changes in practice are expected. However, specific training for senior managers can, for example, enable them to lead the development an organisation-wide assessment policy. This policy should encompass a range of assessment techniques with purposes that are clear to teachers and learners (Harlen & Deakin Crick, 2003).

To support a teacher education framework that incorporates key competences and supporting pedagogy, including assessment, there is a general need for ‘intensive capacity building’ (Halász & Michel, 2011). There seems to be little doubt about the breadth of this need. Teachers in the many countries participating in the OECD’s Teaching and Learning International Survey (TALIS) 2008 survey reported focusing on knowledge transmission to learners in passive roles rather than facilitating learners’ active development of competences. The OECD’s report on the survey concluded that since much of the variation in teachers’ practices and beliefs about learning was between teachers rather than schools and countries, teacher education should be targeted to individual teachers (OECD, 2009). This implies finding a manageable way to individually review teachers’ pedagogic beliefs and practices. CEDEFOP (2011a) similarly asserted the need for ‘...not only the right knowledge and skills, but also the appropriate attitudes...’. In consonance with the key competences, teacher education therefore needs to cover not only the requisite knowledge and skills for assessment but also the supporting attitudes. These attitudes are crucial because teachers ultimately exercise discretion over what is actually implemented through their day-to-day practices, including assessment (Bowe, Ball, & Gold, 1992).

Teacher learning communities

Although the formative assessment research literature has influenced education policy in many countries, the extent to which teaching practices reflect research or policy varies widely (Third International Conference on Assessment for Learning, 2009). Teachers’ pedagogic content knowledge in specific subject domains may be an impediment (Pellegrino & Hilton, 2012) and ambiguity in the research literature may also be a factor (Third International Conference on Assessment for Learning, 2009). However, the research literature suggests that professional networks of teachers called teacher learning communities offer hope for translating the research into practice. Wiliam and Thomson (2007) identify teacher learning communities as: permitting the exercise of professional judgement and sustaining it over time; challenging assessment practices in a non-threatening environment; gaining real examples that motivate action; and, interpreting research in specific circumstances. Furthermore, these communities could also foster and sustain changes in teachers’ summative assessment practices.

With reference to summative assessment, Wiliam and Thomson’s (Ibid.) emphasis on gaining real examples is particularly relevant. The research literature suggests that portfolios of students’ work could be a useful resource for sharing examples of assessing key competences. P. Black (2010, p. 8) argued that: ‘a portfolio can serve as each pupil’s own record of their achievements, and can also be a documentary basis for comparison, between teachers in the same school, and between different schools, to ensure comparability in their standards’. As such, portfolios were 'the optimum way' of allowing evidence of competence to be communicated within a rigorous assessment framework. Since portfolios can represent a collection of entries demonstrating learners’ development of key competences, they can contain information gathered using a range of information sources and assessment techniques. Just as portfolios assessment of key competences could serve formative or summative purposes, portfolios could also provide a focus for teachers to review and develop their own formative and summative assessment practices.

Continuing with the value of examples more generally, Sadler (1989) argued that key examples should be selected to identify high quality assessment. These ‘exemplars’ should be typical of each level of attainment defined in the assessment; a range of exemplars should be to stimulate creativity rather than encouraging conformity; and, new examples should be generated in order to be relevant and interesting. Importantly, these exemplars could provide insights into not only learning outcomes but also learning processes associated with key competences. Assessment exemplars could therefore help teachers to develop their assessment practices and wider pedagogy. However, exemplars could be particularly useful for standardising teachers' assessment for summative purposes – hence Sadler's emphasis on exemplars for each level of attainment or, rather, each level of competence.

Both Fullan (2001) and Wiliam (2007) identified the need for pressure (or 'accountability') and support for changes in teachers' practices through teacher learning communities. This pressure and support can provide the scrutiny and feedback that is necessary for the implementation of changes in assessment practices. Furthermore, where there is constructive discussion of assessment practices for broader learning outcomes such as key competences, there are positive impacts on learners' effort and attainment (Deakin Crick, 2008). Wiliam (Ibid.) also identifies the need for a gradual approach, with each teacher implementing no more than two or three assessment techniques any one time to avoid a loss of routine and disorder. Furthermore, he argued that the approach should be flexible, since techniques that work in one context may not work in others or need adjustment (see also Resnick, Spillane, Goldman, & Rangel, 2010). Lastly, teachers should be able to choose the techniques they use on the basis of their preferences. For example, some teachers will feel suitable for prepared for some techniques (eg orchestrating whole class discussions) than for others (eg facilitating work in small groups).

As in the research on teacher learning communities, Gardner, Harlen, Hayward, and Stobart (2011) also emphasis the value of teachers working together to develop their own assessment practices with the necessary support. The authors summarise the findings of two contrasting research projects which both sought to develop teachers' formative assessment practices. In one project, the researchers trained teachers in practices that the researchers had already developed. This 'transmission' approach did lead to changes in teachers' practices that were consistent with the aims of the training. However, there was evidence that teachers had gained a procedural understanding of the assessment practices but lacked a conceptual understanding of the underlying principles. This resulted in some confusion about the methods and benefits of formative assessment, reducing the likelihood of effective and sustained change in teachers' practices. In the other project, the researchers created a network of teachers in two localities and worked with them to develop effective formative assessment practices, evidence and materials for wider use. This 'constructivist' approach enhanced teachers' understanding of the principles underlying the assessment practices and increased their commitment to sustained changes in their practices.

Although this research suggests that the constructivist approach has more potential than the transmission approach, the literature cautions that it is very resource-intensive, particularly in terms of teachers' and researchers' time. In particular, teachers need time for discussion, reflection and planning. However, the constructivist approach also has the benefit of being more consistent with the type of learner-centred activities associated with key competences (Gordon, et al., 2009). Furthermore, experiencing a learner-centred approach including formative assessment and understanding the principles and rationale positions teachers to take the same approach to their own pedagogic practices, and to be become effective advocates for changes in assessment practices.

Encouragingly, findings from the OECD's TALIS 2008 survey suggest that teachers would welcome working with researchers and other teachers to develop their own assessment practices. Specifically, TALIS found that teachers in many countries saw collaborative research as an effective form of professional development. However, few teachers actually participated in collaborative research. Furthermore, there was a lack of recognition for effective or innovative teaching (OECD, 2009). Teacher learning communities could provide a forum for collaborative research and the informal recognition of effectiveness and innovation. However, support for innovation more generally would therefore need to be coupled with formal recognition through performance reviews, pay and conditions.

5. Conclusions

Assessment is an important agent for change in education. Whilst it lags behind the curriculum, teaching and learning will be hampered. Policy makers and practitioners therefore need to give due attention to the assessment of key competences. This should take the shape of a development process involving the specification of learning outcomes followed by the development of assessment and accompanied by the alignment of related policy areas.

A development process:

Specify learning outcomes

The European Reference Framework provides broad definitions of key competences, for each Member States to interpret them in the specific circumstances of their education system. The literature indicates that this means specifying them as learning outcomes derived from the requirements of real-life contexts. Policy makers need to ensure that these learning outcomes achieve a balance between central prescription and local judgement, which will reflect the assessment purpose. However, the learning outcomes should certainly include not only the knowledge but also the skills and supporting attitudes required in the range contexts envisaged by the curriculum. These learning outcomes will then provide an adequate basis for the development of assessment instruments and practices for the key competences.

Develop assessments

Whilst summative assessments set the direction, formative assessments inform next steps in teaching and learning. It is therefore important to develop assessments of key competences that serve each of these purposes. Specific features designed into a range of conventional or computer-based assessment methods can help to create real or authentic contexts requiring learners to develop and demonstrate the necessary knowledge, skills and attitudes. Teachers' development of formative assessment practices for key competences can position them to contribute their expertise to summative assessment – with safeguards such as external moderation. The process of peer and self-assessment is associated with learners' development of outcomes particularly related to social competence, learning to learn, initiative and entrepreneurship, the constructive management of feelings and decision taking. Careful interpretation of the literature on self- and co-regulated learning may provide further insights into the assessment of these competences.

Align related policies

In the same way that curriculum and assessment need to be aligned with one another, a number of other policies also need to be aligned if key competences are to be translated from policy into practice. This includes incorporating assessment competence into the teacher education framework and developing the competence through teacher education and teacher learning communities. It also includes research and evaluation of assessments of key competences, according to criteria including validity, reliability and equity. Since the successful implementation of key competences ultimately means learners' developing their key competences, assessment has an important contribution to make to the evaluation, and subsequent refinement, of the full range of policies and practices necessary for implementation. However, the 2013 update of this literature review finds no evidence of a surge in rigorous research connected with the development of new assessments of the key competences promoted in EU policies. This suggests that educational assessment research requires additional stimulation by the European Commission and EU Member States.

6. References

Alexander, P. A., & Winne, P. H. (2006). *Handbook of Educational Psychology*. New Jersey: Lawrence Erlbaum.

Allais, S. M. (2007). Why the South African NQF Failed: lessons for countries wanting to introduce national qualifications frameworks. *European Journal of Education, 42*(4), 523-547.

Allal, L., & Lopez, L. M. (2005). Formative assessment of learning: A review of publications in French *Formative assessment: Improving learning in secondary classrooms*. Paris: OECD.

Ananiadou, K., & Claro, M. (2009). *21st Century skills and competences for new millennium learners in OECD countries*. Paris.

Assessment of Key Competences in initial education and training: Policy Guidance. (2012).

Bennett, R. E. (2010). Technology for Large-scale Assessment-

In P. Peterson, E. Baker & B. McGaw (Eds.), *International Encyclopedia of Education* (3 ed., Vol. 8, pp. 48-55). Oxford: Elsevier.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., & Rumble, M. (2010). *Draft White Paper 1: Defining 21st Century Skills*.

Black, B. (2012). An Overview of a Programme of Research to Support the Assessment of Critical Thinking. *Thinking Skills and Creativity, 7*(2), 122-133.

Black, P. (1998). *Testing, friend or foe?: the theory and practice of assessment and testing*: Falmer Press.

Black, P. (2010). *Assessment of and for learning: improving the quality and achieving a positive interaction*: King's College London.

Black, P., Burkhardt, H., Daro, P., Lappan, G., Pead, D., & Stephenson, M. (2011). *High-stakes Examinations that Support Student Learning: Recommendations for the design, development and implementation of the SBAC assessments*: International Society for Design and Development in Education Working Group on Examinations and Policy.

Black, P., & Wiliam, D. (1998a). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7 - 74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. In K. s. C. London (Ed.). London: GL Assessment.

Black, P., & Wiliam, D. (2003). ‘In praise of educational research’: formative assessment. *British Educational Research Journal, 29*(5), 623-637.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31.

Boekaerts, M., & Corno, L. (2005). Self-Regulation in the Classroom: A Perspective on Assessment and Intervention. *Applied Psychology, 54*(2), 199-231.

Bowe, R., Ball, S. J., & Gold, A. (1992). *Reforming education and changing schools: case studies in policy sociology*: Routledge.

Brennan, R. L. (2006). *Educational measurement*: Praeger Publishers.

Bunderson, V. C., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 367-407). New York: Macmillan.

Busuttil-Reynaud, G., & Winkley, J. (2006). *e-Assessment Glossary (Extended)*: HEFCE.

CEDEFOP. (2008a). *The shift to learning outcomes: Conceptual, political and practical developments in Europe*. Luxembourg.

CEDEFOP. (2008b). *Terminology of European education and training policy: A selection of 100 key terms*. Luxembourg.

CEDEFOP. (2011a). *Briefing note: When defining learning outcomes in curricula, every learner matters* Thessaloniki: European Centre for the Development of Vocational Training (Cedefop).

CEDEFOP. (2011b). *Using learning outcomes*. Luxembourg: Publications Office of the European Union.

Claro, M., Preiss, D. D., San Martin, E., Jara, I., Hinostroza, J., Valenzuela, S., et al. (2012). Assessment of 21st century ICT skills in Chile: Test design and results from high school level students. *Computers & Education, 59*(3), 1042-1053.

Commission, E. (2009). *Key competences for a changing world: Draft 2010 joint progress report of the Council and the Commission on the implementation of the "Education & Training 2010 work programme"*.

Commission, E. (2010). *Developing coherent and system-wide induction programmes for beginning teachers: a handbook for policy makers*. Brussels.

Commission, E. (2012). *Assessment of Key Competences in initial education and training: Policy Guidance*. Strasbourg.

Consortium, P. (2010). *PISA 2012 field trial problem solving framework: draft subject to possible revision after the field trial*.

Crooks, T. J. (1988). The Impact of Classroom Evaluation Practices on Students. *Review of educational research, 58*(4), 438-481.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education, 16*(5–6), 523-545.

Deakin Crick, R. (2008). Key competencies for education in a European context: narratives of accountability or care. *7, 3*(311-318).

Delors, J. (1996). *Learning: the treasure within*. Paris: OECD.

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning, 3*(3), 231-264.

Dweck, C. S. (1999). *Self-theories : their role in motivation, personality, and development*. Philadelphia, PA: Psychology Press Ltd.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* Cambridge, MA, US: The MIT Press.

Eseryel, D., Ifenthaler, D., & Ge, X. (2013). Validation Study of a Method for Assessing Complex Ill-Structured Problem Solving by Using Causal Representations. *Educational Technology Research and Development, 61*(3), 443-463.

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

Eurydice. (2012). *Developing key competences at school in Europe: Challenges and opportunities for policy. Eurydice report.* Luxembourg: Publications Office of the European Union.

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-Based Assessment and Instructional Change: The Effects of Testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20*(2), 95-113.

Fredriksson, U., & Hoskins, B. (2008). *Learning to learn : What is it and can it be measured?* Ispra: European Commission JRC.

Fullan, M. (2001). *The new meaning of educational change*: Teachers College Press.

Gardner, J., Harlen, W., Hayward, L., & Stobart, G. (2011). Engaging and Empowering Teachers in Innovative Assessment Practice. In R. Berry & B. Adamson (Eds.), *Assessment Reform in Education: Policy and Practice*: Springer.

Gipps, C. V. (1994). *Beyond Testing: Towards a Theory of Educational Assessment*: Falmer Press.

Gordon, J., Halasz, G., Krawczyk, M., Leney, T., Michel, A., Pepper, D., et al. (2009). *Key competences in Europe: Opening doors for lifelong learners across the school curriculum and teacher education*. Warsaw.

Greene, J. A., & Azevedo, R. (2007). A Theoretical Review of Winne and Hadwin's Model of Self-Regulated Learning: New Perspectives and Directions. *Review of Educational Research, 77*(3), 334-372.

Haggerty, K., Elgin, J., & Woolley, A. (2012). *Social-Emotional Learning Assessment Measures for Middle School Youth*: Social Development Research Group, University of Washington.

Halász, G., & Michel, A. (2011). Key Competences in Europe: interpretation, policy formulation and implementation. *European Journal of Education, 46*(3), 289-306.

Harlen, W. (2005). Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20*(3), 245-270.

Harlen, W. (2007). Assessment of Learning Available from http://kcl.eblib.com/patron/FullRecord.aspx?p=433641

Harlen, W., & Deakin Crick, R. (2003). Testing and Motivation for Learning. [doi: 10.1080/0969594032000121270]. *Assessment in Education: Principles, Policy & Practice, 10*(2), 169-207.

Harlen, W., & James, M. (1997). Assessment and Learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice, 4*(3).

Health, C. o. E. f. P. (2011). *Competencies and Learning Objectives*. Washington.

Hurley, J. J. (2012). Social Validity Assessment in Social Competence Interventions for Preschool Children: A Review. *Topics in Early Childhood Special Education, 32*(3), 164-174.

Jadoul, R., Merche, J.-F., Martin, R., & Latour, T. (Undated). *TAO / Cascade*.

James, M., Black, P., McCormick, R., & Pedder, D. (2007). Promoting learning how to learn through assessment for learning. In M. James, R. McCormick, P. Black, P. Carmichael, M.-J. Drummond, A. Fox, J. MacBeath, B. Marshall, D. Pedder, R. Procter, S. Swaffield, J. Swann & D. Wiliam (Eds.), *Improving learning how to learn: Classrooms, schools and networks*. Abingdon, Oxon: Routledge.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* Praeger Publishers.

Kerka, S. (1998). Competency-Based Education and Training: Myths and Realities., from http://www.calpro-online.org/eric/textonly/docgen.asp?tbl=mr&ID=65

Köller, O. (2005). Formative assessment in classrooms: a review of the empirical German literature *Formative assessment: Improving learning in secondary classrooms*. Paris: OECD.

Koretz, D. (2005). Alignment, High Stakes, and the Inflation of Test Scores. *Yearbook of the National Society for the Study of Education, 104*(2), 99-118.

Kupiainen, S., Hautamäki, J., & Rantanen, P. (2008). EU Pre-Pilot on Learning to Learn, Report on the compiled data. *University of Helsinki, Centre for Educational Assessment, 1190*, 001-001.

Learning, T. I. C. o. A. f. (2009). *Position Paper on Assessment for Learning*, Dunedin, New Zealand.

Leney, T., Gordon, J., & Adam, S. (2008). *The Shift to Learning Outcomes; Policies and Practices in Europe, Cedefop, 2008* Thessalonika.

Looney, J. (2011). *Alignment in Complex Education Systems*. Paris: OECD.

Mays, C. (1995). *Performance standards in education: Report of the seminar*. Paris: OECD.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. New York: Macmillan Publishing Co.

Mislevy, R. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*(4), 439-483.

Morris, A. (2011). *Student Standardised Testing*. Paris: OECD.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice, 14*(2), 149-170.

OECD. (2004). *Learning for tomorrow's world: first results from PISA 2003*. Paris: OECD.

OECD. (2009). *Creating Effective Teaching and learning environments first results from TALIS*. Paris: OECD.

OECD. (2010). *Mathematics Teaching and Learning Strategies in PISA*.

OJEU. (2006). *Recommendation of the European Parliament and of the Council of 18 December 2006 on key competences for lifelong learning (2006/962/EC)*. Retrieved from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:394:0010:0018:en:PDF.

Panadero, E., Tapia, J. A., & Huertas, J. A. (2012). Rubrics and Self-Assessment Scripts Effects on Self-Regulation, Learning and Self-Efficacy in Secondary Education. *Learning and Individual Differences, 22*(6), 806-813.

Pellegrino, J., & Hilton, M. (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington D.C.: National Academy of Sciences.

Pepper, D. (2007). *Assessment for disabled students: an international comparison*. London: QCA.

Pepper, D. (2011). Assessing Key Competences across the Curriculum — and Europe. *European Journal of Education, 46*(3), 335-353.

Pepper, D. (2012a). *Are you thinking what I'm thinking? Students' interpretations of PISA mathematical self-efficacy and self-concept items.* Paper presented at the Association of Educational Assessment - Europe, Berlin.

Pepper, D. (2012b). *Thematic Working Group 'Assessment of Key Competences': Literature review, Glossary and examples.* Brussels: European Commission Directorate-General for Education and Culture.

Popham, W. J. (2001). Teaching to the Test? *Helping All Students Achieve 58*(6), 16-20.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*(1), 4-13.

Redecker, C. (2013). *The Use of ICT for the Assessment of Key Competences*: Institute for Prospective and Technological Studies, Joint Research Centre.

Redecker, C., & Johannessen, O. (2013). Changing Assessment -- Towards a New Assessment Paradigm Using ICT. *European Journal of Education, 48*(1), 79-96.

Resnick, L. B., Spillane, J. P., Goldman, P., & Rangel, E. S. (2010). Implementing innovation: from visionary models to everyday practice. In C. f. E. R. a. Innovation (Ed.), *The Nature of Learning: Using research to inspire practice* (pp. 285-315). Paris: OECD Publishing.

Rychen, D. S., & Salganik, L. H. (2003). *Definition and Selection of Competencies: Theoretical and Conceptual Foundations (DeSeCo). Summary of the final report: "Key Competencies for a Successful Life and a Well-Functioning Society"*. Paris.

Sadler, D. R. (1987). Specifying and Promulgating Achievement Standards. *Oxford Review of Education, 13*(2), 191 - 209.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144.

Sanchez-Nunez, M., Fernandez-Berrocal, P., & Latorre, J. (2013). Assessment of Emotional Intelligence in the Family: Influences Between Parents and Children on Their Own Perception and That of Others. *Family Journal, 21*(1), 65-73.

Saxton, E., Belanger, S., & Becker, W. (2012). The Critical Thinking Analytic Rubric (CTAR): Investigating Intra-Rater and Inter-Rater Reliability of a Scoring Mechanism for Critical Thinking Performance Assessments. *Assessing Writing, 17*(4), 251-270.

Schneider, M., & Stern, E. (2010). The cognitive perspective on learning: ten cornerstone findings. In C. f. E. R. a. Innovation (Ed.), *The Nature of Learning: Using research to inspire practice* (pp. 69-90). Paris: OECD Publishing.

Simon, M., & Forgette-Giroux, R. (2000). Impact of a Content Selection Framework on Portfolio Assessment at the Classroom Level. *Assessment in Education: Principles, Policy & Practice, 7*(1), 83-100.

Spencer, E., Lucas, B., & Claxton, G. (2012). *Progression in Creativity: Developing new forms of assessment*. Newcastle: CCE.

Stanley, G., MacCann, Gardner, J., Reynolds, L., & Wild, I. (2009). *Review of teacher assessment: Evidence of what works best and issues for development*: University of Oxford.

Stobart, G. (2008a). Attitudes and assessment. [Editorial]. *Assessment in Education: Principles, Policy & Practice, 15*(1), 1 - 2.

Stobart, G. (2008b). *Testing times: The uses and abuses of assessment*. Oxfordshire: Routledge.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice*. Norwell, MA: Kluwer.

Wiliam, D. (2007). Content then process: Teacher learning communities in the service of formative assessment. In D. Reeves (Ed.), *Ahead of the curve: the power of assessment to transform teaching and learning* (pp. 182-204). Bloomington, Indiana: Solution Tree.

Wiliam, D., & Black, P. (1996). Meanings and Consequences: A Basis for Distinguishing Formative and Summative Functions of Assessment? *British Educational Research Journal, 22*(5), 537-548.

Wiliam, D., & Thomson, M. (2007). Integrating assessment with instruction: what will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: shaping teaching and learning*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences, 8*(4), 327-353.

Wolf, A. (2001). Competence-based assessment. In J. Raven & J. Stephenson (Eds.), *Competence in the Learning Society* (pp. 453-466). New York: Peter Lang.

Wolters, C. A. (2010). *Self-Regulated Learning and the 21st Century Competencies*