

Report from the FP7 project:

Assess Inquiry in Science, Technology and Mathematics Education



ASSISTME

Report on current state of the art in formative and summative assessment in IBE in STM

Leibniz-Institute for Science and Mathematics Education (IPN)
in cooperation with Pearson Education International (PEI)

Delivery date	15.10.2013
Deliverable number	D 2.4
Lead participant	Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany
Contact person	Silke Rönnebeck (roennebeck@ipn.uni-kiel.de) – IPN (Part I) Rose Clesham (rose.clesham@pearson.com) – PEI (Part II)
Dissemination level	PU

Report from the FP7 project:

Assess Inquiry in Science, Technology and Mathematics Education



ASSISTME

Report on current state of the art in formative and summative assessment in IBE in STM – *Part I* –

Sascha Bernholt, Silke Rönnebeck, Mathias Ropohl,
Olaf Köller, & Ilka Parchmann
with the assistance of Hilda Scheuermann & Sabrina Schütz

Delivery date	15.10.2013
Deliverable number	D 2.4
Lead participant	Leibniz Institute for Science and Mathematics Education (IPN), Kiel, Germany
Contact person	Silke Rönnebeck (roennebeck@ipn.uni-kiel.de)
Dissemination level	PU

Table of Contents

SUMMARY	4
1. INTRODUCTION	5
2. THEORETICAL BACKGROUND	7
2.1 IBE in STM.....	7
2.2 Assessment in education	11
2.2.1 Characteristics of assessment systems	12
2.2.2 Summative and formative assessment	13
2.2.3 Characteristics of formative assessment	14
2.2.4 Assessment methods and techniques	14
2.2.5 Formative assessment – barriers and support.....	15
2.2.6 Links between formative and summative assessment.....	17
2.2.7 Assessment and inquiry	19
3. OBJECTIVES OF THE LITERATURE REVIEW	20
4. PROCEDURE OF THE LITERATURE REVIEW	22
4.1 Searches in data bases.....	22
4.2 Searches in relevant journals	27
4.3 Searches in reference lists.....	28
4.4 Final extract	28
4.5 Expert survey.....	33
5. RESULTS OF THE LITERATURE REVIEW	37
5.1 Which aspects of IBE are emphasized or researched in the study?	38
5.1.1 Diagnosing problems/ Identifying questions.....	38
5.1.2 Searching for information	39
5.1.3 Considering alternative or multiple solutions/ searching for alternatives/ modifying designs	40
5.1.4 Creating mental representations	42
5.1.5 Constructing and using models.....	43
5.1.6 Formulating hypotheses/ researching conjectures	44
5.1.7 Planning investigations.....	46
5.1.8 Constructing prototypes	47
5.1.9 Finding structures or patterns.....	48
5.1.10 Collecting and interpreting data/ evaluating results	49

5.1.11 Constructing and critiquing arguments or explanations, argumentation, reasoning, and using evidence.....	51
5.1.12 Communication/ debating with peers	54
5.1.13 Searching for generalizations.....	55
5.1.14 Dealing with uncertainty	56
5.1.15 Problem solving.....	56
5.1.16 IBE and inquiry process skills in general.....	57
5.1.17 Knowledge/ achievement/ understanding	59
5.1.18 Further aspects focused on or assessed by the studies.....	60
5.2 Which types of assessment are employed in the study?	61
5.2.1 Science.....	62
5.2.2 Technology.....	75
5.2.3 Mathematics	78
6. PERSPECTIVES.....	81
7. APPENDIX.....	84
7.1 Frameworks of inquiry competences and/or assessment.....	84
7.2 Computer-supported inquiry learning environments and computer-based assessment tools	87
7.3 Assessment instruments.....	91
REFERENCES.....	95
FIGURES	120
TABLES.....	121

Summary

The EU project 'Assess Inquiry in Science, Technology and Mathematics Education' (ASSIST-ME) investigates formative and summative assessment methods to support and improve inquiry-based approaches in European science, technology and mathematics (STM) education.

In the first step of the project, a literature review was conducted in order to gather information about the current state of the art in formative and summative assessment in inquiry-based education (IBE) in STM. Searches were conducted in data bases, in the most important journals in the field of STM education, and in the reference lists of relevant publications. This report describes the search strategies used in detail and presents the results of the empirical studies described in the found publications in this field.

Especially in science education, numerous publications were found by the search strategies whereas in technology and mathematics education the numbers of publications are much lower. On the one hand, the chosen keywords and search strategies might be a reason. On the other hand, the research foci of the disciplines might be another reason.

The results of the literature review indicate that only a small number of empirical studies have simultaneously investigated both the use of formative and summative assessment in the learning of inquiry in STM and the influence of this form of assessment on the learning of inquiry in STM. Moreover, most of the studies did not assess inquiry directly, but rather knowledge, understanding or attitudes. Nevertheless, there are examples of methodological approaches which illustrate the successful application of several assessment instruments and explain their advantages or disadvantages.

1. Introduction

The overall rationale for ASSIST-ME is that assessment should enhance learning in STM education. It is well acknowledged that assessment is one of the most important drivers in education and is a defining aspect of any educational system. However, it can be observed that instruction – and especially innovative approaches to instruction – and assessment very often are not aligned. Evaluations of inquiry-based teaching and learning are often based on traditional summative assessments of content knowledge that need not necessarily show achievement gains. Stieff (2011), for instance, found that using an inquiry curriculum in combination with a visualization tool yielded only small to moderate gains in a summative achievement test but significantly increased students' representational competence. In recent years, however, the need to align curriculum, instruction and assessment has become more and more obvious.

One major objective of ASSIST-ME is to develop a set of assessment methods suitable for enhancing IBE with regard to STM related competences. Based on these methods, strategies for the formative and summative assessment of competences in STM will then be identified that are adaptable to various European educational systems (Dolin, 2012). The research into the formative and summative assessment of competences relevant to IBE in STM will be based on an understanding of the concept of competences (both domain-specific and transversal), of IBE and of formative versus summative assessment.

In order to achieve this understanding, work package 2 (WP 2) in the ASSIST-ME project carried out a review of the existing research literature on the formative and summative assessment of IBE in STM. The aim of this review is to summarize what we know about the formative and summative assessment of competences in STM – with a special focus on IBE – and to identify methods that can improve student outcomes. Part II of the review (conducted by Pearson Education International) deals specifically with computer-based assessment and the use of information and communication technology (ICT) tools.

One major challenge for the literature review was that the field of interest is not clearly defined. With respect to science education, there is still disagreement among researchers and educators about what features define the instructional approach of IBE (Furtak, Shavelson, Shemwell, & Figueroa, 2012; Hmelo-Silver, Duncan, & Chinn, 2007). A rich vocabulary is used to describe inquiry-based approaches to teaching and learning, such as inquiry-based teaching and learning, authentic inquiry, model-based inquiry, modelling and argumentation, project-based science, hands-on science, and constructivist science (Furtak, Seidel, Iverson, & Briggs, 2012). These approaches might include characteristics of IBE to a varying degree but they are not necessarily synonyms of IBE. The situation gets even more complicated because, e.g. in the US, the field of science education has moved away from using the term inquiry and now calls it “scientific and engineering practices” (National Research Council, 2012). Moreover, the definitions of IBE or inquiry-based approaches to teaching and learning differ between the three domains of science, technology, and mathematics (see D 2.5).

A similar situation is described by Black and Wiliam (1998) in their meta-analysis of formative assessment in the classroom. They state that a literature search carried out by entering keywords in the ERIC data base was inefficient for their purposes because of “a lack of terms used in a uniform way” (Black & Wiliam, 1998, p. 8). As in the case of IBE, formative assessment may be described with a variety of names, such as classroom evaluation, curriculum-based assessment, feedback or formative evaluation (Black & Wiliam, 1998). With respect to the literature review of WP 2, this had consequences for the search strategies. They will be described in chapter 4. Procedure of the literature review.

In this report, some background information about inquiry-based approaches (see 2.1 IBE in STM) and formative and summative assessment in STM education (see 2.2 Assessment in education) will first be given. With respect to IBE, this report puts a special focus on the aspects and definitions of inquiry competences found in the literature and used by previous EU projects. These definitions form the basis for the data base searches and the analysis of results. A detailed description of the definition of IBE in the three domains is given in deliverable D 2.5 ‘A definition of inquiry-based STM education and tools for measuring the degree of IBE’.

In the paragraphs about the formative and summative assessment in STM, first, the concepts are briefly defined. Afterwards, their role in and their influence on STM teaching and learning and the factors that might support or impede their employment are discussed. The main part of the report, however, deals with the results of the search for empirical studies which have investigated the effects of IBE and assessment methods employed to assess and measure these effects. After describing the methodology of the literature search in section 4, the aspects of inquiry which are assessed in STM education are discussed, along with the formative and summative assessment methods which are used (see section 5). The results of a literature search which focussed on the computer-based assessment of IBE in STM that was performed by the ASSIST-ME partner Pearson are presented in part II of this document.

2. Theoretical background

2.1 IBE in STM

According to Anderson (2002) – whose definition forms the basis of the ASSIST-ME application – inquiry-based STM education includes students' involvement in questioning, reasoning, searching for relevant documents, observing, conjecturing, data gathering and interpreting, investigative practical work and collaborative discussions, and working with problems from and applicable to real-life contexts. Whereas these characteristics generally apply to all three subject areas – science, technology and mathematics – the ASSIST-ME application explicitly acknowledges that various meanings and forms of inquiry are possible in different disciplines and need to be addressed in the project. These different approaches to inquiry, however, need to be aligned with a general definition of the construct that will be produced by the project and form deliverable D 2.5 'A definition of inquiry-based STM education and tools for measuring the degree of IBE'.

Looking at the literature, it seems that IBE has mainly been investigated in the field of science education. Performing a basic search in the Web of Science for the period 1996 to 2012 using the keywords 'science/scientific' crossed with 'teaching', 'learning', 'education' and 'instruction' and crossed with 'inquiry' resulted in 2034 entries. Replacing 'science/scientific' by 'mathematics' reduced the number of results to 218, by 'technology' to 567 with most of the entries in technology dealing with the use of technology in inquiry-based (science) education and not with inquiry in technology education (search performed in November 2012).

This might partly be due to the fact that in mathematics and technology the term 'inquiry' is not common and thus inquiry-based approaches go under different names. In the case of mathematics, for instance, teaching approaches and learning theories that include characteristics of mathematical inquiry are – as named in the ASSIST-ME application – inquiry mathematics (Cobb, Wood, Yackel, & McNeal, 1992), open approach lessons (Nohda, 2000), and problem-centred learning (Schoenfeld, 1985). The Fibonacci-project (Artigue & Baptist, 2012) extends this list towards the Dutch approach of realistic mathematics education (Freudenthal, 1973) and the French theory of didactical situations (Brousseau & Balacheff, 1997). Moreover, they include the Swiss concept of dialogic learning (Gallin, 2012). In dialogic learning, instead of immediately trying to solve the problem, students should instead focus on exploring the question and related aspects in depth, thus relating it to their own world. A decisive factor for dialogic learning is that feedback is provided to the students during the exploration process (Gallin, 2012). Another approach of inquiry in mathematics education is the concept of 'problem-based learning' that is also mentioned in the well-known Rocard report (European Commission, 2007, p. 9): "In mathematics teaching, the education community often refers to 'Problem-Based Learning (PBL)' rather than to IBE. In fact, mathematics education may easily use a problem-based approach while, in many cases, the use of experiments is more difficult. PBL describes a learning environment where problems drive the learning." Problem- or project-based learning is also used in technology education. The closest connection to inquiry, however, is provided by approaches to teaching and

learning using the concept of design that bears close resemblance to IBSE. The main difference is seen in the fact that “‘doing’ holds a central position in all aspects relating to both technology and technological literacy” (Ingerman & Collier-Reed, 2011, p. 138). Action is seen as an important component of technological literacy especially in view of “the need to be able to ‘select, properly apply, then monitor and evaluate appropriate technologies’ ([Hayden, 1989] p. 231 – emphasis added) in a given situation. In this way, technological literacy in a situation is constituted through actions” (Ingerman & Collier-Reed, 2011, p. 138; see also Vries & Mottier, 2006).

A lot of former and on-going EU projects in the field of IBE (e.g. Mind the Gap, S-TEAM, ESTABLISH and Fibonacci) have based their understanding of IBSE on a definition from Linn, Davis and Bell (2004, p. 4):

“[inquiry is] the intentional process of diagnosing problems, critiquing experiments, and distinguishing alternatives, planning investigations, researching conjectures, searching for information, constructing models, debating with peers and forming coherent arguments”.

In IBSE, students should be able to identify relevant evidence and use critical thinking and logical reasoning to reflect on its interpretation. They should develop the skills necessary for inquiry and the understanding of science concepts through their own activity and reasoning. This involves exploration and hands-on experiments (Fibonacci project, not reported). IBSE should foster critical and creative minds, it should encourage students to engage in, explore, explain, extend, and evaluate real-life situations in collaboration and cooperation with their peers (PRIMAS project, 2010). It is thus based on a specific understanding of learning as deliberately involving linguistic processes such as argumentation (Dolin, 2012) and requires students to take charge of their own learning in order to achieve genuine understanding (Harlen, 2009). The ESTABLISH project dissected the definition of Linn, Davis and Bell (2004) and articulated nine aspects or elements of inquiry (ESTABLISH project, 2011):

1. Diagnosing problems
2. Critiquing experiments
3. Distinguishing alternatives
4. Planning investigations
5. Researching conjectures
6. Searching for information
7. Constructing models
8. Debating with peers
9. Forming coherent arguments

These aspects can be regarded as inquiry competences. Because of their prominent role in European IBE projects, it was decided to use them as the foundation of the ASSIST-ME definition of IBE. Comparing them with other definitions of inquiry-based science education (e.g. American Association for the Advancement of Science, 2009; Hmelo-Silver, Duncan, & Chinn, 2007; Kessler & Galvan, 2007; National Research Council, 1996, National Research Council, 2012) and with definitions of inquiry-based approaches in mathematics (Artigue & Baptist, 2012; Artigue, Dillon, Harlen, & Léna, 2012; Hunter & Anthony, 2011; Kwon, Park, & Park, 2006) and technology education (American Association for the Advancement of Science, 2009; National Research

Council, 2012) however, the need to elaborate on and extend the list of aspects became clear.

A characteristic feature of technology education, for instance, is that knowledge, experience and resources are applied purposefully to create products and processes that meet human needs (Davis, Ginns, & McRobbie, 2002). Thus, inquiry-based approaches in technology education often focus on the design process as a process of problem solving consisting of

1. defining the problem and identifying the need,
2. collecting information,
3. introducing alternative solutions,
4. choosing the optimal solution,
5. designing and constructing a prototype, and
6. evaluating and correcting the process (Doppelt, 2005).

Differences and similarities between inquiry-based science and mathematics education have been investigated and discussed within the Fibonacci project. In the Fibonacci Background Resource Booklets 'Learning through Inquiry' (Artigue, Dillon, Harlen, & Léna, 2012) and 'Inquiry in Mathematics Education' (Artigue & Baptist, 2012), the authors present the similarities and specificities of mathematical inquiry compared to scientific inquiry:

"Like scientific inquiry, mathematical inquiry starts from a question or a problem, and answers are sought through observation and exploration; mental, material or virtual experiments are conducted; connections are made to questions offering interesting similarities with the one in hand and already answered; known mathematical techniques are brought into play and adapted when necessary. This inquiry process is led by, or leads to, hypothetical answers – often called conjectures – that are subject to validation." (Artigue & Baptist, 2012, p. 4)

The main differences between mathematical and scientific inquiry are based on the type of questions or problems they address and the processes they rely on for answering or solving them. These are aspects that characterize mathematical inquiry: the distinction between mathematical and extra-mathematical systems, a need to construct mental representations, a search for structure, patterns, and relationships and the principal aim of generalization (Hunter & Anthony, 2011; Mathematical Sciences Education Board, 1990).

Table 1 gives an overview of the similarities and differences between aspects of IBE within the three domains (The origin of the table is explained in D 2.5). The term 'aspects' was chosen in order to avoid overlaps to constructs such as 'abilities', 'competences', 'skills', 'standards' etc. Often they are not used distinct. The listed aspects might be skills, competence or abilities. The different aspects can principally be regarded as steps in the inquiry process that have a chronological order. However, an important characteristic of inquiry processes is that they are seldom linear. Students continually (or at least frequently, at different stages) have to check their progress or results with the plan they made in the beginning and make corrections or adaptations if necessary so that steps can be repeated or left out.

Table 1: Aspects of IBE in STM

Science	Technology	Mathematics
diagnosing problems and identifying questions	diagnosing problems and identifying needs	diagnosing problems
searching for information	searching for information	searching for information
	considering alternative solutions	considering multiple solutions
	creating mental representations	creating mental representations
formulating hypotheses	formulating hypotheses in view of the function of a device	formulating hypotheses
planning investigations	planning design	planning investigations
constructing and using models	constructing and using models	constructing and using models
researching conjectures		researching conjectures
	constructing prototypes/a prototype	
		finding structures/patterns
collecting and interpreting data		
evaluating results	evaluating results	
searching for alternatives	modifying designs	
		searching for generalizations
		dealing with uncertainty
constructing and critiquing arguments or explanations/argumentation/reasoning/using evidence	constructing and critiquing arguments or explanations/argumentation/reasoning/using evidence	constructing and critiquing arguments or explanations/argumentation/reasoning/using evidence
debating with peers/communicating	debating with peers/communicating	debating with peers/communicating
<i>Notes.</i>		
	Aspect of IBE in STM	
	Aspect of IBE in TM, SM or ST	
	Domain-specific aspects	

Although aspects have the same name, they might have slightly different meanings in the different domains and even within one domain (e.g. reasoning in science). Different frameworks might exist which have to be taken into account when comparing assessment methods and results between different studies. A detailed description of the different frameworks is beyond the scope of this report. A summary of theoretical papers dealing with different frameworks that were found during the review, however, is given in section 7.1 Frameworks of inquiry competences and/or assessment together with theoretical papers focusing on assessment methods.

In addition to these domain-specific skills, there are also transversal competences that are ascribed to inquiry. For example, the Benchmarks for Science Literacy (American Association for the Advancement of Science, 1998) pay special attention to the so-called 'habit of mind' which describes problem-solving skills that are relevant in all subjects. These skills are computation and estimation, manipulation and observation, communication and quantitative thinking, critical response skills (evaluating evidence and claims) and creativity in designing experiments and solving mathematical or scientific problems; the competence of the students is reflected in the quality of questions they pursue and the rigor of their methodology (American Association for the Advancement of Science, 1998). Moreover, a habit of mind also includes values and attitudes like honesty, curiosity, open-mindedness and scepticism. The key competences for lifelong learning described in the Recommendation of the European Parliament (European Parliament, 2006) supplement this list by the ability of learning to learn and a sense of initiative and entrepreneurship (creativity, innovation and risk-taking, as well as the ability to plan and manage projects in order to achieve objectives).

Attitudes investigated in the context of inquiry-based approaches to teaching and learning include, e.g., enjoyment, value, interest, and self-efficacy expectations. In mathematics, Schukajlow et al. (2012) found that student-centred, modelling-based teaching approaches most beneficially affected students' attitudes towards mathematics. Similar results were obtained for science (e. g. Gibson & Chase, 2002). Nolen (2003) investigated the relationship between learning environment, motivation and achievement in high school science. She found that task orientation and the value of deep-processing strategies are mediated by a learning environment that supports deep understanding and independent thinking. Moreover, a focus on science learning combined with a shared belief in the teacher's desire for student understanding and independent thinking accounted for all the predictable variation in satisfaction with learning. In technology education, there is still a lack of research on learning and instruction (Miranda, 2004). A recent review came to the conclusion that technology education research is still dominated by descriptive studies that rely on self-reports and perceptions (Johnson & Daugherty, 2008). However, an appreciation of the interrelationships between technology and individuals, society and the environment (International Technology Education Association, 1996) as well as of the concepts of sustainability, innovation, risk, and failure (Rossouw, Hacker, & Vries, 2011) is regarded as an important goal of technology education.

2.2 Assessment in education

Assessment is one of the most important driving forces in education and a defining aspect of any educational system. Assessment signals priorities for curricula and instruction since teachers and curriculum developers tend to focus on what is tested rather than on underlying learning goals which encourage a one-time performance orientation (Binkley et al., 2012; Gardner, Harlen, Hayward, Stobart, & Montgomery, 2010). However, assessment can be regarded from different perspectives. The European report "Europe needs more scientists" (European Commission, 2004, p. 137) distinguishes between three perspectives: (1) traditionally, as the function of evaluating stu-

dent achievement for grading and tracking, (2) as an instrument for diagnosis to give students and teachers continual feedback about learning outcomes and difficulties, and (3) as a means to enable broader knowledge about the conditions behind and influences on students' understanding and competence (e.g. in international large-scale assessments). In the last decades, accountability has become an increasingly important issue in assessment that strongly influences teaching practice – especially when high stakes are connected to it. Educational research in the United States and the United Kingdom has provided empirical evidence that high stakes, standard-based assessment systems have negative effects (for reviews see Cizek, 2001; Nichols, Glass, & Berliner, 2006; Pellegrino, Chudowsky, & Glaser, 2001). Given the anticipated consequences of their students' test results, it has been shown that teachers adapt their classroom activities to the test, often devoting a considerable proportion of instructional time to test preparation. This could be seen in a positive light if the student competencies as assessed by the test were actually fostered but comparisons between the assessment systems of different US states showed that such positive effects rarely exist (Nichols et al., 2006). A similar result is reported by Anderson (2012) who argues that under accountability policies, many research-based reform efforts in science have become side-tracked and disrupted. Teacher practice has become more fact-based, science is taught less, teachers are less satisfied, and many students' needs are not met.

2.2.1 Characteristics of assessment systems

There is general agreement in the literature about the characteristics that define 'good' assessment systems. An important feature of assessment systems that support learning is coherence – classroom and external assessments have to share the same or compatible underlying models of student learning. Moreover, the design of international, national, state, and classroom-level assessments must be clarified and aligned (Bernholt, Neumann, & Nentwig, 2012; Mislevy, Steinberg, Almond, Haertel, & Penuel, 2001; Pellegrino et al., 2001; Quellmalz & Pellegrino, 2009; Waddington, Nentwig, & Schanze, 2007). The alignment of learning goals, instructional activities, and assessment is also stressed by Krajcik, McNeill, and Reiser (2008). Another important issue is instructional sensitivity. Ruiz-Primo et al. (2012) proposed an approach for developing and evaluating instructionally sensitive assessments in science called DEISA (Developing and Evaluating Instructionally Sensitive Assessments). The development approach considered three dimensions of instructional sensitivity; that is, assessment items should represent the curriculum content, reflect the quality of instruction, and have formative value for teaching. A similar point is made by Pellegrino et al. (2001). Items should be selected or combined in such a way that they provide additional information useful for diagnosis, feedback, and the design of next steps in instruction. Shepard (2003) focused on the student level and defined effective assessment as an assessment that makes students' thinking visible and explicit, engages students in the self-monitoring of their learning, makes the features of good work understandable and accessible to students, and provides feedback specifically targeted toward improvement (Shepard, 2003 and references therein).

2.2.2 Summative and formative assessment

Assessment always involves the collection, interpretation and use of data for some purpose. The purpose and often also the manner of data collection may differ. These different purposes are often summarized under the terms of summative and formative assessment.

Summative assessment has the purpose of summarizing and reporting learning at a particular time and, for this reason, it is also called ‘assessment *of* learning’. It involves processes of summing up by reviewing learning over a period of time or checking up by testing learning at a particular time. Summative assessment has an undeniably strong impact on teaching methods and content (Harlen, 2007), especially if high stakes are connected to it. This is also emphasized in the European report mentioned above: “Although the results [of large international assessments like PISA and TIMSS] may be used to identify strengths and weaknesses in each country, there is a danger that these studies may trivialize the purpose of schooling by its implicit definition of how educational ‘quality’ might be understood, defined and measured. It is likely that national school authorities put undue emphasis on these comparative studies, and that curricula, teaching and assessment will be ‘PISA-driven’ in the years to come” (European Commission, 2004, p. ix). The dominance of external summative assessment leads to situations where testing remains distinct from learning in the minds of most students and teachers. Thus, when teachers are required to implement their own assessments they tend to imitate external assessments and think only in terms of frequent summative assessment (American Association for the Advancement of Science, 1998; Black & William, 1998).

Formative assessment, in contrast, is “the process used by teachers and students to recognize and respond to student learning in order to enhance that learning, during the learning” (Bell & Cowie, 2001, p. 536). It thus has the purpose of assisting learning and, for this reason, it is also called ‘assessment *for* learning’. The term formative with respect to evaluation and assessment was first used by Scriven (1967) and Bloom (1969) in the late 1960s. According to Black and William (1998) and William (2006), assessments are formative if, and only if, something is contingent on their outcome and the information is actually used to alter what would have happened in the absence of that information – it thus shapes subsequent instruction. In their 1998 review of formative assessment, Black and William (1998) were able to show that formative assessment methods and techniques produce significant learning gains that are among the largest ever identified for educational interventions (Looney, 2011). As a consequence, formative assessment attracted a considerable amount of research interest because of its potential to improve student learning and to achieve a better alignment between learning goals and assessment (for reviews see Bennett, 2011; Dunn & Mulvenon, 2009; Kingston & Nash, 2011). Nevertheless, in one of the most recent reviews of formative assessment, (Bennett, 2011) states that “the term formative assessment does not yet represent a well-defined set of artefacts or practices” (p. 19). He observes a ‘split’ between those who regard formative assessment as referring to an instrument and those who understand it as a process; in his view, each view point is an oversimplification. Moreover, he regards the distinction between assessment ‘for’ and ‘of’ learning

as problematic since it absolves summative assessment from any responsibility to support learning.

2.2.3 Characteristics of formative assessment

Although a variety of methods, techniques, and instruments exists for formative assessment purposes, the methods show some common characteristics. Formative assessment has to be an integral part of teaching and learning (Bell & Cowie, 2001; Birnbaum et al., 2006). It has to be continuous, it has to actively engage students by peer- and self-assessment, and it has to provide feedback and guidance to learners on how to improve their learning by scaffolding information and focusing on the learning process (Looney, 2011; Wilson & Sloane, 2000).

Feedback has to be specific, has to be given in a timely manner, and has to be linked to specific criteria (Sadler, 1989). Not only is its quantity important but also its quality with respect to its technical structure (e.g. accuracy, appropriateness, and comprehensiveness), its accessibility to the learner and its catalytic and coaching value (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Sadler, 1998). Reviews of feedback aspects and their effects on education have been conducted, e.g., by Hattie and Timperley (2007), Kluger and DeNiSi (1996), and Shute (2008). The desired learning outcomes are clearly specified in advance which makes the learning process more transparent for students by establishing and communicating clear learning goals (Looney, 2011). The methods to be employed are deliberately planned but still allow teachers to adjust their teaching and vary their instruction method to meet individual student needs (OECD, 2005).

Formative assessment can be distinguished by its time frame (short – within/between lessons; medium – within/between teaching units; long – over semesters/years) and its amount of formality. The amount of formality ranges on a continuum from informal to formal depending on the amount of planning involved, the nature and quality of the data sought, and the nature of the feedback given to students by the teacher. Shavelson et al. (2008) describe three anchor points on the continuum: (1) 'on-the-fly', (2) planned-for-interaction, and (3) formal and embedded in the curriculum. The amount of planning is also defined by the distinction of Bell and Cowie (2001) between planned and interactive formative assessment. Whereas the former tends to be carried out with the whole class and involves the teacher in eliciting and interpreting assessment information and then taking action, the latter involves the teacher in noticing, recognizing and responding, and tends to be carried out with some individual students or small groups.

2.2.4 Assessment methods and techniques

In the preparation phase of the review, one goal was to find out which methods and techniques are used in formative and summative assessment in STM. It is a characteristic of formative assessment that it uses multiple instruments and techniques ranging from traditional paper and pencil tests to student observations. In general, this is also true for summative assessment, although, especially in large-scale assessments (e.g. PISA), a tendency to use multiple-choice, constructed-response or short open-ended questions can be observed. In contrast to, e.g., extended essays, student notebooks or

performance assessments, these questions can be comparatively easily and reliably scored. Alternative assessment methods in STM include, e.g., quizzes (e. g. Hickey, Taasobshirazi, & Cross, 2012), portfolios (e. g. Gitomer & Duschl, 1995), learn logs or student notebooks (e.g. Barron & Darling-Hammond, 2008), artefacts (e. g. Kyza, 2009), concept or mind maps (e. g. Ruiz-Primo & Shavelson, 1997), performance assessments (e.g. Barron & Darling-Hammond, 2008), and different methods of assessment discourse such as effective questioning (Learning how to Learn Project, 2002), assessment conversations (e. g. Ruiz-Primo & Furtak, 2006), or accountable talk (e. g. Michaels, O'Connor, & Resnick, 2008). Often, these methods are accompanied or complemented by techniques of student observation like video, audio, or field notes (see 5.2.1 Science; e. g. Vellom & Anderson, 1999). Moreover, interviews are employed to gain deeper insights into student thinking (see 5.2.1 Science, e. g. Berland, 2011). In computer-assisted learning and assessment environments, information from log-files can provide additional information. If the assessment method is more open (in contrast, e.g., to multiple-choice items), general or specific rubrics often exist to make a valid and reliable analysis and scoring of student responses possible (e.g. Barron & Darling-Hammond, 2008). Rubrics are also employed in student peer- and self-assessment (Toth, Suthers, & Lesgold, 2002). A summary of assessment instruments found during the literature review is given in Appendix 8.2 and 8.3.

2.2.5 Formative assessment – barriers and support

Recent OECD publications stress the importance of formative assessment and its integration with summative assessment (Looney, 2011; OECD, 2005). They also realize, however, that assessment in many countries still seems to be dominated by summative assessment (see D 2.3 'National reports of partner countries reviewing research on formative and summative assessment in their countries'). Looney (2011) attributes this, among other things, to a perceived tension between formative and highly-visible summative assessments. Moreover, many logistical barriers to making formative assessment a regular part of teaching practice exist.

In order to foster the use of formative assessment, it is essential to first enable teachers to change their deeply held pedagogical beliefs of assessment as a tool for teacher use and accountability rather than as a method to involve students in a constructivist assessment environment. The understanding and acceptance of innovations by the teachers is crucial to the ultimate success of change (Wilson & Sloane, 2000). This can be supported by:

- **Integrating assessment and instruction**
Assessment still often remains distinct from learning in the minds of most students and teachers (American Association for the Advancement of Science, 1998).
Assessment is discussed in terms of particular strategies, techniques, and procedures, distinct from other teaching and learning activities (Coffey, Hammer, Levin, & Grant, 2011).
- **Embedding formative assessment in the curriculum**
The effectiveness of an assessment depends, to a large part, on how well it aligns with the curriculum to reinforce common learning goals (Pellegrino et al., 2001; Shavelson et al., 2008). In order for assessment to become fully and

meaningfully integrated into the teaching and learning process, it must be curriculum dependent i.e. linked to a specific curriculum (Wilson & Sloane, 2000).

- **Fostering the collaboration between curriculum and assessment experts as well as teachers**

Building stronger bridges between research, policy and practice is essential for success but is also challenging (Shavelson et al., 2008).

Teachers should review the assessment questions that they use and discuss them with peers (Ayala et al., 2008; Black & William, 1998).

- **Enhancing accountability**

Teachers must feel confident that new assessment methods will be accepted for accountability purposes by school administrators and the public at large (American Association for the Advancement of Science, 1998).

- **Supporting teachers by teacher professional development (TPD)**

(Pedder, 2006; William, 2006). William considers “the task of improving formative assessment [to be] substantially, if not mainly, about TPD”. The provision of tools for formative assessment – although a necessary condition – will only improve formative assessment practices if teachers can integrate them into their regular classroom activities. To reach this goal, teachers need help to change the perception of their own role (American Association for the Advancement of Science, 1998). Moreover, TPD could foster the integration of assessment into instruction by combining work on assessment with work on instruction and materials.

In her report about the integration of formative and summative assessment, Looney (2011) identifies barriers to an implementation of formative assessment as well as policies that might support it. Although ASSIST-ME is primarily interested in approaches or policies for fostering the implementation of formative assessment, the perceived barriers can provide valuable information that has to be kept in mind when developing assessment methods.

Barriers to an implementation of formative assessment are seen in large classes, extensive curriculum requirements, the difficulty of meeting diverse and challenging student needs, fears that formative assessment is too resource-intensive and time consuming to be practical, a lack of coherence between assessments and evaluations at the policy, school and classroom level, the perception of formative assessment methods as ‘soft’, non-quantifiable assessments by policy makers/administrators, and a perceived tension between formative assessment and highly visible summative assessment (see above). Within the ‘Learning How to Learn’ project, Pedder (2006) found that classroom assessment practices are influenced and defined by conflicting and quite separate principles, namely assessment for learning principles (making learning explicit and promoting learning autonomy) and assessment of learning principles (performance orientation). Teachers’ assessment practices were often out of step with their teaching values.

Difficulties in informal assessment of mathematics are the focus of a study by Watson (2006). In this theoretical paper, the informal assessment practices of two experienced lower secondary mathematics teachers are used as cases for generating questions about future developments in formative assessment practice. In their instruction, both teachers maintain a consistent formative assessment focus on the development of their students as inquirers which one of them supplements with explicit self-assessment

activities. Nevertheless, there are differences in their teaching styles and in the ways in which they assess and describe their students (e.g. levels of formality, amount of content focus or opportunities for self-audit). One conclusion of the author is that a mixture of observation, interaction and judgment that is informed by belief, image and purpose is typical of teachers' informal assessment habits. From the analysis, several questions emerge with respect to the future of formative assessment practice: (a) Can ways be found to use performance data from large-scale studies to construct relevant information for individual teachers? (b) Can non-linear pathways of mathematical development be described?, and (c) How can such descriptions be used by teachers and students without reducing mathematical inquiry to a rubric without purpose?

In contrast, formative assessment practices could be supported by fostering teachers' and school leaders' assessment literacy (i.e. an awareness of the different factors that may influence the validity and reliability of results, the capacity to make sense of data, to identify appropriate actions and to track processes (Alkharusi, 2011 and references therein; American Federation of Teachers, National Council on Measurement in Education, & National Education Association, 1990; Brookhart, 2011; Looney, 2011; OECD, 2005). This could be accomplished by investing in teacher training and support, e.g. by providing guidelines and tools to facilitate formative assessment practice, by encouraging innovation and creating opportunities for teachers to innovate, and by developing clear definitions of learning goals and a theoretical framework of how that learning is expected to unfold as the student progresses through the instructional activity. Policy makers and administrators have to be convinced that formative assessment methods are not 'soft' but rather that they measure the development of higher order thinking skills (American Association for the Advancement of Science, 1998). Educational systems should build stronger bridges between research, policy and practice and should actively involve students and parents in the formative process to ensure that classroom, school, and system level evaluations are linked and are used formatively to shape improvements at every level of the system.

2.2.6 Links between formative and summative assessment

Finally, the links between formative and summative assessment could be strengthened by drawing on advances in the cognitive sciences to strengthen the quality of formative and summative assessment (Shepard, 2000 and references therein), by developing curriculum-embedded or 'on-demand' assessments, by taking advantage of technology, by using population instead of census sampling (Chudowsky & Pellegrino, 2003), by developing complementary diagnostic assessments for students at lower proficiency levels to identify specific learning difficulties (Looney, 2011), and by ensuring that standards of validity, reliability, feasibility, and equity are met (American Association for the Advancement of Science, 1998). Moreover, teachers' assessment roles should be strengthened (see assessment literacy above). Heritage, Kim, Vendlinski, and Herman (2009) found that teachers are quite competent in identifying the key mathematical principles being assessed and characterizing the students' level of understanding but had problems determining appropriate next instructional steps. As a last point, the strengthening of teacher appraisal is mentioned (Looney, 2011). There are a number of challenges to the development of coherent and valid measures in the formative as-

assessment practice as it involves several steps, including the assessment process, the interpretation of the evidence of students' learning, and the development of next steps for instruction (Herman, Osmundson, & Silver, 2010).

There is some argumentation in the literature about how close the link between formative and summative assessment might – or should – be. In principal, the term 'formative' is not a property of an assessment; the same test could be used for formative or summative purposes (Bloom, 1969; Wiliam, 2006). Harlen and James (1997), however, argue that the requirements of assessment for formative and summative purposes differ in several dimensions (e.g. reliability, reference base, etc.). They thus challenge the assumption that summative judgments can be formed by the simple summation of formative ones. On the other hand, Black, Harrison, and Hodgen (2010) consider a positive link between formative and summative assessment as going beyond the simple formative use of summative tests. This could be achieved by making use of peer- and self-assessment, thus engaging students in a reflective review of the work they have done, encouraging them to set questions and mark answers, and applying criteria to help them understand how their work could improve (Black, Harrison, Lee, Marshall, & Wiliam, 2004). Looney (2011), moreover, states that especially large-scale summative tests often do not reflect the promoted development of higher-order skills such as problem solving, reasoning, and collaboration – which are key competences in IBE. This is supported by William (2008) who finds that assessments such as PISA are usually relatively insensitive to high-quality instruction. This leads to technical barriers to a more close integration of formative and summative assessment because large-scale summative assessment data are often not detailed enough to diagnose individual student needs or they are not delivered in a time frame which enables them to have an impact on the students assessed. Moreover, creating reliable measures of higher-order skills is still a challenge. Related to this, Looney (2011) sees three major challenges: (1) Developing assessments that measure not only 'what' but also 'how to', (2) Reporting results in a 'criterion-referenced' way instead of a 'norm-referenced' way, including the development of focused reporting scales in criterion-referenced systems to provide diagnostic information (especially for weak students), and (3) Finding a balance between generalizability, reliability, and validity (e. g. Wilson & Sloane, 2000).

Nevertheless, in the literature, some attempts to use summative assessment data formatively (or vice versa) can be found. William and Ryan (2000) analysed the performance of 7 and 14 year old students in the 1997 UK mathematics tests. They tried to describe the children's progression in thinking as it related to their test performance; however, the authors found that the items often were not diagnostic enough. An attempt to combine formative and summative assessment in inquiry-learning environments was also made by Hickey et al. (2012) who used the concept of close, proximal, and distal assessment items. Modest empirical evidence was found that improvement in (formative) feedback conversations leads to gains in external (summative) achievement tests. Pellegrino et al. (2001) described examples in which alternative assessment approaches were successfully used to evaluate individuals and programmes in large-scale contexts in the US.

2.2.7 Assessment and inquiry

Some references looking at the relationship between assessment and inquiry could be found. According to Barron and Darling-Hammond (2008), assessment systems that support inquiry approaches share three characteristics. They contain intellectually ambitious performance assessments, evaluation tools such as guidelines and rubrics, and formative assessments to guide the feedback to the students and shape instructional decisions. As types of assessments that could be used in inquiry lessons the authors name: rubrics (must include scoring guides that specify criteria for students and teachers), solution reviews, whole class discussions, performance assessments, written journals, portfolios, weekly reports, and self-assessments. The authors claim that “most effective inquiry approaches use a combination of on-going informal formative assessment and project rubrics that communicate high standards” (Barron & Darling-Hammond, 2008, p. 3); however, no references are given. The Principled Assessment Designs for Inquiry project (PADI) aimed to provide a practical, theory based approach to developing high-quality assessments of science inquiry by combining developments in cognitive psychology and research on science inquiry with advances in measurement theory and technology. The centre of attention was a rigorous design framework for assessing inquiry skills in science which are highlighted in standards but difficult to assess (Mislevy et al., 2003; SRI International, 2007). The difficulty of assessing inquiry skills is also addressed by Hume and Coll (2010) who conclude that standards-based assessments using planning templates, exemplar assessment schedules and restricted opportunities for full investigations in different contexts tends to reduce student learning about experimental design to an exercise in 'following the rules'.

The relation between inquiry-based science education (IBSE) and assessment, especially formative assessment, was the focus of a conference held in York in 2010 titled “Taking IBSE into secondary education”. As an outcome of the conference, it was stated that “implementation of IBSE will require some fundamental changes particularly in [...] the form and use of assessment and testing” (INQUIRE project, 2010, p. 6). The participants agreed that a full implementation of inquiry will involve the use of formative assessment since the aims of formative assessment and IBSE coincide in helping students to take responsibility for their own learning; however, introducing inquiry-based science education and formative assessment both require a considerable change in pedagogy (INQUIRE project, 2010). The shared potential of formative assessment and inquiry to develop understanding through students taking charge of their own learning is also stressed by Harlen (2009). Delandshere (2002) argues that formative assessment itself can be understood as a form of inquiry (e.g. asking questions, defining criteria, interpreting data, coming to conclusions, communicating results, etc.). In their investigation of problem and project based learning, Barron and Darling-Hammond (2008) eventually state that formative assessment might provide a kind of scaffolding that supports student learning. Scaffolding is defined as a “process that helps a child or novice to solve a problem, carry out a task, or achieve a goal which would be beyond his unassisted efforts” (Barron & Darling-Hammond, 2008, p. 276).

3. Objectives of the literature review

The first phase of ASSIST-ME, including WP 2, focused on producing the knowledge base necessary for a research-based design of assessment methods, followed by a trial implementation of these methods. Therefore, the development of a baseline definition of IBE in STM (see D 2.5 ‘A definition of inquiry-based STM education and tools for measuring the degree of IBE’) and the identification of a set of assessment methods suitable for enhancing inquiry-based learning in STM were the starting point, as described above. The literature review takes up on these definitions and aims to answer the following research questions:

- Which aspects of IBE are investigated by empirical studies in STM?
- What formative and summative assessment methods are used in STM with respect to the aspects of IBE?
- How are these methods used?

Thus, this report is a review of existing knowledge about the formative and summative assessment of knowledge, as well as the competences and/or attitudes in IBE in STM. It focuses on the findings of empirical studies which are related to the research questions mentioned above. The report presents the findings from a comprehensive analysis of existing research on how the summative and formative assessment of knowledge, and the competences and/or attitudes in STM can be linked to aspects of IBE. The focus lies on methods which improve students’ outcomes.

Table 2 shows the intended objective. On the one hand, there are aspects of IBE (see also Table 1) and, on the other hand, there are different formative assessment methods. The question is: Which formative assessment methods are suitable for the assessment of specific aspects of IBE? For example, portfolios are used for the assessment of the aspect ‘planning investigations’ or ‘constructing prototypes’ in order to understand the procedure which the students use (Dori, 2003; Samarapungavan, Mantzicopoulos, & Patrick, 2008; Samarapungavan, Patrick, & Mantzicopoulos, 2011; Williams, 2012).

Table 2: Starting point for the identification of possible connections between IBE and formative assessment

Inquiry-based education	Connections between inquiry-based education and assessment methods	Formative assessment
Diagnosing problems	?	Concept maps
Critiquing experiments		Mind maps
Distinguishing alternatives		Portfolios
Planning investigations		Science notebooks
Researching conjectures		Multiple-choice
...		...

To reach this objective, a literature review was conducted. Its search strategies are presented in section 4. Procedure of the literature review. By categorizing the publications found, information was gathered about IBE and formative and summative assessment. Possible connections will be discussed in report D 2.6 'Report of outcomes of the expert workshop on assessment in STM and IBE' and recommended in report D 2.7 'Recommendation report from D 2.1 – D 2.6'.

4. Procedure of the literature review

The starting point of the literature review was – as described in D 2.2 ‘Synopsis of the literature review’ – the appointment of appropriate keywords. However, a systematic search using keywords faces several challenges.

Above all, these challenges are caused by the diversity of terms and instructional or teaching approaches that include characteristics of IBE. A literature search just using ‘inquiry’ as the keyword would, on the one hand, miss a lot of relevant publications. On the other hand, it would find an unmanageable number of publications. Besides, not only IBE comes under a variety of terms and approaches, but also some of the outcome variables like formative assessment. Therefore, relatively open keyword approaches do not seem to be feasible for the work in the ASSIST-ME project.

For this reason and due to the experience gained in the synopsis (see D 2.2 Synopsis of the literature review), a large number of relevant keywords were defined. Then, three different search strategies were applied to conduct the literature review:

1. Searches in data bases,
2. Searches in relevant journals,
3. Searches in reference lists.

These searches yielded approximately 200 results as a final extract which was managed in a Citavi-project file and evaluated in an Excel file (see 5. Results of the literature review). The following sections describe how these nearly 200 publications were extracted and how the searches were carried out. In addition, an expert survey was realized in order to validate the results and in order to receive recommendations of further relevant and/or influential publications in the field of formative and summative assessment as well as in IBE or problem-solving in STM.

The search concerning ICT-assisted assessment was conducted and documented by Pearson Education International as their contribution to the work of WP 2 in the ASSIST-ME project. The results are presented in part II of this report.

4.1 Searches in data bases

The search in databases allows for the systematic and simultaneous search in a collection of most of the important journals within a specific field of interest. According to the ASSIST-ME proposal (Dolin, 2012), two data bases were selected for this literature review. The first one is ‘Web of Science’ provided by Thomson Reuters. Web of Science includes the ‘Science Citation Index Expanded’ covering over 8500 major journals across 150 disciplines (including education in the scientific disciplines) from 1900 to present as well as the ‘Social Sciences Citation Index’ covering over 3000 journals across 55 social science disciplines (including education and educational research) as well as selected items from 3500 of the world’s leading scientific and technical journals from 1900 to present. Within the Social Sciences Citation Index, the following journals are e.g. listed:

- Review of Educational Research
- Learning and Instruction

- American Educational Research Journal
- Journal of the Learning Sciences
- Educational Researcher
- Journal of Research in Science Teaching
- Science Education

These journals have impact factors that are among the top ten in the 2012 Thomson Reuters Journal Citation Reports (JCR) Social Science Edition. “Journal Citation Reports® is a comprehensive and unique resource that allows for evaluating and comparing journals using citation data drawn from over 11000 scholarly and technical journals from more than 3300 publishers in over 80 countries. It is the only source of citation data on journals, and includes virtually all areas of science, technology, and social sciences” (Thomson Reuters, 2012).

Other journals included in the Web of Science database are e.g. in the field of technology education:

- Journal of Engineering Education,
- Journal of Science Education and Technology,
- International Journal of Technology and Design Education,
- International Journal of Engineering Education,

and in the field of mathematics education:

- Journal for Research in Mathematics Education,
- Educational Studies in Mathematics,
- International Journal of Science and Mathematics Education.

The second database that was used is ‘Education Resources Information Center’ (ERIC). In contrast to Web of Science that presents a broad range of science journals, ERIC focuses specifically on the field of general education and provides access to education literature and resources. It contains more than 1.4 million records and links to more than 337.000 full-text documents from ERIC.

For the literature review, the last 15 years, from April 1st 1998 till April 1st 2013, were chosen as the time span. The selection of the keywords was based on the collection of definitions in the ASSIST-ME project proposal (Dolin, 2012) and on a first unsystematic literature review which is described in D 2.2 ‘Synopsis of the literature review’. Furthermore, a first list of keywords was presented and discussed with the project partners at the WP 2 workshop during the ASSIST-ME kick-off conference in Copenhagen on January 26th 2013. The feedback was considered when the final list of keywords was built. Then, one expert from each subject approved the list. Afterwards, the keywords were grouped into six topics. Each topic is related to an aspect of ASSIST-ME (see Table 3). For example, topic 1 is related to the aspect of IBE. Furthermore, topics 1 and 2 cover domain-specific aspects by considering subject-specific keywords for IBE and alternative keywords for mathematics, science or technology education.

Table 3: Keywords for searches in data bases

Topics	Keywords		
	Science	Technology	Mathematics
Topic 1: inquiry	Inquiry-based learning OR inquiry OR collaborative learning OR discovery learning OR cooperative learning OR constructivist teaching OR problem-based learning OR argumentation	inquiry OR design OR problem-based learning OR project-based learning OR argumentation OR collaborative learning	inquiry OR didactical learning OR didactical situations OR open approach OR problem based-learning OR problem centred learning OR "realistic mathematics education" OR argumentation
Topic 2: subject	science education OR science instruction OR science teaching and learning	technology education OR engineering education OR technology instruction OR technology teaching OR technology learning	mathematics education OR mathematics instruction OR mathematics teaching OR mathematics learning
Topic 3: school	classroom OR teacher OR student	classroom OR teacher OR student	classroom OR teacher OR student
Topic 4: objective	assessment OR evaluation OR validation OR achievement OR feedback	assessment OR evaluation OR validation OR achievement OR feedback	assessment OR evaluation OR validation OR achievement OR feedback
Topic 5: type of assessment	formative OR embedded OR summative	formative OR embedded OR summative	formative OR embedded OR summative
Topic 6: method of assessment	discourse OR effective questioning OR assessment conversations OR accountable talk OR quizzes OR self-assessment OR peer-assessment OR portfolio OR learn log OR mind map OR concept map OR rubrics OR science notebook OR multiple-choice OR constructed-response OR open-ended response	discourse OR effective questioning OR assessment conversations OR accountable talk OR quizzes OR self-assessment OR peer-assessment OR portfolio OR learn log OR mind map OR concept map OR rubrics OR science notebook OR multiple-choice OR constructed-response OR open-ended response	discourse OR effective questioning OR assessment conversations OR accountable talk OR quizzes OR self-assessment OR peer-assessment OR portfolio OR learn log OR mind map OR concept map OR rubrics OR science notebook OR multiple-choice OR constructed-response OR open-ended response

For the searches in the data bases, the topics were combined to achieve a high correlation between the content of the literature found and the objectives of the ASSIST-ME project. The five combinations are presented in Table 4. The first search resulted in a very large number of references. By checking the content of the literature found, it became obvious that most of the publications did not meet the aims of the ASSIST-ME project. Therefore, the search strategy was changed. In order to focus on the intended objectives, the keywords of topic 5 were added (search 2). As a result, the number of references substantially decreased which increased the danger of missing relevant

publications. Thus, topic 5 was exchanged for topic 6 (search 3) and the explicit mentioning of the terms formative and summative was avoided. The third search strategy led to a better result in view of relevant literature. Searches 4 and 5 were carried out in order to verify the search strategy. By deleting the keywords of topic 1, the literature found once again did not meet the objectives of the ASSIST-ME project. Thus, search strategy 3 was used for the data base searches. With regard to the WP 2 time frame, it led to a manageable number of publications while, at the same time, yielded results that are relevant with respect to the project objectives.

The results of the searches were refined in the data bases by the following categories: 'education educational research', 'education scientific disciplines', 'education special', 'computer science interdisciplinary applications', 'psychology educational'. In addition, the chosen document types were articles, book chapters or reviews.

There is an overlap between the results of the two data bases within a subject. However, it is quite low. Therefore, these findings confirm that carrying out a search in two different data bases was worthwhile. Ultimately, 331 publications in science, 88 in mathematics and 68 in technology were found. The references were imported to a Citavi-project file.

Table 4: Results of the searches in data bases

Web of Science									
Search	Variations						Results		
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	S	M	T
1	Inquiry-based learning OR ...	science education OR ...	classroom OR ...	assessment OR ...			790	171	249
2	Inquiry-based learning OR ...	science education OR ...	classroom OR ...	assessment OR ...	formative OR ...		69	11	25
3	<i>Inquiry-based learning OR ...</i>	<i>science education OR ...</i>	<i>classroom OR ...</i>	<i>assessment OR ...</i>		<i>discourse OR ...</i>	163	34	50
4		science education OR ...	classroom OR ...	assessment OR ...		discourse OR ...	513	181	64
5		science education OR ...	classroom OR ...			discourse OR	1253	423	105
Education Resources Information Center									
1	Inquiry-based learning OR ...	science education OR ...	classroom OR ...	assessment OR ...			1105	482	220
2	Inquiry-based learning OR ...	science education OR ...	classroom OR ...	assessment OR ...	formative OR ...		82	23	17
3	<i>Inquiry-based learning OR ...</i>	<i>science education OR ...</i>	<i>classroom OR ...</i>	<i>assessment OR ...</i>		<i>discourse OR ...</i>	+183	+56	+25
4		science education OR ...	classroom OR ...	assessment OR ...		discourse OR ...	749	526	49
5		science education OR ...	classroom OR ...			discourse OR	1255	888	84
Search 3: Results of both data bases									
						Duplicates	-15	-2	-7
						Total	= 331	= 88	= 68

4.2 Searches in relevant journals

In addition to the searches in the data bases, searches in relevant journals were conducted as a result of the discussion about the search strategies at the ASSIST-ME Kick-off meeting in Copenhagen. The journals in Table 5 were considered as relevant in view of the objectives of the ASSIST-ME project or even as the most important for each subject or research field. If available, the impact factors of each journal are presented for the last year and the last five years, indicating their importance. Those journals that have an impact factor are also included in the Science Citation Index or in the Social Science Citation Index and are thus regarded by searches in the data base Web of Science.

However, the impact factors were not the only criterion for the selection of the journals. In addition, publications about the importance of journals were considered. For example, Johnson and Daugherty (2008) asked key leaders in the field of technology education to identify what they consider the top research-focused journals in the field. "The following four technology education journals were consistently mentioned by the panel of experts: (a) the International Journal of Technology and Design Education (ITDE), (b) the Journal of Industrial Teacher Education (JITE), (c) the Journal of Technology Studies (JTS), and (d) the Journal of Technology Education (JTE). This is essentially the same list of refereed journals that Zuga analysed in her 1994 study. The only difference is that Zuga included 'The Technology Teacher' while this study included the 'International Journal of Technology and Design Education'." Journals focusing on teachers or teacher education were excluded because ASSIST-ME focuses mainly on students.

Table 5: Relevant journals and their impact factors

Subjects	Journals	Impact factor ¹	
		Last year	Last five years
Science	Journal of Research in Science Teaching	2.55	3.23
	Science Education	2.38	2.71
Technology	Int. Journal of Technology and Design Education	0,34	0.42
	Journal of Technology Education	-	-
	Journal of Technology Studies	-	-
Mathematics	Educational Studies in Mathematics	0.77	-
	Int. Journal of Science and Mathematics Education	0.46	-
	Journal for Research in Mathematics Education	1.55	2.08
Assessment	Applied Measurement in Education	0.58	0.74
	Assessment in Education	-	-
	Educational Assessment	-	-

¹(according to Thomson Reuters, 2013)

Both methods led to the list of journals in Table 6. The articles of all issues published during the last 10 years were scanned by using the homepages of the publishers and the two data bases mentioned above. Compared to the search in the data bases, the numbers of references were much lower. But, the differences between the subjects were also much smaller. Thus, this search was able to improve the quantity and quality of the literature basis.

Table 6: Results of the searches in the issues of relevant journals by subject

Subjects	Journals	Results	
		Per journal	Per subject
Science	Journal of Research in Science Teaching	44	63
	Science Education	19	
Technology	Int. Journal of Technology and Design Education	14	24
	Journal of Technology Education	9	
	Journal of Technology Studies	1	
Mathematics	Educational Studies in Mathematics	11	30
	Int. Journal of Science and Mathematics Education	10	
	Journal for Research in Mathematics Education	9	
Assessment	Applied Measurement in Education	9	41
	Assessment in Education	19	
	Educational Assessment	13	
Total		158	158

4.3 Searches in reference lists

To guarantee that important literature with regard to IBE and formative or summative assessment was considered, an additional, more unsystematic search was carried out. Following the pyramid scheme, the reference lists of the literature found were scanned in view of frequently recurring publications which might have a high impact on research on IBE and formative or summative assessment. As well as the publications from the search in relevant journals, the references were added to the Citavi-project file. For science, there were 32 additional references that focused on students in school. For mathematics, there were only 10 publications, and for technology and assessment none.

4.4 Final extract

Finally, the literature collected by the different search strategies and searches was imported into one Citavi-project file. This file contained 732 references. However, 31 duplications resulted from the parallel searches. They were deleted from the project file. In the end, the Citavi-project file contained 701 entries.

Up to this point, a deeper analysis of all publications had not been carried out. Therefore, the titles and abstracts of the publications were read and categorized in order to further identify the relevant literature. Table 7 shows the categories and the numbers of

references for each category by subject. Only the publications in the category ‘focus students (school)’ should meet the objectives of the ASSIST-ME project. The other publications addressed the learning process of university students or its assessment; others contributed to the research on teacher education or development and some others did not report findings from an empirical study but only theoretical aspects. Therefore, these publications did not meet the core objectives of the ASSIST-ME project at the current stage of the project and were no longer regarded for this review. Nevertheless, the found publications focusing on teachers’ professional development should be evaluated at a later stage of the project when teacher training courses will be developed.

Table 7: Categorization of literature

Categories	Science	Mathematics	Technology	Assessment	Total
<i>Focus students (school)</i>	152	44	23	16	235
Focus students (university)	19	4	23	-	46
Focus teacher	57	38	14	5	114
No study ¹	58	12	28	13	111
Review	5	2	1	4	12
Book (Monograph)	15	2	1	-	18
Book (Serial)	11	6	5	-	22
Dissertation	9	6	2	-	17
Proceeding	-	6	2	-	8
Not relevant ²	94	18	3	3	118
Total	420	138	102	41	701

¹e.g. policy or methodological frameworks, description of approaches, theoretical discussions, or presentation of explorative investigations
²The content or focus of the publications is not connected to the objectives of ASSIST-ME.

In order to achieve a deeper analysis of the relevant literature from the category ‘focus students (school)’, all 235 publications were read and evaluated with a coding scheme. The results were filed in an Excel file. Table 9 shows the titles and contents of each column in the Excel file. First, the aim of this step in the analysis procedure was to gather information about the whole content of the publications. In addition, this step analysed the extent to which the literature met the objectives of the ASSIST-ME project. The second aim was to categorize the results with respect to the research questions:

- Which aspects of IBE are investigated by empirical studies in STM?
- What formative and summative assessment methods are used in STM with respect to the aspects of IBE?
- How are these methods used?

Besides, it was recorded which domain and grade level the studies address. Furthermore, the literature derived from the three assessment journals was reassigned to the three subject domains.

Table 8: Final extract for the literature review

Category	S	M	T	Total
<i>Focus students (school)</i>	148	30	13	191

Even though the literature was categorized by reading the titles and abstracts in advance, 42 references were identified which did not belong to this category but to one of the others. The remaining 191 references are the publications which meet the objectives of the ASSIST-ME project and thus form the final extract for this report (see Table 8). Even though there was a partial selection before, 510 of all 701 publications were excluded. Chapter 5. Results of the literature review summarizes the empirical results of the 191 publications. Obviously, the three search strategies resulted in a huge number of publications in science education but only in a few number of publications in mathematics and especially technology education. Reasons might be that IBE as a teaching and learning approach is best developed and investigated in science education. In technology education there might be less research on IBE as technology is not a common school subject in a lot of countries. In mathematics education there is huge range of different teaching and learning approaches or theories which might include aspects of inquiry (see D 2.5). Therefore, the strongly focused search strategy applied within this review might not reflect this diversity and thus lead to the small number of publications in mathematics.

Some of the aspects of IBE focused on by the interventions and learning environments or by the assessment are conceptually not distinguishable. Therefore, ‘considering alternative or multiple solutions’, ‘searching for alternatives’ and ‘modifying designs’ are combined in one paragraph. The aspects ‘formulating hypotheses’ and ‘researching conjectures’ are evaluated in one section as well. Third, ‘collecting and interpreting data’ and ‘evaluating results’ are also described within one section.

Table 9: Scheme for the evaluation of the literature

Column	Content
Literature	author(s)
General information about the investigation/ analysis	year
	country
	design (Survey, Intervention, Evaluation, Case Study, Meta-analysis)
	domain (Science, Technology, Mathematics)
	sample(s) size (N)
	sample characteristics: grade (school type)
	sample characteristics: age
Content focus of the investigation/ analysis (either as focus of the intervention/learning environment/curricula or as focus of the assessment)	scientific inquiry/science process skills
	diagnosing problems/ identifying questions
	searching for information
	considering alternative or multiple solutions
	creating mental representations
	constructing and using models
	formulating hypotheses
	planning investigations
	constructing prototypes
	finding structures or patterns
	researching conjectures
	collecting and interpreting data
	evaluating results
	searching for alternatives/ modifying designs
	constructing and critiquing arguments or explanations/ argumentation/ reasoning/ using evidence
	debating with peers/ communication
	searching for generalizations
	dealing with uncertainty
	knowledge/ achievement/ understanding/ conceptual change
	problem solving
other	

Assessment: method/ practice	Multiple-choice
	constructed-response/ open-ended
	concept map
	mind map
	portfolios
	learn log
	notebook
	effective questioning
	discourse/ assessment conversations/ accountable talk
	heuristics
	quizzes
	performance assessment/ experiments
	interviews
	observation/ field notes
	video tapes
	audio tapes
questionnaires	
written materials	
artefacts	
other	
Assessment: character/ type	summative assessment
	formative assessment
	embedded assessment
	computer-based/-assisted assessment
	software or learning environment used or curriculum
Assessment: additional information	feedback
	peer-assessment
	self-assessment
	rubrics
	other
Assessment instru- ments given?	yes
	examples
	no
Rubrics given?	yes
	examples
	no
Important outcome	

4.5 Expert survey

The comparably small number of publications found in the field of mathematics education lead to concerns within the project that mathematics might not be adequately represented in the literature review. In order to validate the results from the review and to ensure that no relevant literature is missing, an expert survey was conducted. Experts from all three subject domains were asked to name those ten publications that they regarded as the most important or relevant in the field of formative and summative assessment or IBE and problem-solving, respectively.

In total, at the end of August 2013 twelve experts were contacted, four from the field of science education, two from the field of technology education and five from the field of mathematics education. Until the beginning of October, four experts had responded to the survey, three from mathematics and one from science.

Most of the recommended publications are theoretical articles, reviews or books within the above mentioned research fields. Only very few publications refer to empirical studies.

In science, almost three quarter of the recommended publications had previously been found in the literature review. The additional publications are all theoretical papers dealing either with certain aspects within the field of IBE (e.g. the role of teachers or model-based inquiry as a new paradigm in school science) or the role of feedback in out of school contexts (management theory, communication networks and decision processes). Another additional paper by Wiliam (2007) investigated the relationship between classroom assessment and the regulation of learning and was also recommended by one of the mathematics experts.

Due to time constraints, it was not possible to include the additional empirical studies recommended by the mathematics experts within the results section of this review. They will thus be shortly described in the following. The theoretical publications about IBE or problem-solving are included in D 2.5 'A definition of inquiry-based STM education and tools for measuring the degree of IBE'.

In the field of mathematics education, the majority of recommended papers refers to formative assessment (34 compared to 18 in IBE). Compared to science, a smaller amount of publications had already been found within the literature review (12 papers). However, summarizing all publications, there is also only small agreement among the experts with only five papers being named by more than one expert.

Among the empirical studies, Elia, Gagatsis, Panaoura, Zachariades, and Zoulinaki (2009) investigated three different dimensions of grade 12 students' understanding of the concept of limit and their interrelations. These dimensions are students' conceptions concerning the meaning of the concept of limit; their competence in converting a certain expression of limit from a geometric to an algebraic representation and vice versa, and their problem solving abilities with respect to limits. Since no representation can fully reflect a mathematical construct and each form of representation has its advantages but also its limitations, especially the ability to flexibly use and convert representations is regarded as a prerequisite for the acquisition of conceptual understand-

ing. The assessment instrument consisted of a questionnaire that involved ten tasks related to the above mentioned dimensions of conceptual understanding and their interrelations. The results of the analysis indicated that students who had constructed a conceptual understanding of limit were more likely to accomplish the conversions of limits from the algebraic to the geometric representations and vice versa.

Verschaffel, Corte, and Vierstraete (1999) performed an error analysis to investigate grade five to six students' difficulties in modelling and solving nonstandard additive word problems involving ordinal numbers. The backdrop of their study was that in traditional instructional practice realistic modelling and interpreting are often missing. Students are not aware of the possibly problematic modelling assumption underlying their proposed solutions which leads them to approach arithmetic word problems in superficial, mindless and routine-based ways. The assessment instrument consisted of a 17-item paper & pencil word problem test in which tasks were deliberately formulated in a way that the addition/subtraction of two numbers will give either the correct result or a wrong result that differs ± 1 from the correct response. One example for such a task is e.g.: "In September 1995 the city's youth orchestra had its first concert. In what year will the orchestra have its fifth concert if it holds one concert every year?" (Verschaffel et al., 1999, p. 267). Related to the mathematical structure, the nature of the unknown quantity and the size of the number difference involved, nine different problem types of items were defined. The findings showed that the students had great difficulties in solving the items often resulting from a superficial, stereotyped approach of adding/subtracting two numbers without thinking about the appropriateness of the approach in the given situation.

Rodríguez, Bosch, and Gascón (2008) used the Anthropological Theory of the Didactic to analyse metacognition in problem solving in mathematics. Their theoretical considerations were supported by an empirical study in grade 11 focusing on the problem of comparing mobile phone tariffs which constitutes a complex problem with a multitude of variables. Students were asked to keep a portfolio including the progressive productions of their work; in addition field notes and video tapes were used as assessment instruments. The analysis of the 'didactic moments' in the process revealed that (a) teachers often destroyed them by wanting to make 'progress' and (b) that self- and peer-evaluation appeared naturally during the collaborative course work. At the end of the process, the students were asked to answer an individual written test on the comparison of fixed phone tariffs with some novelties. The results showed that the students were able to approach a question similar to the one previously studied, explain the process followed and use the comparison techniques constructed during their previous work in a flexible way.

Another aspect of problem solving that causes problems even for high performing calculus students was investigated by Moore and Carlson (2012). They looked at students' ability to model relationships between two dynamically varying quantities. This is regarded as a critical reasoning ability for thinking about and representing the quantitative relationships described in a problem statement which in turn provides the basis for future constructions and reflection during the problem solving process. The study focused on undergraduate pre-calculus students at university (age 18-25) which are be-

yond the age range addressed by the ASSIST-ME project. It has to be seen during the future work of the project whether the results are transferable to the school context or not. The students were assessed using structured, task-based clinical interviews. The authors found a positive correlation between the ability to mentally construct a robust structure of the related quantities and the production of meaningful and correct solutions. They concluded that it is critical that students first engage in mental activity to visualize a situation and construct relevant quantitative relationships prior to determining formulas or graphs.

The assessment of mathematical problem solving ability was also the focus of a study by Collis, Romberg, and Jurdak (1986). They reported the developing, administering, and scoring of a set of mathematical problem-solving items – so-called ‘superitems’ – and examined their construct validity using the ‘Structure of the Learned Outcomes – SOLO’ taxonomy. Each superitem included a mathematical situation and a structured set of questions about that situation that reflected the SOLO levels. The items belonged to six content categories (numbers and numeration; variables and relationships; size, shape, and position; measurement; statistics and probability; and unfamiliar) and were designed in a way that within any item a correct response to a question would indicate an ability to respond to the information in the stem at least at the level reflected in the SOLO structure of that question. Two test versions were constructed, one for 17-year-olds and one for nine to thirteen year-olds. The results showed that to construct valid items required input from three significant groups of people: (a) mathematicians, mathematics educators, and mathematics teachers; (b) people with expertise in interpreting the theoretical model in a practical situation and (c) students for whom the finished test was intended. Following this recommendation, however, the SOLO model proved viable for devising a construct valid test in mathematical problem solving suggesting that this kind of response model approach may be very useful for educators and researchers who have the task of describing levels of reasoning on school-related tasks.

The last two empirical studies recommended by the mathematics experts are examples for one of the key findings of the literature review presented in this report: the evaluation of an inquiry-based teaching approach by using standardized achievement measures. Both publications refer to a problem-centred mathematics program in the United States. Within the program, special emphasis was placed on e.g. the development of thinking strategies and the development of algorithms within the instructional activities as well as providing opportunities for collaborative working and whole-class discussions. The first paper by Cobb et al. (1991) compares results for ten grade two classes who had been participating in the program for one year with the results of eight non-program classes. Means for the comparison were two arithmetic competence tests: a standardized achievement test (the state-mandated multiple-choice standardized achievement test – ISTEP) and another arithmetic test developed by the program. Within the latter, items had been constructed in a way that they could be coded for the use of a standard algorithm or that incorrect answers would reveal the use of e.g. a figurative rule. Moreover, students had to fill in a questionnaire about personal goals and beliefs about the reasons for success in mathematics. Results showed that the

levels of computational performance were comparable between program and control group. However, qualitative differences in the use of arithmetical algorithms could be observed. Program students “had higher levels of conceptual understanding; held stronger beliefs about the importance of understanding and collaborating; and attributed less importance to conforming to the solution methods of others, competitiveness, and task-extrinsic reasons for success.” (Cobb et al., 1991, p. 3). In a later publication, Wood and Sellers (1997) presented results from a longitudinal analysis of grade three and four students within the same teaching program (and using the same assessment instruments). The study yielded similar results. Compared to students in textbook instruction, students in problem-centred classrooms had significantly higher arithmetic achievement, better conceptual understanding and more task-oriented beliefs.

Summarizing the outcomes of the expert survey, it can be said that for science the literature review seems to reflect the state-of-the-art of formative and summative assessment in IBE. For mathematics, the survey further emphasizes the importance of problem solving and its components in inquiry-based approaches to mathematics education. However, as far as assessment methods are concerned, the applied methods are in line with those identified within the literature review.

5. Results of the literature review

The identified publications were read by four researchers to extract the study's aim, design and results. The analysis focused on three questions:

1. Which aspects of IBE are emphasized or researched in the study?
2. Which types of assessment are employed in the study?
3. Which connections can be found between the emphasis on particular aspects of IBE and specific assessment instruments?

The following two chapters of report D 2.4 will be structured in line with the first two questions. The interrelatedness between the diverse aspects of IBE and assessment will be described in the recommendation report D 2.7 that will be based on all prior reports from WP 2. Then, connections made in the publications will be displayed to show which aspects are often bound and researched together.

When reading the next sections, it is important to keep in mind that in technology and mathematics education the number of found publications is rather low. Therefore, the findings from this literature review cannot be generalized for these two subjects. Nevertheless, in science education a sufficient number of publications was found.

As a kind of disclaimer, it is important to mention two issues for those reading this report. First, in line with the description of both IBE and formative and summative assessment stated above, the findings of the literature review are presented in a rather fragmented way. For instance, the different aspects of IBE are presented one after another, including specific foci and interpretations as extracted from the different papers in this review. Thereby, the interconnections between the different aspects are partly lost.

Second, the following description of findings mainly focuses on details of the different aspects of IBE and assessment instruments. However, for the purpose of better readability, not all studies relevant to a particular aspect are cited each time. We tried to include citations from relevant or representative papers, but no effort is made to achieve a balanced citation of all studies.

5.1 Which aspects of IBE are emphasized or researched in the study?

5.1.1 Diagnosing problems/ Identifying questions

Finding, identifying, and/or formulating a research question are certainly major steps in scientific inquiry processes, whereas diagnosing problems is mostly related to mathematics (e. g. Chang, Wu, Weng, & Sung, 2012) and technology education (e. g. Mioduser & Betzer, 2007). Accordingly, the aspect of diagnosing problems or identifying questions is present in many IBE studies. 44 publications of this review explicitly explored this aspect as part of a learning environment or as part of the assessment.

While the relevance of identifying the research problem and formulating a research question is intuitively clear to every researcher, the manner in which students come to a problem or question of interest makes a difference. Studies explicitly including this step of problem identification focus on/consider instruction that introduces students to a challenging problem (Toth et al., 2002), student-generated problems in science (Zhang & Sun, 2011), or students' ability to identify a situation in technology which demands a design (Mioduser & Betzer, 2007). As can be seen from Table 10, this aspect of inquiry has mainly been investigated in the field of science education. Highlighting personal relevance aims to stimulate students' engagement in the task so that they then take personal ownership of a problem (Silk, Schunn, & Cary, 2009).

For the evaluation of students' ability to diagnose problems and to identify research questions, Ebenezer, Kaya, and Ebenezer (2011) formulated two scoring criteria:

“Criterion 1: ‘Define a scientific problem based on personal or societal relevance with need and/or source’ means that students ought to identify and accurately define a community-based problem that is meaningful to them. The problem must have personal or societal relevance. Students should defend the problem based on the need for the study or because they have identified the problem from a reliable source.

Criterion 2: ‘Formulate a statement of purpose and/or scientific question’ means students should write the purpose and state a scientific question with clarity and precision.” (p. 102).

Regarding students' ability and results when asked to identify research questions of interest or relevance, different approaches can be identified. Dori and Herscovitz (1999) investigated students' question-posing capability as an alternative evaluation method. They used two case studies (dealing with rain forests and the threat of health hazard problems caused by the ozone layer) and asked students to pose as many questions as possible related to these two cases. The results of both case studies were analysed according to the number of questions posed by each student, the orientation of each question (differentiating between phenomena and/or problem descriptions, descriptions of hazards, and treatment and/or solution), the relation to the case study (establishing whether the answer is provided in the case study, a part of the answer is provided in the case study, or the answer cannot be found in the case study), and the complexity of each question (distinguishing between application and/or analysis, inter-

disciplinary approaches, judgement and/or evaluation, and taking a stance and/or forming a personal opinion).

Similarly, Chin and Osborne (2010) analysed students' questions and derived five categories of questions to classify the kind of questions students came up with: "(a) key inquiry; (b) basic information; (c) unknown or missing information; (d) conditions under which the heating was carried out; and (e) others" (p. 891). Key inquiry questions sought explanations. Basic information questions addressed the most basic, factual information students needed to know. Unknown or missing information questions asked for any information not given in the task sheet but which students felt was necessary. Questions in the conditions category included students' predictive thinking in terms of asking what would happen if the conditions of the experiment were altered.

Aguiar, Mortimer, and Scott (2010) analysed the impact of students' questions on the discourse of the lesson. The authors tried to reveal the 'teaching explanatory structure' (cf. Ogborn, Kress, Martins, & McGillicuddy, 1996) of a lesson, as it provides a way to conceptualize the teaching discourse which the students are responding to with their questions.

In general, students' ability to identify research questions was explicitly addressed in 44 publications (see Table 10). However, the majority of these publications included this introductory step of scientific inquiry processes only as a facet of the learning environment, while less than one third of the publications tried to explicitly assess students' ability in this step.

Table 10: Number of studies investigating 'diagnosing problems/ identifying questions'

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	5	21	1	27
Focus on assessment	1	10	1	12
Focus on both	0	5	0	5
Studies per subject [M]	6	36	2	44

5.1.2 Searching for information

Searching for information is an important and relevant step in each inquiry process. Missing information needs to be looked up, to be evaluated, and to be integrated into existing knowledge and inferences. The self-evident relevance of this step might be the reason for why it has only been researched by few studies.

Toth et al. (2002) distinguish between an information search and an evaluation of information. Additionally, the information search measure has two sub-items: "(1) How many topic-relevant information pieces were recorded and (2) How many topic-relevant

information pieces were labelled as data and hypotheses” (p. 274). The scoring revealed a broad use of categories by students, including theory, hypotheses, idea, fact, data, and evidence (Toth et al., 2002).

Regarding the evaluation of information, the amount of topic-relevant inferences was analysed. Three kinds of inferences were differentiated between: Consistency inferences (‘for’ inferences), indicating a supportive relationship between data and hypotheses; inconsistency inferences (‘against’ inferences), indicating disparities between hypotheses and data; and conjunction inferences (‘and’ inferences), indicating that two information pieces should be considered together during reasoning (Toth et al., 2002).

In general, only few studies focused on students’ search for information, especially as a facet of the respective assessment procedures, and they were almost exclusively located in the field of science education (see Table 11).

Table 11: Number of studies investigating ‘searching for information’

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	1	12	0	13
Focus on assessment	0	3	0	3
Focus on both	0	1	0	1
Studies per subject [M]	1	16	0	17

5.1.3 Considering alternative or multiple solutions/ searching for alternatives/ modifying designs

This aspect of IBE can play a role in different points of the inquiry process. Especially if the inquiry tasks involve ill-structured problems, students are required to consider alternative pathways towards a solution at an early stage of the process (e. g. MacDonald & Gustafson, 2004). After conducting the investigation and evaluating the results, however, the necessity to consider alternative solutions might also arise if the results do not yield the desired outcome. Especially in technology education, the improvement of an artefact after its construction is an important aspect (e. g. Hong, Yu, & Chen, 2011; MacDonald & Gustafson, 2004). In any case, the identification or evaluation of alternative or multiple solutions to an inquiry problem is a challenging step.

In addition, considering alternatives also deals with the use of a variety of investigation technologies. Accordingly, students should be able to decide between different tools to support their investigation (e.g., hand tools; measuring instruments and calculators; electronic devices; and computers for the collection, analysis, and display of data; (Ebenezer et al., 2011)). But, the challenges and sacrifices on the side of both the students and the researchers are quite high:

“To make sensible decisions about experimental designs that test the multitude of ideas they hold, learners need to combine their knowledge of combinatorial reasoning and controlling variables with methods for sorting out their disciplinary knowledge and identifying compelling questions. Learners must weigh multiple sources of knowledge to conduct informative experiments” (McElhaney & Linn, 2011, p. 748).

These high affordances might be the reason for the small number of studies identified which include this facet of IBE.

In their study within the field of science education, McElhaney and Linn (2011) asked students to develop a series of consecutive trials for the same investigation. Each trial was scored using a knowledge integration rubric from zero to five, reflecting the strength of the link between students' investigation goals and their variable choices in several ways. The authors describe three objectives of the rubric as it was used within the study:

“First, the rubric rewards conducting at least two unique trials for a particular investigation question, as comparisons between multiple trials are essential for illustrating variable relationships. Second, the rubric rewards varying the variable that corresponds to the chosen investigation question for that comparison. Third, the rubric rewards controlled comparisons that produce evidence for a variable effect, as measured by achieving opposite outcomes (safe or unsafe).” (McElhaney & Linn, 2011, p. 755).

In a similar manner, students in engineering classes in Australia were asked to design a product that would enable someone stranded on a beach with no drinking water to use the power of the sun to produce drinkable water from the sea water (Williams, 2012). The task required students to produce four alternative designs that were supposed to show revised and improved solutions to the problem.

In mathematics, only one study addressed this issue by asking students to find multiple answers or to apply multiple strategies to open-ended questions (Kwon et al., 2006). One example given was that students should choose from a list of numbers one number that was different from the others and explain their choice. They were instructed to try to find as many cases or answers as possible.

In total, 26 studies could be identified that incorporated students dealing with alternative or multiple solutions, either as part of a learning environment or as part of the assessment (see Table 12). Again, this facet of scientific inquiry was mainly incorporated within a learning environment, probably because of the high complexity of the analysis when carried out as part of the assessment.

Table 12: Number of studies investigating ‘considering alternative or multiple solutions/ searching for alternatives/ modifying designs’

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	0	11	2	13
Focus on assessment	1	5	2	8
Focus on both	0	3	2	5
Studies per subject [M]	1	19	6	26

5.1.4 Creating mental representations

The use of mental representations is a vast research area in itself (cf. Genter & Stevens, 1983). The power of internal and external representations “originates from the unique characteristic of each form of inscription – table, graph, picture – to guide the user’s attention towards employing specific strategies of extracting information encoded in these representations” (Toth et al., 2002, p. 266). Hence, the use of representations influences scientific inquiry processes by making ideas perceptually salient (Koedinger, 1992; Larkin & Simon, 1987). In mathematics, this aspect is often closely related to the aspect of finding patterns or structures (see 5.1.9 Finding structures or patterns). For example, Lin, Yang, and Chen (2004) investigated the relationship between reasoning, proving, and understanding proof in a number of patterns. This investigation was closely related to the process of representation, which incorporates exploring and searching for geometric number patterns, and explaining patterns verbally or diagrammatically.

Oh et al. (2012) analysed the impact of using simulation applets to facilitate students’ understanding of gas and liquid pressure concepts. The analysis indicated significant improvements in understanding when using the applets compared to didactic instruction. In addition, students were interested in the use of simulation applets and perceived them to be useful.

In general, the use of mental representations seems to be a characteristic feature of mathematics and science education. The studies extracted in these reviews are almost evenly distributed between these two domains, as well as between the adoption of mental representations as part of the learning environment or as part of the assessment (see Table 13).

Table 13: Number of studies investigating 'creating mental representations'

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	2	2	0	4
Focus on assessment	1	3	0	4
Focus on both	2	1	0	3
Studies per subject [M]	5	6	0	11

5.1.5 Constructing and using models

Analogous to the creation of mental models, the construction and usage of models is an important part of scientific reasoning. An indicator of students' understanding of scientific models is their ability to apply them to reasoning about scientific phenomena, patterns, and data (Anderson, 2003). In this regard, models can be used to explain or predict patterns or relations.

Schwarz and White (2005) developed curriculum material to foster students' learning about the nature of scientific models and to engage them in the process of modelling, especially by creating computer models that express students' own theories of force and motion, by evaluating their models using criteria such as accuracy and plausibility, and by engaging them in discussions about models and the process of modelling. In an evaluation study, students working with these materials wrote significantly better conclusions in an inquiry test and performed better in some far-transfer problems. In addition, the results suggest that developing knowledge of modelling and inquiry can be transferred to the learning of science content within such a curriculum.

In the field of chemistry, Kaberman and Dori (2009) developed curriculum material that integrates computerised hands-on experiments with molecular modelling. The material was evaluated with regard to its impact on students' higher-order thinking skills of question-posing, inquiry, and modelling. Their findings indicate that the experimental group of students performed significantly better than their comparison peers in all three examined skills. With regard to modelling skills, students in the experimental group significantly improved in making transfers from 3D models to structural formulae. But, in total, only about half of them were able to transfer from formulae to 3D models.

Zhang, Wilson, and Manon (1999) analysed gender differences in problem-solving strategies for two extended constructed-response mathematics questions. The analysis revealed different patterns, e.g. more boys than girls used approaches of higher sophistication, yet, overall, more boys were unsuccessful in accomplishing the task. The girls were more likely to use a visual, more concrete approach, and a lot more girls than boys did not give a sufficient explanation for the strategy used to solve the problem.

In total, students' ability to construct and use models was explicitly addressed in 17 publications (see Table 14). Between the adoption of modelling as part of the learning environment or the assessment, the studies extracted in this review are almost evenly distributed.

Table 14: Number of studies investigating 'constructing and using models'

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	1	5	2	8
Focus on assessment	1	4	2	7
Focus on both	0	2	0	2
Studies per subject [M]	2	11	4	17

5.1.6 Formulating hypotheses/ researching conjectures

The formulation of (testable) hypotheses is a major facet of scientific practice (Klahr & Dunbar, 1988; Kuhn, 1962). "In the end, there are a relatively small number of characteristics that define the enterprise we call science. The central ideas involve observation of the world and the constant testing of theories against nature, with the requirement that everything that is to be called science must be testable" (Trefil, 2008, p. 19). In this 'enterprise', meaningful and well-founded hypotheses are at the centre of scientific knowledge and progress.

With regard to students' ability in formulating a testable hypothesis, Ebenezer et al. (2011) expect students to "be able to state a hypothesis that lends itself to testing. Also, the hypothesis should be accompanied by coherent explanation(s)" (p. 103).

Burns, Okey, and Wise (1985) used multiple-choice items to analyse students' ability to identify and select testable hypotheses. Using constructed-response items, Lavoie (1999) examined the effects of adding a prediction or discussion phase at the beginning of a learning cycle. He asked students to individually write out predictions with explanatory hypotheses concerning problems in genetics, homeostasis, ecosystems, and natural selection. By introducing this phase, the authors intended to prompt students to construct and deconstruct their procedural and declarative knowledge. The evaluation of this intervention revealed significant gains in the use of process skills, logical-thinking skills, understanding scientific concepts, and scientific attitudes.

Kyza (2009) examined students' inquiry practices in considering alternative hypotheses. She analysed students' discourse, actions, inquiry products, and interactions with their teacher and peers. Despite significant learning gains when implementing a supportive learning environment, the authors point out several epistemological problems relating to students' perception of the usefulness of examining and communicating alternative explanations, e.g. about what constitutes a convincing explanation of a com-

plex problem or what counts as evidence. Their findings indicate the importance of epistemologically targeted discourse alongside guided inquiry experiences for overcoming these challenges.

The researching of conjectures is explicitly only part of the research by Reiss, Heinze, Renkl, and Groß (2008). The authors refer to three phases: (1) The production of a conjecture is the first step which includes the exploration of the problem leading to the conjecture as well as the identification of arguments to support its evidence; (2) The second step is the precise formulation of a conjecture as a basis for all future activities; (3) The third phase combines the exploration of the (precisely stated) conjecture, the identification of appropriate mathematical arguments for its validation, and the generation of a rough proof idea. In other publications, the researching of conjectures is implicitly part of the aspect 'formulating hypotheses' and is not an aspect by itself (e. g. Gobert, Pallant, & Daniels, 2010; Toth et al., 2002).

In the field of scaffold inquiry, Pine et al. (2006) asked students why an ice cube melts much more slowly in salt water than in tap water. After the replication of an experiment with ice cubes made of tap water coloured with red dye and the subsequent observations of the flow of the coloured melt water, students were asked to try to present/give/offer/provide an initial explanation for the difference in melting times. Furthermore, on successive days, students studied coloured water dropped from an eyedropper into fresh and salt water, and the effect of stirring on the difference in melting times in fresh and salt water. They again were asked to provide an explanation for the difference in melting times observed at the beginning.

In total, students' ability to formulate hypotheses or research conjectures was explicitly addressed in 38 publications (see Table 15). Despite this large number of studies, only a small number of studies disentangled this aspect of inquiry in detail. Additionally, no study in the field of technology education explicitly referred to the formulation of hypotheses as an important step of inquiry. This might be due to the nature of technological inquiry itself. In solving design problems, e.g., students generally do not have to formulate a hypothesis in its classical sense since this hypothesis would be that the design they are proposing will work and will fulfil the specified requirements and constraints.

Table 15: Number of studies investigating ‘formulating hypotheses/ researching conjectures’

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	0	17	0	17
Focus on assessment	2	12	0	14
Focus on both	0	7	0	7
Studies per subject [M]	2	36	0	38

5.1.7 Planning investigations

Similar to the formulation of hypotheses, planning an investigation is at the core of inquiry, especially in science. To develop appropriate investigations, students need to demonstrate logical connections between their conceptual understanding, their guiding hypothesis, and the research design. This means that “students should identify the scientific concepts and create a conceptual system that will guide the hypothesis and research design” (Ebenezer et al., 2011, p. 103).

The reviewed publications differ - especially with regard to the mode in which students approach the planning of their investigations. For example, McElhany and Linn (2011) used a computer simulation in which students conducted experiments to answer different investigation questions. The questions could be selected from a drop down menu or students could choose an alternative such as ‘just exploring’. While students conducted their experiments, the software logged the investigation question and the variable values that the students selected for each trial. Students’ choice of an investigation question was used to infer their intentions in each trial.

Other studies used open questions that students had to answer by planning their own, hands-on investigations, or these studies analysed differences between hands-on investigations and surrogates (e.g. simulations) (Baxter, Shavelson, Goldman, & Pine, 1992; Shavelson, Baxter, & Pine, 1991; Williams, 2012). Furthermore, White and Frederiksen (1998) investigated the effect of reflective assessments on inquiry units. Overall, students’ performance improved significantly and a controlled comparison revealed that students’ learning was greatly facilitated by reflective assessment. Interestingly, adding this metacognitive process to the curriculum was particularly beneficial for low-achieving students: Performance in their research projects and inquiry tests was significantly closer to that of high-achieving students than was the case in the control classes.

In total, the planning of investigations represents a broad research area with many different facets. 39 publications that included planning as part of a learning environment or as part of the assessment were found (see Table 16). Most of these publications stem from the field of science education (in which there is generally a larger number of

publications than in other fields) and reflect the importance of this inquiry aspect for science.

Table 16: Number of studies investigating 'planning investigations'

	Mathematics	Science	Technology	Studies per focus [N]
Focus on learning environment	2	26	0	28
Focus on assessment	0	10	0	10
Focus on both	0	0	1	1
Studies per subject [N]	2	36	1	39

5.1.8 Constructing prototypes

The construction of prototypes is predominantly addressed in publications from the field of technology education (see Table 17). Eight out of the twelve technology publications that were found investigated this issue, which shows the predominant role that this aspect plays in technological inquiry. MacDonald and Gustafson (2004) describe a project in which the children designed, made, and tested model parachutes. The intention was to analyse the characteristics of the design technology drawings that the children made before entering a construction phase. The results indicate that drawing was conceived by the children solely as representation. It was not used to indicate initial thoughts, to explore and form ideas, or as a vehicle for thinking, but was used exclusively to depict the completed product. Thus, the function of prototypes was not well understood by the children. Gustafson, MacDonald, and Gentilini (2007) extended this study to students' talking and drawing. However, no studies were identified in which students constructed prototypes in hands-on activities.

Table 17: Number of studies investigating 'constructing prototypes'

	Mathematics	Science	Technology	Studies per focus [N]
Focus on learning environment	0	2	3	5
Focus on assessment	0	0	3	3
Focus on both	0	2	2	4
Studies per subject [N]	0	4	8	12

5.1.9 Finding structures or patterns

As the Mathematical Sciences Education Board states, ‘mathematics is a science of patterns and relationships’ (Mathematical Sciences Education Board, 1990). Finding patterns or structures is seen by several authors as being closely related to processes of mathematical thinking (Lin et al., 2004; Tzur, 2007), reasoning and proving (Lin et al., 2004), problem solving (Zhang et al., 1999), and to the ability to use mental strategies and to make use of mathematical symbols (Britt & Irwin, 2008). It is considered to play an important role in students’ ability to generalize. For example, Britt and Irwin (2008) investigated the use of ‘tens frames’ in primary mathematics classrooms and found that their use and understanding supported children’s generalization ability and thus engaged them in mathematical thinking. Lin et al. (2004) analysed the relation between students’ understanding of number patterns and their abilities in proving, reasoning, and algebraic thinking. To assess students’ reasoning in geometric number patterns, they used four types of items: understanding the task, generalizing the number pattern, representing this pattern with symbols, and checking if a given number fits into this pattern. The relation between students’ ability to identify and generalize patterns was also an important aspect in the study of Zhang et al. (1999). They used two everyday situations (sorting eggs into egg cartons and estimating the number of beans in a jelly jar). Students had to identify the pattern, generalize it, and then apply it to reach the solution.

In science, the publications dealing with the aspect of finding structures or patterns are mostly related to the identification of patterns in data (Gobert et al., 2010; Ketelhut & Nelson, 2010). In the study of Gobert et al. (2010), e.g., students were required to analyse earthquake patterns, use these patterns to explain their data, and relate them to plate interactions.

Wilson, Taylor, Kowalski and Carlson (2010) compared inquiry-based and commonplace science teaching with respect to students’ knowledge, reasoning, and argumentation. They used an inquiry unit dealing with sleep disorders that was based on the BSCS 5E model. Within this model, they specifically focused on the ‘explore’ activity. Students should find patterns and negotiate those with their peers.

The small number of studies addressing this aspect of inquiry (see Table 18) might be due to the fact that it cannot be clearly separated from, e.g., ‘searching for generalizations’ in mathematics or ‘collecting and interpreting data’ in science.

Table 18: Number of studies investigating 'finding structures or patterns'

	Mathematics	Science	Technology	Studies per focus [N]
Focus on learning environment	1	5	0	6
Focus on assessment	1	0	0	1
Focus on both	2	2	0	4
Studies per subject [N]	4	7	0	11

5.1.10 Collecting and interpreting data/ evaluating results

Collecting and interpreting data, thus, the experiment itself, is certainly at the core of inquiry in science. Thousands of articles have been published about the role of the experiment in science education, as well as its benefits and relevance for students' understanding of science. Most of these publications regard the experiment as a fixed procedure; some even talk about THE scientific procedure. In several studies, experimenting means controlling variables. Therefore, fewer studies aim to describe the steps that must be taken in order to collect data that can be interpreted in a scientific way.

Designing and conducting experiments related to a hypothesis requires making a logical outline of methods and procedures, using proper measuring equipment, heeding safety precautions, and conducting a sufficient number of repeated trials to validate the results (Ebenezer et al., 2011). In addition, appropriate tools, methods, and procedures are necessary to collect and analyse data systematically, accurately, and rigorously. In some cases, this can include the use of mathematical tools and statistical software, e.g. to analyse and display data in charts or graphs or to test relationships between variables (Ebenezer et al., 2011).

Several studies in this review aimed to describe the different steps that must be taken in the collection and interpretation of data. Toth et al. (2002) used a 'design experiment' approach to develop an instructional framework that lends itself to authentic scientific inquiry. A technology-based knowledge-representation tool called 'Belvedere' enabled students to relate hypotheses to data by constructing so-called 'evidence maps'. Students formulated scientific statements by using 'hypotheses' (oval shapes) and 'data' (square shapes) and indicated the relation between these with 'for' (support) and 'against' (refutation) links. Additionally, 'and' links could be used to conjoin statements. "The results indicated that in real-life-like classroom investigations designed to teach students how to evaluate data in relation to theories, the use of evidence mapping is superior to prose writing. Furthermore, this superior effect of evidence mapping was greatly enhanced by the use of reflective assessment throughout the inquiry process." (Toth et al., 2002, p. 264).

Lubben, Sadeck, Scholtz, and Braund (2010) investigated the untutored ability of grade 10 students to engage in argumentation about the interpretation of experimental data. The authors analysed students' written interpretations of experimental data and their justifications for these interpretations based on evidence and concepts of measurement. The results revealed an initial low level of argumentation, which was considerably improved through small group discussions unsupported by the teacher. The authors concluded that several factors impact on students' argumentation ability, such as experience with practical work, or students' language ability to articulate ideas.

Further studies focused on interventions to foster students' ability in collecting and interpreting data. Mattheis and Nakayama (1988) investigated the effects of a laboratory-centred inquiry programme on laboratory skills, science process skills, and understanding. The Foundational Approaches in Science Teaching (FAST) programme was compared with a traditional science textbook approach. These results indicate that the FAST instruction especially affects laboratory skills (e.g. measuring height, area, mass, volume displacement, and calculation of density) and specific process skills (e.g. identifying experimental questions, formulating hypotheses, identifying variables), although no significant effects were found on process skills and understanding in general contexts.

Zion, Michalsky, and Mevarech (2005) investigated the effects of four different learning methods on students' scientific inquiry skills. The 2x2-design included metacognitive-guided inquiry vs. unguided inquiry and the usage of asynchronous learning networked technology vs. face-to-face interaction. The study examined general scientific ability and domain-specific inquiry skills in microbiology. The group using metacognitive-guided inquiry within asynchronous learning networked technology outperformed all other groups, while the face-to-face group without metacognitive guidance acquired the lowest scores. The authors concluded that the use of metacognitive training within a learning environment enhances the effects of asynchronous learning networks on students' achievements in science.

After having conducted an experiment, the interpretation of the obtained data is an important step. However, it seems that only few studies focus on students' ability to make logical connections between evidence and scientific explanations. Ebenezer et al. (2011) emphasized that students should be able to connect evidence from their investigations to explanations based on scientific theories.

Ruiz-Primo, Li, Ayala, and Shavelson (2004) analysed students' notebooks in science for, among other things, entries on interpreting data and/or concluding. They interpreted these entries as indicators of students' conceptual understanding. They found high and positive correlations between the derived notebook scores and other performance assessment scores. However, students' communication skills and understanding differed greatly from the expected maximum scores and did not improve over the course of the study that lasted for one school year.

The evaluation of results is included in many publications as a step of inquiry, but often only as a buzzword or by-product of a more general view on inquiry. Most of these pub-

lications stem from the field of science education (in which there is generally a larger number of publications than in other fields) and reflect the importance of this inquiry aspect for science. In total, 81 studies focused on students' ability to collect and interpret data or evaluate results, 73 of them in the field of science education (see Table 19).

Table 19: Number of studies investigating 'collecting and interpreting data/ evaluating results'

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	5	45	0	50
Focus on assessment	0	20	1	21
Focus on both	1	8	1	10
Studies per subject [M]	6	73	2	81

5.1.11 Constructing and critiquing arguments or explanations, argumentation, reasoning, and using evidence

Studies including argumentation, explanation, or reasoning as part of an inquiry process make up the largest group of studies in this review, leading to a broad array of theoretical and empirical papers. None of the other aspects is researched in the same detail.

The construct understood as argumentation varies slightly between studies. Two major conceptualizations can be identified: argumentation as students' general use of data and scientific concepts to construct arguments or explanations about the phenomenon under study (e. g. Linn, Songer, & Eylon, 1996; Smith, 1991; Strike & Posner, 1985); and argumentation as students' competitive interaction in which participants present claims, defend their own claims, and rebut the claims of their opponents until one participant (or side) 'wins' and the other 'loses' (e. g. Driver, Newton, & Osborne, 2000; Duschl, 2000; Kuhn, 1962; Latour, 1980; Toulmin, 1972). The difference between these conceptualizations depends upon the question of whether explanation and argumentation are treated as separate categories or as a single practice (Berland & Reiser, 2009).

The process of reasoning is often researched as part of an explanatory and argumentative discourse, often without any differentiation between or definition of these modes of communication (Bielaczyc & Blake, 2006; Hogan, Nastasi, & Pressley, 1999). Scardamalia and Bereiter (1994) refer to this combination as 'knowledge building'. While the combination of explanation and argumentation certainly makes sense in terms of their related goals and processes, it results in a practice with multiple instructional goals, with some of them more challenging for students than others (Berland & Reiser, 2009).

In a theoretical paper, Berland and Reiser (2009) identified “three distinct goals for constructing and defending scientific explanations: (1) using evidence and general scientific concepts to make sense of the specific phenomena being studied; (2) articulating these understandings; and (3) persuading others of these explanations by using the ideas of science to explicitly connect the evidence to the knowledge claims” (p. 29). When emphasizing the goal of persuasion, students are intended to go beyond articulating explanations by engaging with the ideas of others, receiving critiques, and revising their ideas (Driver, Newton, & Osborne, 2000; Duschl, 1990; Duschl, 2000). Thus, the goal of persuasion is to shift classroom interactions involving the practice of constructing and defending scientific explanations from ‘doing school’ to ‘doing science’ (Berland & Reiser, 2009; Jimenez-Aleixandre, Rodriguez, & Duschl, 2000).

In addition, the goal of persuasion signals the overlap to the conceptualization of argumentation as a comparative interaction. In this line of research, most studies refer to Toulmin’s model of argumentation (1958). For example, McNeill (2011) analysed students’ written argumentations and differentiated between a claim (a statement that answers a question or problem), evidence (scientific data that supports the claim), and reasoning (scientific knowledge that is/can be used to solve the problem and to explain why the evidence supports the claim). Toulmin (1958) originally included three more components of an explanation: qualifiers (statements about how strong the claim is), backings (assumptions or reasons to support the claim), and rebuttals (statements that contradict the data, warrants, qualifiers, or backings). These components have also been researched by other authors (Ruiz-Primo, Li, Tsai, & Schneider, 2010).

Studies differ not only with regard to the conceptualization of argumentation, but also with regard to the different methods used to assess students’ abilities in argumentation. While most studies use the verbal data of students’ discourse, many studies focus on students’ written argumentation. Ebenezer et al. (2011) even claim that “students should be able to write a clear scientific paper with sufficient details so that another researcher can replicate or enhance the methods and procedures” (p. 103).

A major difficulty in analysing students’ argumentations is the differentiation between the structure and components of argumentation and its accuracy. McNeill (2011) used four different codes (argument, just claim, informational text, personal narrative) to evaluate the writing style of students’ arguments. These codes were used regardless of the accuracy of the science content. Similarly, Ruiz-Primo et al. (2010) coded the accuracy of a claim as a separate measure. In addition, the authors analysed the focus (whether the claim addressed the main issues of the investigation question), and three aspects of the quality of the evidence (type: what type of evidence the student provided - anecdotal, concrete examples, or investigation-based; nature: did the student focus on patterns of data or isolated examples?; and sufficiency: did the student provide enough evidence to support the claim?) (Ruiz-Primo et al., 2010).

Toth et al. (2002) put an emphasis on analysing students’ reasoning and their final conclusions. The authors scored students’ written conclusions based on three components: (1) whether the information in the conclusion was based on information previously explored, (2) whether the conclusion contained any data to support the main hy-

pothesis, and (3) whether the conclusion indicated evidence 'going against' the accepted hypothesis (p. 275). The authors detailed different strategies the students used to structure their reasoning process. Several groups of students approached the inquiry problem by listing all the hypotheses they could think of or all the hypotheses they found in the web-based materials, and then continued with exploring data ('reasoning from hypothesis' approach to scientific reasoning). "Other groups started with data recording, and only after they had collected several data pieces did they start recording hypotheses, indicating a strategy resembling a 'reasoning from data' approach to scientific reasoning." (Toth et al., 2002, p. 280).

Wilson et al. (2010) investigated students' ability to construct and critique arguments. The authors used standardized open-ended interviews, in which students were asked to develop explanations for patterns in given data, as well as critique given explanations for those patterns. The results of a control-group comparison indicated

"that students receiving inquiry-based instruction reached significantly higher levels of achievement than students experiencing commonplace instruction. The superior effectiveness of the inquiry-based instruction was consistent across a range of learning goals (knowledge, scientific reasoning, and argumentation) and time frames (immediately following the instruction and 4 weeks later)" (Wilson et al., 2010, p. 292).

A further approach used to foster students' engagement in argumentation and explanation is to put student explanations in opposition to each other so that they are in positions to persuade one another (e. g. Bell & Linn, 2000; Hatano & Inagaki, 1991; Osborne, Erduran, & Simon, 2004). Using this approach, the role of argumentative discourse is emphasized while scientific explanations are a by-product of this process. Using a control-group design, Osborne, Erduran and Simon (2004) analysed the effect of fostering argumentation in science lessons. Teachers taught the experimental groups a minimum of nine lessons which involved socio-scientific or scientific argumentation. In addition, the same teachers taught similar lessons to a comparison group at the beginning and end of the year. Results from analysing small groups of four students engaging in argumentation over the course of 33 video-taped lessons indicated that there was improvement in the quality of students' argumentation, albeit not significant. In addition to the difficulties in fostering students' ability to engage in high-quality argumentation, the authors also concluded that supporting and developing argumentation in a scientific context is significantly more difficult than enabling argumentation in a socio-scientific context.

In mathematics, reasoning has been investigated in relation to proof competence (Heinze, Cheng, Ufer, Lin, & Reiss, 2008; Reiss et al., 2008). Boesen, Lithner, and Palm (2010) analysed the relation between the proximity of assessment tasks to the textbook and the mathematical reasoning students use. They thereby extended the relationship between reasoning and proof to understanding reasoning as "the line of thought adopted to produce assertions and reach conclusions. Argumentation is the substantiation, the part of the reasoning that aims at convincing oneself or someone else that the reasoning is appropriate". Their results show that when confronted with test tasks that are closely related to tasks in the textbook, students solved them by try-

ing to recall facts or algorithms. Surprisingly, more distant tasks mostly elicited creative mathematically founded reasoning.

All in all, 106 publications included aspects of argumentation, constructing and critiquing arguments or explanations (see Table 20). Among these studies, both the fostering of students' content knowledge by improving their argumentation skill and the fostering of argumentation skills as a merit/value on its own can be found. Again, the majority of publications can be found in the field of science.

Table 20: Number of studies investigating 'constructing and critiquing arguments or explanations, argumentation, reasoning, and using evidence'

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	6	24	0	30
Focus on assessment	4	36	1	41
Focus on both	3	31	1	35
Studies per subject [M]	13	91	2	106

5.1.12 Communication/ debating with peers

Scientific knowledge is socially and culturally constructed through negotiation (Alexopoulou & Driver, 1996; Kelly & Green, 1998). "A key element of this negotiation is oral discourse. Group processes therefore are central to understanding how knowledge is created in a science classroom" (Baker et al., 2009). These group processes go beyond the individual construction of conceptual understanding, but also build a scientific community in the classroom (Newton, Driver, & Osborne, 1999).

Cavagnetto, Hand, and Norton-Meier (2010) analysed students' interactions in small groups in a primary school utilising the Science Writing Heuristic approach. Their results indicate that students worked on tasks 98% of the time, engaging in generative talk about 25% and in representational talk about 71% of the time. The authors emphasized that students' talk was dominated by the informative function (i.e. representing one's idea) and that students spent less time on the heuristic function (i.e. inquiring through questions) or on challenging each other's ideas.

Toth et al. (2002) investigated the processes of peer communication in four ninth grade science classrooms. In their study, student groups in different classrooms shared their research results and conclusions with peer groups at the end of their inquiry. Both the peer groups and the teacher used rubrics to score each team's performance as well as the artefacts (evidence maps and reports) they developed during their inquiry. The use of rubrics was a form of reflective assessment used to provide clear expectations for optimal progress throughout the entire process of inquiry. The results showed that the

use of these reflective assessments improved students' performance in evaluating data in relation to theories.

In total, 70 studies included facets of communication processes, although the majority of them only included them as part of the learning environment (see Table 21). Interestingly, several studies which included communication as part of the assessment tended to analyse written artefacts.

Table 21: Number of studies investigating 'communication/ debating with peers'

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	5	31	1	37
Focus on assessment	2	21	0	23
Focus on both	0	10	0	10
Studies per subject [M]	7	62	1	70

5.1.13 Searching for generalizations

The facet of generalizing findings and implications as part of the inquiry process has seldom been researched. Only a small number of studies were found that explicitly entailed this step. For example, Woods, Williams, and Mc Neal (2006) analysed students' mathematical thinking as apparent in video-taped classrooms. Students' synthetic-analysing, which is Woods' et al. (2006) category to represent the production of independent generalizations, made up between 0 and 16 % of the time in different classrooms. Further analysis revealed major differences between conventional and reform-oriented classrooms in the quality of mathematical thinking.

In total, only five studies included the facet of searching for generalizations in the learning environment, only one as part of the assessment (see Table 22). However, as can be seen above, the aspect of searching for generalizations is, especially in mathematics, often closely related to the aspect of finding patterns (see 5.1.9 Finding structures or patterns).

Table 22: Number of studies investigating 'searching for generalizations'

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	2	3	0	5
Focus on assessment	1	0	0	1
Focus on both	1	1	0	2
Studies per subject [M]	4	4	0	8

5.1.14 Dealing with uncertainty

Similarly, students' dealing with uncertainty has also seldom been researched (see Table 23). Only two studies were identified that included this aspect of inquiry. One example is Liedtke's (1999) study about two projects in Victoria (British Columbia) primary schools that tried to promote positive attitudes towards mathematical tasks and problem solving. The authors used open-ended tasks with multiple solutions to stimulate curiosity, group discussions, and risk taking. The case study revealed positive changes in the classroom behaviour of several students; they became more willing to ask questions and volunteer answers.

Table 23: Number of studies investigating 'dealing with uncertainty'

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	1	1	0	2
Focus on assessment	0	0	0	0
Focus on both	0	0	0	0
Studies per subject [M]	1	1	0	2

5.1.15 Problem solving

Problem solving is part of the inquiry process but it affects more than one aspect of IBE. Usually, several aspects are combined within the studies found. For example, in mathematics education, Chang, Wu, Weng, and Sung (2012) investigated students' problem posing by analysing four phases: (1) 'posing problems' (problem-posing activity); (2) 'planning' (verifying self-posed problems and revising self-posed problems according to the teacher's feedback); (3) 'solving problems' (solving posed problems); and (4) 'looking back' (obtaining teacher's feedback and getting new ideas to create new problems). This example illustrates that the process of problem solving covers more than just identifying a problem. The phases originally derive from Polya's (1957)

work which defined the phases: understanding, planning, carrying out the plan and looking back. Other studies also refer to this definition (e. g. Lorenzo, 2005). As students have to learn the complex process of problem solving, research projects investigate the methodological approach of scaffolding (e. g. Simons & Klein, 2007).

In total, 13 studies from mathematics and science education were found (see Table 24). However, none were found in the field of technology education.

Table 24: Number of studies investigating ‘problem solving’

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	1	0	0	1
Focus on assessment	5	7	0	12
Focus on both	0	0	0	0
Studies per subject [M]	6	7	0	13

5.1.16 IBE and inquiry process skills in general

While many of the reviewed publications focused on the development and evaluation of learning environments for IBE or the assessment of certain aspects of IBE, some studies took a broader perspective on IBE and inquiry process skills. These studies used inquiry as a ‘black box’ category. The problem is that these approaches do not allow “for distinctions between activities that are guided more by the teacher and those guided more by the student” (Furtak and Seidel et al., 2012, p. 304). While mostly taking inquiry as a single construct, the studies differ in their research intentions.

A central field of research is the question of whether inquiry skills and content knowledge can be separated within a domain. Gobert et al. (2010), for example, designed a supplemental instructional and assessment module for enhancing middle school students’ content knowledge and inquiry skills in the domain of geosciences. By using factor analysis, the authors intended to demonstrate the separation of content knowledge and inquiry skills. They found five factors, some reflecting content knowledge exclusively, some representing inquiry skills exclusively, and some including both content and inquiry within the same strand. The authors concluded that content knowledge and inquiry skills can partly be separated, but are also partly interrelated.

Beyond the analysis of the ‘construct’ inquiry, several publications investigated the comparison of IBE with other forms of teaching, often referred to as ‘direct’, ‘traditional’ or ‘commonplace’ teaching. For instance, Cobern et al. (2010) designed a controlled experimental study which compared inquiry instruction and direct instruction in realistic science classroom situations in middle school grades. The results indicate that “inquiry and direct methods led to comparable science conceptual understanding in roughly

equal instructional times. Gain differences between instructional modes were not statistically significant within the observed natural variation of students, teachers and classrooms.” (Cobern et al., 2010, p. 92).

In contrast, Furtak and Seidel et al. (2012) critique that “insufficient attention has been given to the operationalization of the inquiry construct in the case of prior meta-analyses of inquiry-based teaching and that this has masked important differences in the efficacy of distinct features of this instructional approach” (p. 304). Thus, the generalizability of the inferences one can make after combining effect sizes depends on “the way that the sample of students has been selected, the way that the outcome variable has been measured, and the way that the treatment under investigation has been defined” (Furtak and Seidel et al., 2012, p. 304). Therefore, Ruiz-Primo et al. (2012) present an approach which considered three aspects of quality in terms of the assessment items: (1) representing the curriculum content, (2) reflecting the quality of instruction, and (3) having formative value for teaching.

But, of course, there are studies which provide evidence that IBE has positive effects on students’ learning. For example, Gibson and Chase (2002) concluded that “a 2-week summer science programme which used an inquiry-based approach may have helped middle school students, who had a high level of interest in science, maintain their interest during their years in high school” (p. 704). Additionally, Hofstein, Navon, Kipnis, and Mamlok-Naaman (2005) present evidence that students can improve their ability to ask relevant questions as a result of gaining experience with inquiry-type experiments. Furthermore, students who were involved in these experiences were more motivated to pose questions regarding scientific phenomena. Even if the results are related to the aspect of identifying questions, general process skills are also included in the experiments.

Baker et al. (2009) developed the Communication in Science Inquiry Project which aims to create science classroom discourse communities (SCDCs): “a community of learners who create a culture that reflects literacy practices in science. The culture promotes norms of interaction that foster scientific discourse, use of notebooks, scientific habits of mind, and scientific language acquisition through inquiry. Central to a SCDC are experiences for students to communicate, create, interpret, and critique scientific arguments using scientific principles and data from inquiry activities.” (Baker et al., 2009, p. 260). The evaluation of this project focused on student perceptions of the teacher’s use of instructional strategies (i.e. scientific inquiry, learning expectations, writing, and use of science notebooks).

Further studies analysed the effect of curricular reforms. For example, Reys, Reys, Lapan, Holiday, and Wasman (2003) investigated the impact of standards-based mathematics curriculum material for middle grades on student achievement. The mathematics section/part of the Missouri Assessment Program (MAP) was used to measure students’ achievement. This included aspects of IBE, for example, defending data predictions, recognizing dependent and independent variables, using diagrams, patterns or functions in problem solving, and solving problems by using strategies (Reys et al., 2003). Differences were found between students who used the standards-based mate-

rials for at least 2 years and students from comparison districts who used other materials.

In total, 55 of the reviewed publications included a broader focus on IBE in STM; most of them in science education (see Table 25).

Table 25: Number of studies investigating 'IBE and inquiry process skills in general'

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	0	32	2	34
Focus on assessment	2	14	3	19
Focus on both	0	2	0	2
Studies per subject [M]	2	48	5	55

5.1.17 Knowledge/ achievement/ understanding

There are 96 studies that focused on the assessment of students' knowledge, achievement or understanding in the context of IBE, mainly in science education (see Table 26). This indicates that these variables are seen as control variables or dependent variables which are presumably influenced by any kind of an intervention including inquiry-based learning environments (e. g. Birchfield & Megowan-Romanowicz, 2009; Chen & Klahr, 1999; Santau, Maerten-Rivera, & Huggins, 2011).

The use of central examinations is one example for a frequently used assessment strategy. Schneider, Krajcik, Marx, and Soloway (2002) investigated the effect of a project-based science programme using the twelfth grade 1996 National Assessment of Educational Progress (NAEP) science test. This test includes the assessment of knowledge or understanding, as well as the assessment of aspects of scientific inquiry.

As the assessment of knowledge, achievement, and understanding is strongly related to the assessment methods and instruments, they are presented in Section 5.2 Which types of assessment are employed in the study?

Table 26: Number of studies investigating 'knowledge/ achievement/ understanding

	Mathematics	Science	Technology	Studies per focus [M]
Focus on learning environment	2	0	0	2
Focus on assessment	6	81	5	92
Focus on both	0	2	0	2
Studies per subject [M]	8	83	5	96

5.1.18 Further aspects focused on or assessed by the studies

Despite the broad definition of inquiry which led the focus of this review, several publications included further aspects. Some of these aspects are domain-specific, for example, proof competence as part of inquiry in mathematics education (Heinze et al., 2008; Lin et al., 2004; Reiss et al., 2008). Representing data by graphs (Burns, Okey, & Wise, 1985; McElhaney & Linn, 2008), visualizing data, drawing, and graphing (Gobert et al., 2010; Ruiz-Primo & Furtak, 2007), or using visualizations in general (Hamilton, Nussbaum, & Snow, 1997) are also partly linked to mathematics but, without doubt, these aspects are relevant for the domains of science and technology too.

In addition, epistemological aspects were also addressed in several publications. Epistemic understanding was either regarded as domain-specific, e.g. the nature of science (Akerson & Donnelly, 2010; Herrenkohl, Palincsar, DeWater, & Kawasaki, 1999; Khishfe, 2008; Vellom & Anderson, 1999), or as more general, e.g. epistemic understanding (Ryu & Sandoval, 2012) or the nature of modelling (Schwarz & White, 2005).

Interdisciplinary relevance is also significant for abilities such as divergent thinking and creativity (Doppelt, 2009; Kwon, Park, & Park, 2006) or critical thinking (Kim et al., 2012). However, these aspects are not only limited to the domains of STM. In fact, they are more closely related to aspects of general cognitive abilities.

Beyond these cognitive abilities, affective aspects are also addressed in certain publications, although to a smaller extent. Enjoyment, interest, value, self-efficacy (Schukajlow et al., 2012), motivation (Butler & Lumpe, 2008; Shavelson et al., 2008), and confidence (Klahr, Triona, & Williams, 2007), but also attitudes towards science (Burghardt, Hecht, Russo, Lauckhardt, & Hacker, 2010; Gibson & Chase, 2002; Lavoie, 1999; Mislter Jackson & Songer, 2000; White & Frederiksen, 1998) are analysed in relation to different aspects of inquiry.

5.2 Which types of assessment are employed in the study?

First of all, for the analysis of the assessment practices, the frequency of the assessment types used was compared between science, technology and mathematics. Table 27 shows the results. In three quarters of all studies, methods of summative assessment were employed. Methods of formative assessment were not very common among the empirical studies found, especially in science education. However, nearly 15% of the studies in science combined methods of summative and formative assessment. Furthermore, in science education, some studies dealt with embedded assessment (see Table 28). Peer- and self-assessment played a subordinate role. In combination with IBE, neither was explored very often. In contrast, rubrics were a common instrument used for the evaluation and analysis of varying assessment situations.

When comparing the results, one has to keep in mind that there were only 13 studies in technology and 30 in mathematics, but 148 in science. This made it difficult to determine subject-specific main focuses, especially in technology and mathematics.

Table 27: Assessment practices by subject

Type of assessment	Science		Technology		Mathematics	
	N	%	N	%	N	%
Summative assessment	108	73.0	10	76.9	23	76.7
Formative assessment	9	6.1	2	15.4	6	20.0
Summative and formative assessment	22	14.8	1	7.7	-	-
Neither summative nor formative assessment	9	6.1	-	-	1	3.3
Total	148	100.0	13	100.0	30	100.0

Table 28: Character of the assessment

Character of assessment	Science		Technology		Mathematics	
	N	%	N	%	N	%
Embedded assessment in combination with summative assessment	5	3.4	1	7.7	1	3.3
Embedded assessment in combination with summative and formative assessment	8	5.4	-	-	-	-
Feedback	12	8.1	-	-	2	6.7
Peer-assessment	8	5.4	1	7.7	1	3.3
Self-assessment	11	7.4	1	7.7	4	13.3
Rubrics	51	34.5	6	46.2	5	16.7

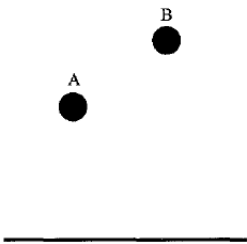
In view of the objectives, it is important to know which assessment methods are frequently employed in the studies and which assessment methods are less common. Furthermore, the purpose of the assessment methods is of importance. In the following three chapters, these aspects are addressed for every subject by analysing the purpose of each assessment method exemplarily. One has to note that the focus of the search strategy was on IBE and assessment methods. Therefore, most of the studies

using assessment methods have to be seen against the background of IBE and related aspects and competences.

5.2.1 Science

Multiple-choice items and constructed-response or open-ended items used as a summative assessment tool dominate the assessment methods in research on IBE in science education (see Table 30). The reasons are obvious as these items have many advantages. In particular, the analysis of multiple-choice items is more objective and the results are easier to compare and to interpret than other more complex assessment methods. Figure 1 shows an example from a research project in physics education by White and Frederiksen (1998) which combined both item formats for the assessment of physics knowledge.

Imagine that you drop two identical balls from different heights. Both balls are dropped at exactly the same time.



Which ball hits the floor first? (circle your choice below)

- (A) The lower ball.
- (B) The higher ball.
- (C) Both balls hit the floor at the same time.

Explain the reasons for your choice: _____

Which ball is going faster when it hits the floor? (circle your choice below)

- (A) The lower ball.
- (B) The higher ball.
- (C) Both balls are going the same speed when they hit the floor.

Explain the reasons for your choice: _____

Figure 1: A sample gravity problem from a physics test (White & Frederiksen, 1998, p. 60)

However, even though the items have advantages in view of summative assessment, they are less frequently used for formative assessment. Four studies used multiple-choice items and five studies constructed-response or open-ended items. Hickey and Zuiker (2012) provided an example of open-ended items supporting feedback conversations (see Figure 2). The explanations were the basis of the following conversations in biology learning.

Section 3A: Assessment
From Offspring Phenotypes to Mode of Inheritance I

In dragons, a single gene with two possible alleles determines visual ability. Use what you know about pedigrees to help you figure out these other things about visual ability.

female sighted

male sighted

female blind

male blind

PEDIGREE FOR VISUAL ABILITY IN DRAGONS

1.1 Do the alleles for visual ability show complete dominance or incomplete dominance?

1.2 Explain what it is about the pedigree that distinguishes between complete and incomplete dominance.

2.1 Is the allele for blindness dominant, recessive, or incompletely dominant to the sighted allele? (Hint: use the circled part of the pedigree.) _____

2.2 Explain what it is about the circled part of the pedigree that tells you the answer, and explain why the parents and offspring that were not circled do not tell you the answer.

Figure 2: Formative assessment item on dominance relationships (Hickey & Zuiker, 2012, p. 24)

To assess students' understanding of key concepts, concept maps instead of items are often used for a summative assessment. For example, Brandstädter, Harms, and Großschedl (2012) investigate concept maps as an assessment tool for system thinking in biology education. As the process of the concept map development is quite complex, some approaches use computer-assisted methods (e. g. Schaal, Bogner, & Girwidz, 2010).

On the other hand, concept maps can be used for formative assessment. In this case, the focus lies on checking students' progress in understanding key concepts at several times during a treatment (e. g. Furtak et al., 2008). The analysis of concept maps can be organised by rubrics as shown in Table 29 (e. g. Nantawanit, Panijpan, & Ruenwongsa, 2012).

In general, it is important to train students in the procedure of making a concept map (Nantawanit et al., 2012). One possible way is the think-pair-share method: First, students make an individual map, then, they build a map in a small group, and finally, they construct a concept map as a class (e. g. Furtak et al., 2008). Another common method is to give the concepts and linking words to the students (see Figure 3). Both approaches have a more formative than summative character.

Table 29: Holistic concept mapping scoring rubric (Nantawanit et al., 2012)

Score	Content	Logic and Understanding	Presentation
5	All relevant concepts (14) of plant responses to biological factors are correct with multiple connections.	Understanding of facts and concepts of plant responses to biological factors is clearly demonstrated by correct links.	Concept map is neat, clear, and legible, has easy-to-follow links and has no spelling errors.
4	Most relevant concepts (10-13) of plant responses to biological factors are correct with multiple connections.	Understanding of facts and concepts of plant responses to biological factors is demonstrated by a few error links.	Concept map is neat, clear, and legible, has easy-to-follow links and has some spelling errors.
3	Few relevant concepts (6-9) of plant responses to biological factors are correct with two or more connections.	Understanding of facts and concepts of plant responses to biological factors is demonstrated but with some incorrect links.	Concept map is neat, legible but with some links difficult to follow and has some spelling errors.
2	Few relevant concepts (3-5) of plant responses to biological factors are correct with no connection.	Poor understanding of facts and concepts of plant responses to biological factors with significant errors.	Concept map is untidy with links difficult to follow and has some spelling errors.
1	1-2 relevant concepts are linked via the linking words.		

<u>Set of given concepts:</u>	<u>Set of linking words:</u>
Blue mussel	Becomes
Byssus	Develops
Egg	Eats
Eider duck	Protects
Human	Displaces
Larvae	Breeds
Mussel breed	Sticks together
Oyster	Lives in
Sea star	Builds
Young mussel	Warms
Worm	Pulls

Figure 3: Given concepts and linking words for the construction of a concept map in biology (Brandstädter et al., 2012, p. 2167)

The publication about the advantages of mind maps does not report any empirical data (Goodnough & Long, 2006). However, the authors state that mind mapping is a tool that can be used to ascertain students' developing ideas about scientific concepts. Furthermore, similar to concept mapping, the technique makes the exploration of prior knowledge possible, as well as an assessment of students' overall performance from the viewpoint of specific learning outcomes.

Notebooks are a science-specific assessment method used in formative assessment. They are supposed to monitor and facilitate students' understanding of complex scientific concepts and especially inquiry processes. To achieve this, the method includes the collection of student writing before, during, and after hands-on investigations (Aschbacher & Alonzo, 2006). As notebooks are an embedded part of the curriculum, they can obtain information about students' understanding at any point without needing additional time and expertise to create quizzes.

Baxter, Shavelson, Goldman, and Pine (1992) were able to confirm that notebooks are a valid tool for a summative assessment of hands-on activities. They compared the analysis of notebooks with results from an observation and from multiple-choice items. However, field observations are a more reliable tool than notebooks.

As well as notebooks or science journals, portfolios summarize the inquiry process, for example, in a laboratory or learning environment (Dori, 2003; Zhang & Sun, 2011). Portfolios are normally compiled individually to measure knowledge growth over a certain period of time. Thus, they are used for summative assessment.

Hands-on activities like experiments are often used as for performance assessment in a summative manner. They are supposed to be an alternative to more traditional paper and pencil assessment methods (Shavelson et al., 1991). However, in comparison to these methods, performance assessment requires more complex scoring or evaluation systems. Baxter et al. (1992) recommend field observations instead of notebooks.

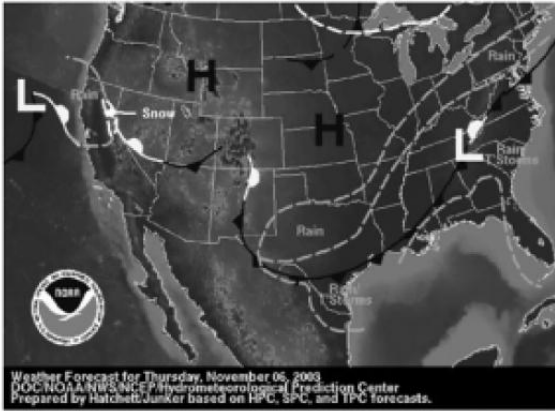
For example, Hofstein, Navon, Kipnis, and Mamlok-Naaman (2005) investigated the ability of students to ask questions related to their observations and findings in an inquiry-type experiment. Providing students with opportunities to engage in inquiry-type experiments in the chemistry laboratory improved their ability to ask high-level questions, to hypothesize, and to suggest questions for further experimental investigations (Hofstein et al., 2005). In this case, the experiments were a method to provoke a more realistic assessment situation. The purpose of the study of Kelly, Druker, and Chen (1998) was quite similar; they investigated the reasoning processes students use while solving electricity performance assessments (Kelly et al., 1998). In contrast, Ruiz-Primo, Li, Tsai, and Schneider (2010) conducted a study on various types of assessment and their advantages compared to others. With regard to performance assessment, students were asked to design and conduct an investigation to solve a problem with given materials.

There was one study which really meets the objectives of ASSIST-ME (Pine et al., 2006). By conducting a performance assessment, the inquiry skills 'planning an inquiry', 'observation', 'data collection', 'graphical and pictorial representation', 'inference' and 'explanation based on evidence' were measured.

Among the publications, quizzes were only used by one research group (Cross, Taasobshirazi, Hendricks, & Hickey, 2008; Hickey et al., 2012; Taasobshirazi & Hickey, 2005; Taasobshirazi, Zuiker, Anderson, & Hickey, 2006). Ultimately, the quizzes developed by Hickey, Taasobshirazi and Cross (2012) were a combination of multiple-choice and open-ended items (see Figure 4). Each quiz consisted of three to four two-part items, with the first part requiring a short answer, and the second part requiring an explanation to support that answer. Students completed the quizzes individually. Then, pairs of students joined with other pairs to engage in a structured argumentation review routine to discuss the answers. The questions focused on activities completed during several units of a software-based learning environment. Each quiz was aligned to the specific activities the students had completed for that particular unit.

Figure 5 shows guidelines for the feedback conversation which structured the argumentation process.

In *Weather or Not?*, you predicted the weather using satellite images and other kinds of weather maps. The weather map below shows the weather forecast for Thursday, November 6, 2003. The temperature on that day was 73°. Using the map, predict the weather in northern Georgia for Friday, November 8, 2003.



1a What will the weather be like on Friday?

- a low pressure system is moving in with warm temperatures
- a high pressure system is moving in with warm temperatures
- there will be a low pressure system with the possibility of rain
- there is the possibility of a cold front but it should be sunny

1b Explain your answer:

Figure 4: Activity-oriented quiz (Hickey et al., 2012, p. 1247)

Usually, conversations or discussions are carried out to enhance students' argumentation, reasoning or communication skills. Mainly, the discussions take place in small groups. These students' discussions indicate an alternative didactical approach in contrast to the more traditional discourse where the teacher dominates classroom dialogue mainly to transmit information and requires students to use oral discourse only to show acquired knowledge. In order to distinguish between the approaches, it is important to know that the term 'discourse' includes a broader set of practices than the language-intensive ones usually associated with discussion or argumentation (van Aalst & Mya Sioux Truong, 2011).

Feedback conversation guidelines as shown in Figure 5 support collective discourse (Hickey et al., 2012; Hickey & Zuiker, 2012). This approach suggests that the most valuable function of feedback is fostering participation in discourse. Furthermore, formative discussions can help students in IBE. For example, the consideration of multiple solutions can be followed by a classroom discussion in which students present their solutions, share information, reflect on things, raise questions, and receive feedback on their proposed solutions (Valanides & Angeli, 2008).

- FOR EACH QUIZ ITEM**
1. **EXPLAIN AND COMPARE EACH ANSWER:**
Everybody should *share* what they wrote (provide claims) and explain why they wrote it **and** how they knew it (warrant claims using data).
 2. **REACH INITIAL CONSENSUS:**
Through evaluation of the warrants, the group should try to agree on the most sensible claim, agree to disagree, or agree that for some questions there could be multiple answers. Everyone should understand why one (or more) claims are sensible.
 3. **REVIEW ANSWER EXPLANATION:**
As a group, read aloud the answer explanation sheet. Use the written explanation and the *language of science* to provide warrants and data to help your group agree on the most sensible claims.
 4. **CONFIRM GROUP UNDERSTANDING:** Everyone should understand and agree on a final sensible claim. Take your time! It's more important that everyone understands than to finish quickly.

Figure 5: Feedback conversation guidelines (Hickey et al., 2012, p. 1248)

Apart from a formative character, one can use discussions with a more summative character with regard to the assessment. One evaluating study used students' small group discussions to address four aspects of IBE: "(a) expressing and comparing prior knowledge on a specific phenomenon or situation to create a common ground for the collaborative construction of knowledge; (b) formulating and comparing hypotheses before performing an experiment; (c) examining empirical data in the light of previous predictions; (d) and making a shared synthesis to propose a final explanation for an examined phenomenon" (Mason, 2001, p. 315). A qualitative analysis of the collected data was then carried out to analyse the collaborative discourse-reasoning.

In biology education, students are trained in discussing socio-scientific issues – such as whether to allow human gene therapy (Nielsen, 2012). This kind of issue calls for a discussion about what to do and not merely about what is true. Socio-scientific issues seem to be a good theme or opportunity for discussions. The first and final lessons of an intervention by Osborne et al. (2004) were devoted to the discussion of whether zoos should be permitted, whereas the remaining lessons were devoted solely to discussion and arguments of a scientific nature. The authors used a generic framework for the materials that supported and facilitated argumentation in the science classroom. The starting point was a table of statements on a particular topic in science which was given to students. They were asked to say whether they agreed or disagreed with the statements and argue for their choices. Based on this starting point, one can build discussions and initiate IBE learning.

Ruiz-Primo's and Furtak's (2006) approach to exploring teachers' questioning practices is based on viewing whole-class discussions as assessment conversations. Assessment conversations consist of four-step cycles: 1. The teacher elicits a question; 2. The student responds, 3. The teacher recognizes the student's response; 4. The teacher uses the information collected to assist/initiate student learning. Thus, these kinds of conversations permit teachers to gather information about the status of students' con-

ceptions, mental models, strategies, language use, or communication skills and enable them to use these to guide instruction.

Closely related to discourses, assessment conversations or accountable talks can also be employed as assessment methods, just like field notes or video tapes. As well as observations or field notes, video and audio tapes are mostly conducted as a form of summative assessment. These methods are used with a variety of purposes because they allow the measurement of certain constructs and the description of learning and teaching processes in retrospect.

Communication processes are often observed, for example, to assess students' argumentation within discussions or classroom interaction (e. g. Abi-El-Mona & Abd-El-Khalick, 2006; Lavoie, 1999). Moreover, observations provide records of the order in which students carried out certain activities in learning environments and the time they spent on these activities (e. g. Hamilton et al., 1997; Kubasko, Jones, Tretter, & Andre, 2008). For some reasons, it is necessary to combine both purposes. For example, in the study of Harskamp, Ding and Suhre (2008) the observers' task was to use observation log files to document and log individual student's time on the task, as well as cooperative actions and the type of interaction.

The application of video and audio tapes aims more at the observation and analysis of learning and teaching processes than at the assessment of learning or teaching outcomes (Valanides & Angeli, 2008), even though they are generally used for summative assessment. Moreover, they are used as a further tool in addition to other research methods or in explicit combination with other tools, e.g. field notes, written materials or multiple-choice pre- and post-tests (e. g. Vellom & Anderson, 1999). Which tool is used depends on the objectives and design of the study.

The time scale of video or audio-taped classroom or learning environment interaction varies. Some studies collected data daily from whole class sessions for longer periods. However, some studies only collected data from selected student groups for a few hours (e. g. Southerland, Kittleson, Settlage, & Lanier, 2005).

In order to achieve a deeper analysis, video or audio tapes are usually transcribed using repeated viewings or hearings of video or audio segments (e. g. Aguiar et al., 2010). Sometimes, annotations about important contextual factors such as actions, gestures, and other classroom interactions were added to the transcripts (e. g. Vellom & Anderson, 1999).

One major purpose of video and audio tapes is the observation of class or group interaction, discussions or dialogues (Schnittka & Bell, 2011; Southerland et al., 2005). For example, Shemwell and Furtak (2010) investigated the quality of argumentation in classroom discussion by analysing the support of argumentation by evidence. In another study, McNeill (2009) analysed the instructional practices teachers use to introduce scientific explanations by videotaping classroom interaction. Another purpose is the observation of students' performance in a certain task (Sampson, Grooms, & Walker, 2011).

In cases in which only audio tapes were used, the focus was on the talk especially on the amount of on/off task talk and the categorization of task talk (Cavagnetto et al., 2010). Chin and Teou (2009) audiotaped conversation from one group to provide a record of students' thinking in a form that was accessible to the teacher for monitoring and feedback purposes. This is an example of a formative use of audio tapes. Students' assertions and questions had formative potential as they encouraged discourse by drawing upon each other's ideas.

Even though there are so many publications that include video and audio tapes, the purpose of their use and the way in which they can be analysed often remain unclear (e. g. Harris, McNeill, Lizotte, Marx, & Krajcik, 2006; Tytler, Haslam, Prain, & Hubber, 2009). Obviously, video and audio tapes provide background information that is not described and explained in detail.

In addition, field notes are a method which combines both observations and video or audio tapes. For instance, they provide general descriptions of the most salient instructional events during an observed session (e. g. Abi-El-Mona & Abd-El-Khalick, 2006) or provide information about events that occur outside the range of a video camera (e. g. Ryu & Sandoval, 2012). Furthermore, field notes can be taken as events unfold, and recorded with time indices for later matching with video segments (e. g. Vellom & Anderson, 1999). However, in view of performance assessment, notebooks are a reliable tool that can be used for formative teacher feedback (Ruiz-Primo et al., 2004).

Appendix A. Interview questions that elicited argument

Actual questions from interview transcripts included:

- What do you think about biotechnology?
- Can you tell me about any problems that might come from biotechnology?
- What are your feelings about cloning?
- Can you see any other problems with cloning?
- What about cloning of extinct animals or endangered species?
- Can you tell me about any problems with genetics engineering?
- What do you feel about genetically modified (GM) foods?
- Do you have any views on forensic testing?

Figure 6: Examples of questions for a semi-structured interview (Dawson & Venville, 2009, p. 1445)

Similar to any kind of observation, the objectives of interviews are also manifold and, similar to field notes, they are an additional tool that is usually combined with other methods such as observation, video tapes (e. g. Berland, 2011) or audio tapes (e. g. Dawson & Venville, 2009). Interviews are an assessment and research method that is usually qualitatively analysed. Therefore, in most of the studies, only some students from the total samples were interviewed in order to acquire additional information on the explored aspects. For example, after responding to a questionnaire, students were asked to explain their answers in order to gather information about existing misconceptions (White & Frederiksen, 1998). Furthermore, pre- and post-interviews provide another possibility for evaluating the intervention part of a case study (Berland, 2011).

A possibility which makes interviews and especially their content more comparable is the realization of semi-structured interviews, as they were conducted by Dawson and Venville (2009) who, for example, asked questions about students' understanding and views of biotechnology, cloning, and genetic testing for diseases.

Ash (2008) gives an example of how interviews can be used as a kind of formative assessment. An interviewer provided biological dilemmas as thought experiments, described the context, and then asked questions. The formative character was introduced by further questions or hints: After the student had answered, the interviewer provided a hint if the student was on the wrong track or a challenge if the student gave an appropriate answer. The hint determined what a student might achieve with appropriate help, while the challenge helped determine whether understanding was robust. The goal was to measure students' competence in solving biological dilemmas (Ash, 2008). Unfortunately, the purposes of the interviews were often not explained in detail within the publications (e. g. Tytler et al., 2009). Therefore, it is difficult to provide a detailed overview.

Artefacts are used quite rarely as an assessment method for research on IBE in STM. Only two publications referred to their use when collected as written material (Harris et al., 2006; Kyza, 2009).

Rubrics are a common tool for the analysis of several assessment methods, as described above. Figure 7 shows another example which illustrates the use of rubrics in students' self-assessment to enhance students' self-reflection with regard to the learning process.






RUBRIC FOR SELF ASSESSMENT (ACTIVITY 13 – INITIAL IDEA GENERATION) My name and surname: _____ Date of self assessment: _____ Be honest with yourself and indicate if you could do activity 13. Cross (X) the dog of your choice. The first dog indicates that you could easily do the work. The second dog indicates that the work was a bit difficult but that you completed it satisfactory. The third dog indicates that you could not do the work at all.		
Stage outcomes for activity 13 Consider as many as possible solutions for the problem (Specifically look at the three possible solutions you have generated)	I assess myself 	
RUBRIC FOR SELF ASSESSMENT (ACTIVITY 13 – INITIAL IDEA GENERATION) My name and surname: _____ Date of self assessment: _____ Be honest with yourself and indicate if you could do activity 13. Cross (X) the face of your choice. The meaning of each face is explained with each statement.		
I assess myself		
Consider as many as possible solutions for the problem (Specifically look at the three possible solutions you have generated)		
No solutions for the problem is considered 	One or two good solutions for the problem are considered 	One, two or three excellent solutions for the problem are considered 

Figure 7: Assessment rubric for self-assessment (van Niekerk, Piet Ankiewicz, & Swardt, 2010, p. 213)

Table 30: Frequency of assessment methods in the studies from the field of science education

Assessment method	SA [N]	References	FA [N]	References
Multiple-choice	63	Acar & Tarhan, 2007; Baxter et al., 1992; Blanchard et al., 2010; Burns, Okey, & Wise, 1985; Chen & Klahr, 1999; Cobern et al., 2010; Cross et al., 2008; Ding & Harskamp, 2011; Dori & Herscovitz, 1999; Ebenezer et al., 2011; Furtak & Ruiz-Primo, 2008; Geier et al., 2008; Gerard, Spitulnik, & Linn, 2010; Gibson & Chase, 2002; Gijlers & Jong, 2005; Gotwals & Songer, 2009; Hamilton et al., 1997; Harris et al., 2006; Hickey et al., 2012; Hmelo, Holton, & Kolodner, 2000; Jang, 2010; Ketelhut & Nelson, 2010; Kyza, 2009; Lavoie, 1999; Lee & Liu, 2010; Lee, Brown, & Orrill, 2011; Linn, 2006; Liu, Lee, & Linn, 2011; Liu, O. L., Lee, H.-S., & Linn, M. C., 2010a; Liu, O. L., Lee, H.-S., & Linn, M. C., 2010b Mattheis & Nakayama, 1988; McNeill & Krajcik, 2007; McNeill, 2009; Mistler Jackson & Songer, 2000; Nantawanit et al., 2012; Oh et al., 2012; Osborne, Simon, Christodoulou, Howell-Richardson, & Richardson, 2013; Pifarre, 2010; Pine et al., 2006; Repenning, Ioannidou, Luhn, Daetwyler, & Repenning, 2010; Rivet & Kastens, 2012; Rivet & Krajcik, 2004; Ruiz-Primo & Furtak, 2006; Ruiz-Primo & Furtak, 2007; Ruiz-Primo et al., 2010; Ruiz-Primo et al., 2012; Ryu & Sandoval, 2012; Schneider et al., 2002; Schnittka & Bell, 2011; Schwarz & White, 2005; Shavelson et al., 1991; Shavelson et al., 2008; Shymansky, Yore, & Anderson, 2004; Silk et al., 2009; Simons & Klein, 2007; Spires, Rowe, Mott, & Lester, 2011; Steinberg, Cormier, & Fernandez, 2009; Taasoobshirazi & Hickey, 2005; Taasoobshirazi et al., 2006; Tsai, Hwang, Tsai, Hung, & Huang, 2012; Wilson et al.,	4	Aschbacher & Alonzo, 2006; Birchfield & Megowan-Romanowicz, 2009; Hickey et al., 2012; White & Frederiksen, 1998

		2010; Wong & Day, 2009; Young & Lee, 2005; Zion et al., 2005		
Constructed-response / Open-ended	65	Acar & Tarhan, 2007; Brown et al., 2010; Ding & Harskamp, 2011; Dori, 2003; Dori & Herscovitz, 1999; Furtak & Ruiz-Primo, 2008; Geier et al., 2008; Gerard et al., 2010; Gijlers & Jong, 2005; Gobert et al., 2010; Gotwals & Songer, 2009; Hamilton et al., 1997; Harris et al., 2006; Harskamp et al., 2008; Hickey et al., 2012; Hickey & Zuiker, 2012; Hmelo et al., 2000; Jang, 2010; Kaberman & Dori, 2009; Khishfe, 2008; Kubasko et al., 2008; Kyza, 2009; Lee & Liu, 2010; Lee et al., 2011; Lin & Mintzes, 2010; Linn, 2006; Liu et al., 2011; Liu, O. L. et al., 2010a; Liu, O. L. et al., 2010b; Lorenzo, 2005; Lubben et al., 2010; Mason, 2001; Mattheis & Nakayama, 1988; McElhanev & Linn, 2008; McNeill & Krajcik, 2007; McNeill, 2009; McNeill, 2011; Mistler Jackson & Songer, 2000; Pifarre, 2010; Rivet & Kastens, 2012; Rivet & Krajcik, 2004; Ruiz-Primo et al., 2010; Ryu & Sandoval, 2012; Schneider et al., 2002; Schwarz & White, 2005; Shavelson et al., 1991; Shavelson et al., 2008; Shemwell & Furtak, 2010; Shymansky et al., 2004; Siegel, Hynds, Siciliano, & Nagle, 2006; Simons & Klein, 2007; Stecher et al., 2000; Steinberg et al., 2009; Tsai et al., 2012; Valanides & Angeli, 2008; van Aalst & Mya Sioux Truong, 2011; Veal & Chandler, 2008; Wilson & Sloane, 2000; Wilson et al., 2010; Winters & Alexander, 2011; Wirth & Klieme, 2003; Wong & Day, 2009; Yoon, 2009; Young & Lee, 2005; Zion et al., 2005	5	Hickey et al., 2012; Hickey & Zuiker, 2012; van Niekerk et al., 2010; White & Frederiksen, 1998; Wilson & Sloane, 2000
Concept map	8	Brandstädter et al., 2012; Brown et al., 2010; Butler & Lumpe, 2008; Dori, 2003; Nantawanit et al., 2012; Schaal et al., 2010; Vasconcelos, 2012; Yin, Vanides, Ruiz-Primo, Ayala, & Shavelson, 2005	3	Furtak & Ruiz-Primo, 2008; Furtak et al., 2008; Okada & Shum, 2008; Yin et al., 2005



Mind map	1	Goodnough & Long, 2006	-	-
Portfolios	2	Dori, 2003; Zhang & Sun, 2011	-	-
Notebook	8	Baxter et al., 1992; Kelly et al., 1998; Ruiz-Primo et al., 2004; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002; Ruiz-Primo et al., 2010; Shavelson et al., 1991; Simons & Klein, 2007; So, 2003	4	Aschbacher & Alonzo, 2006; Tytler et al., 2009; van Niekerk et al., 2010; White & Frederiksen, 1998
Effective questioning	-	-	2	Chin & Teou, 2009; Wong & Day, 2009
Discourse / assessment conversations/ accountable talk	10	Lyon, Bunch, & Shaw, 2012; Mason, 2001; Nielsen, 2012; Osborne, Erduran, & Simon, 2004; Reyes, 2008; Ruiz-Primo & Furtak, 2006; Ruiz-Primo & Furtak, 2007; van Aalst & Mya Sioux Truong, 2011; Winters & Alexander, 2011; Zhang & Sun, 2011	4	Chen & Klahr, 1999; Hickey et al., 2012; Hickey & Zuiker, 2012; Valanides & Angeli, 2008
Quizzes	1	Cross et al., 2008	3	Hickey et al., 2012; Taasobshirazi & Hickey, 2005; Taasobshirazi et al., 2006
Performance assessment / experiments	13	Baxter et al., 1992; Hofstein et al., 2005; Kelly et al., 1998; Lyon et al., 2012; McElhaney & Linn, 2011; Pine et al., 2006; Ruiz-Primo et al., 2002; Ruiz-Primo et al., 2010; Schneider et al., 2002; Shavelson et al., 1991; Shavelson et al., 2008; Stecher et al., 2000	2	Chen & Klahr, 1999; Sampson et al., 2011
Interviews	24	Acar & Tarhan, 2007; Akerson & Donnelly, 2010; Berland & Reiser, 2009; Berland, 2011; Carruthers & Berg, 2010; Dawson & Venville, 2009; Gibson & Chase, 2002; Gijlers & Jong, 2005; Gotwals & Songer, 2009; Hamilton et al., 1997; Hmelo et al., 2000; Jang, 2010; Khishfe, 2008; Kim & Song, 2006; Lin & Mintzes, 2010; Mistler Jackson & Songer, 2000; Schnittka & Bell, 2011; Schwarz & White, 2005; Southerland et al., 2005; van Niekerk et al., 2010; Veal & Chandler, 2008; Vellom & Anderson, 1999; White & Frederiksen, 1998; Wilson et al., 2010	3	Ash, 2008; Goodnough & Long, 2006; Tytler et al., 2009

Observation / field notes	13	Abi-EI-Mona & Abd-EI-Khalick, 2006; Aguiar et al., 2010; Carruthers & Berg, 2010; Hamilton et al., 1997; Harskamp et al., 2008; Kubasko et al., 2008; Lavoie, 1999; Mistler Jackson & Songer, 2000; Ryu & Sandoval, 2012; Southerland et al., 2005; Valanides & Angeli, 2008; van Niekerk et al., 2010; Vellom & Anderson, 1999	3	Goodnough & Long, 2006; Harris et al., 2006; Tytler et al., 2009
Video tapes / audio tapes	25	Abi-EI-Mona & Abd-EI-Khalick, 2006; Aguiar et al., 2010; Berland & Reiser, 2009; Berland, 2011; Birchfield & Megowan-Romanowicz, 2009; Cavagnetto et al., 2010; Chen & Klahr, 1999; Chen & Looi, 2011; Chin & Osborne, 2010; Erduran, Simon, & Osborne, 2004; Harris et al., 2006; Kelly et al., 1998; Kim & Song, 2006; Kubasko et al., 2008; Kyza, 2009; McNeill, 2009; Mistler Jackson & Songer, 2000; Ryu & Sandoval, 2012; Sampson et al., 2011; Schnittka & Bell, 2011; Shemwell & Furtak, 2010; Southerland et al., 2005; Taasoobshirazi & Hickey, 2005; Valanides & Angeli, 2008; Vellom & Anderson, 1999	6	Ash, 2008; Chin & Teou, 2009; Furtak & Ruiz-Primo, 2008; Furtak et al., 2008; Tytler et al., 2009; White & Frederiksen, 1998
Questionnaires	8	Brandstädter et al., 2012; Butler & Lumpe, 2008; Kim & Song, 2006; McNeill, 2009; Mistler Jackson & Songer, 2000; Shavelson et al., 2008; Southerland et al., 2005; Winters & Alexander, 2011	-	-
Artefacts	2	Harris et al., 2006; Kyza, 2009	-	-



5.2.2 Technology

In total, empirical studies on IBE and assessment methods in technology education are rare. Obviously, in contrast to science and mathematics education, this research field is not particularly dominant. One reason is that technology is not a common subject in European schools (see D 2.3, National reports of partner countries reviewing research on formative and summative assessment in their countries) or in American schools.

Table 31: Frequency of assessment methods in the studies from the field of technology education

Assessment method	SA [M]	References	FA [M]	References
Multiple-choice	3	Burghardt et al., 2010; Doppelt, 2003; Klahr et al., 2007	-	-
Constructed-response / Open-ended	6	Burghardt et al., 2010; Doppelt, 2003; Fox-Turnbull, 2006; Klahr et al., 2007; Mioduser & Betzer, 2007; Merrill, Custer, Daugherty, Westrick, & Zeng, 2008	-	-
Portfolios	2	Doppelt, 2009; Williams, 2012	3	Barak & Doppelt, 2000; Doppelt, 2003; Hong et al., 2011
Discourse / assessment conversations / accountable talk	1	MacDonald & Gustafson, 2004	-	-
Performance assessment / experiments	2	Mioduser & Betzer, 2007; Williams, 2012	-	-
Interviews	1	Davis et al., 2002	2	Barak & Doppelt, 2000; Doppelt, 2003
Observation / field notes	2	Doppelt, 2003; Doppelt, 2009	1	Barak & Doppelt, 2000
Audio tapes	1	Gustafson et al., 2007	-	-
Questionnaires	1	Doppelt, 2003	-	-

With regard to summative assessment, the most important methods are, similar to science education, constructed-response or open-ended items and multiple-choice items (see Table 31). In most cases, they were used for the assessment of knowledge, achievement or understanding. Furthermore, they measured students' motivation or attitudes towards technology (Burghardt et al., 2010; Doppelt, 2003; Klahr et al., 2007).

When looking at formative assessment, the most important methods are portfolios and interviews (see Table 31). Obviously, the advantage of portfolios is their ability to reconstruct a process when solving a problem or designing a prototype (Barak & Doppelt, 2000; Doppelt, 2003; Hong et al., 2011).

Interviews should usually follow guidelines. Davis, Ginns and McRobbie (2002, p. 39) give examples of questions designed to probe the students' understandings of materials and stability:

- “Tell me as much as you can about this object, what it is, how it is made, and what it is made out of. (At the same time students were shown an artifact such as a model bridge constructed out of wood.)
- If you were building this bridge [type] to carry cars and/or pedestrians, what material(s) would you build it out of and why?
- Is this bridge stable? If not, explain how you would make it more stable.
- How do the changes you have suggested make the bridge more stable?”

One major field of research is problem- or project-based learning. In the first case, the starting point is the presentation of a technical problem (see Figure 8). Students have to find an answer and consider alternative solutions (Fox-Turnbull, 2006). In the second case, the starting points are the presentation of a target setting and of materials which can be used to reach this target (see Figure 9). One of the studies focused on the comparison between a hands-on and a virtual construction of a prototype (Klahr et al., 2007).

Help Me Peel

The photo shows a person who's wanting to peel a potato.
It is very hard because the person can only use one hand.
The other hand is in plaster.
This person needs to have a way for making it easier to peel a potato.

1. Draw a plan of your idea for something to make it easier to peel a potato.
2. On your plan, write how your idea makes it easier to peel a potato.

Figure 8: Help me peel task and photo (Fox-Turnbull, 2006, p. 59)

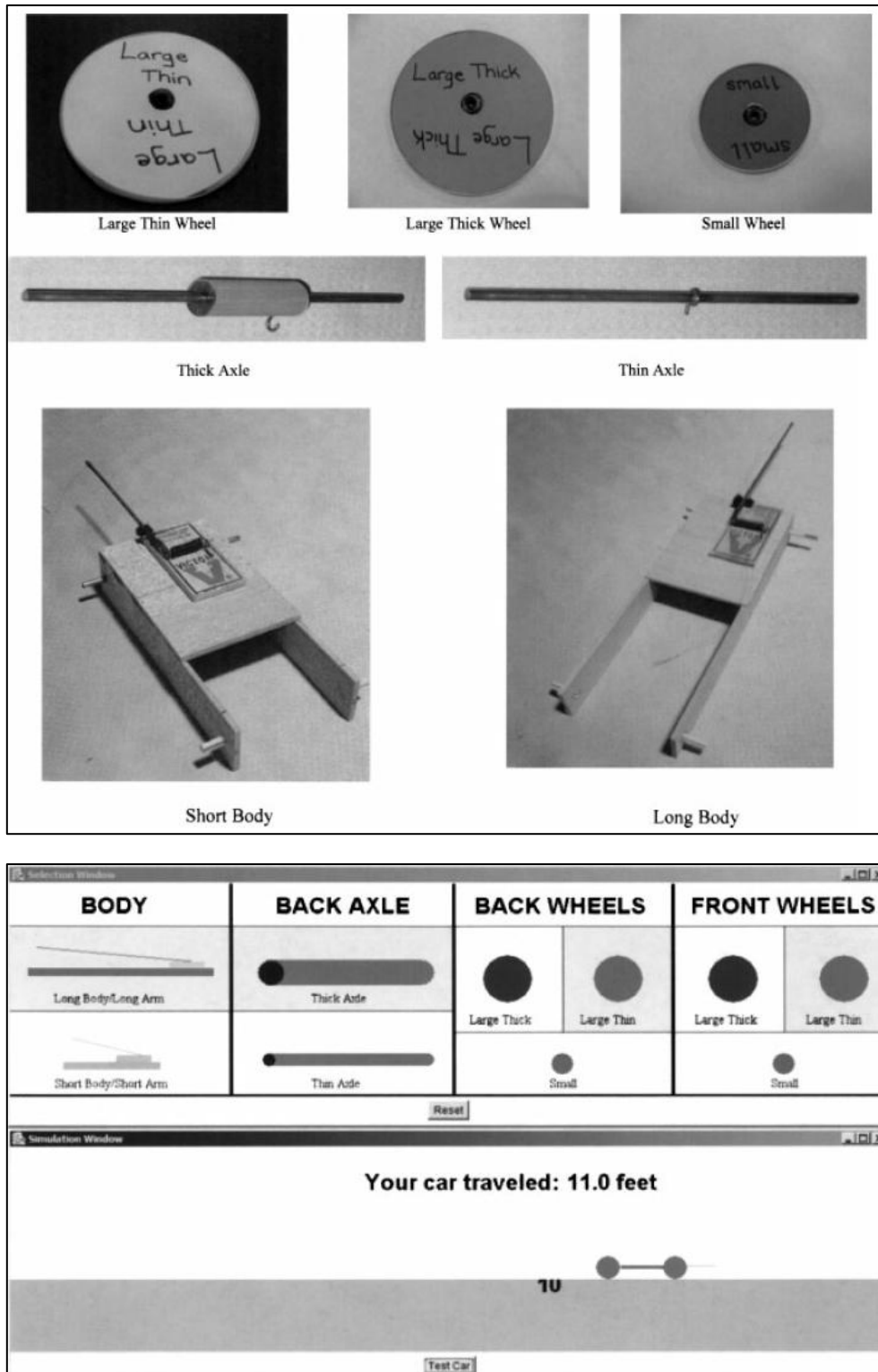


Figure 9: Hands-on and virtual mousetraps (Klahr et al., 2007, pp. 188–189)

The reported studies did not use the methods concept map, mind map, learn log, notebook, effective questioning, heuristics, quizzes, video tapes, written materials, or artefacts.

5.2.3 Mathematics

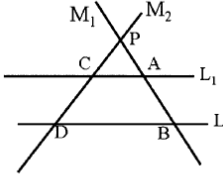
In mathematics, the emphases lay on constructed-response or open-ended items - especially for a summative assessment (see Table 32). The purpose of the items was often the evaluation of an intervention by a pre-post-design. The items ascertained students' reasoning or problem-solving skills and their mathematical knowledge.

Table 32: Frequency of assessment methods in the studies from the field of mathematics education

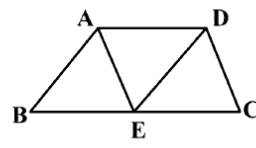
Assessment method	SA [N]	References	FA [N]	References
Multiple-choice	2	Bouck & Kulkarni, 2009; Reys et al., 2003	1	Cross, 2009
Constructed-response / open-ended	14	Boesen et al., 2010; Bouck & Kulkarni, 2009; Britt & Irwin, 2008; Chang et al., 2012; Heinze et al., 2008; Knuth, Alibali, McNeil, Weinberg, & Stephens, 2005; Kwon et al., 2006; Liedtke, 1999; Lin et al., 2004; Reiss et al., 2008; Reys et al., 2003; Rubel, 2007; Wood & Sellers, 1997; Zhang et al., 1999	3	Phelan et al., 2012; Ross, Hogaboam-Gray, & Rolheiser, 2002; Tzur, 2007
Portfolios	1	Koretz, 1998	-	-
Discourse / assessment conversations / accountable talk	3	Martin, McCrone, Bower, & Dindyal, 2005; Pijls, Dekker, & van Hout-Wolters, 2007; Woods et al., 2006	1	Tzur, 2007
Performance assessment / experiments	1	Linn, Burton, DeStefano, & Hanson, 1995	-	-
Interviews	1	Boaler, 1998	1	Ai, 2002
Observation / field notes	1	Boaler, 1998	2	Ai, 2002; Tzur, 2007
Video tapes / audio tapes	2	Chiu, 2008; Webb, Nemer, & Ing, 2006	2	Tzur, 2007; Woods et al., 2006
Questionnaires	3	Boaler, 1998; Chiu, 2008; Schukajlow et al., 2012	-	-
Artefacts	-	-	1	Tzur, 2007

The use of constructed-response or open-ended items is not surprising as, in mathematics education, students usually have to calculate and write down the calculation or prove and explain a given problem. Among the studies, Heinze et al. (2008) gave examples of test items which measure students' proof competence (see Figure 10). Knuth et al. (2005) also gave examples of test items (see Figure 11). Both studies illustrate the character of this assessment method. The example from Schukajlow et al. (2012) focused more on the assessment of problem-solving skills (see Figure 12).

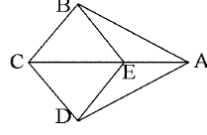
In contrast to science and technology education, multiple-choice items are less common in mathematics education. It is assumed that they would simplify the tests by providing different answer options. Therefore, they are not suitable for the assessment of problem-solving skills.



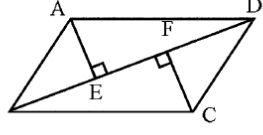
11. L_1 is parallel to L_2 . M_1 intersect L_1 and L_2 at A, B. M_2 intersect L_1 and L_2 at C, D. M_1 intersect M_2 at P, and $PA=PC$. Prove that $AB=CD$



12 Line BC. ABE is congruent to DEC. Prove that AB is parallel to DE.



13 A, E, C lies on a line. $\triangle ABC$ is congruent to $\triangle ADC$. Prove that $BE=DE$



14 In the parallelogram ABCD. AE is perpendicular to BD and CF is perpendicular to BD. Prove that $AE=CF$

Figure 10: The items of the pre-test (Heinze et al., 2008, p. 448)

Is the number that goes in the the same number in the following two equations? Explain your reasoning.

$2 \times \square + 15 = 31$ $2 \times \square + 15 - 9 = 31 - 9$

Figure 11: Using the concept of mathematical equivalence (Knuth et al., 2005, p. 70)

Football Pitch

Trainer Manfred would like to carry out a diagonal run with his team. To do so he would like to know how long the diagonal of the football pitch is. Can you help him?

Calculate the length the diagonal of the football pitch.

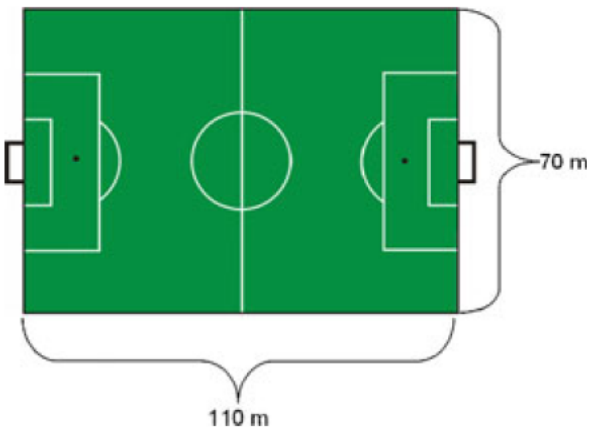


Figure 12: “Dressed up” world problem “football pitch” (Schukajlow et al., 2012, p. 225)

Another emphasis lay on the observation of lessons or learning situations by observations, field notes, video tapes and audio tapes. The application of these methods was not described in detail. As these methods were used in a more qualitative way, the focus of the respective publications was on the description of the observed learning or teaching processes (e. g. Boaler, 1998). Other studies focused on the analysis of discourse, assessment conversations or accountable talk in connection with collaborative learning (e. g. Pijls et al., 2007).

The methods concept map, mind map, learn log, notebook, effective questioning, heuristics, quizzes and written materials were not used within the context of the studies found. Admittedly/In fact/Indeed, these methods are more suitable for a formative assessment (s. Chapter 2). Obviously, there is a need for more research on formative assessment in connection with IBE in mathematics learning.

The GPAR reflection sheets are different from all other methods. They ask students to write responses to the questions presented in Figure 13 (Brookhart, Andolina, Zuza, & Furman, 2004). Students have to reflect on their learning process. Therefore, this method is useful in view of formative assessment.

<p>GOAL – What did you want to learn? [space provided] LOCATION – Right now I can do ___ facts in five minutes. PLAN – My goal is to get ___ /100 facts on my next test. I need to improve in [space provided] ACTION – When will you begin? Starting ___ I will use these study strategies to improve (study flashcards, play multiplication games, study with parents, etc.) [space provided] I will use these problem-solving strategies to improve (write a number sentence, use repeated addition, draw a picture, make a model, array) [space provided] RESULTS – Did you follow through with your plan? What happened? Did you see improvements? [space provided]</p>
<p>Plan for Learning Multiplication Facts My last test score was ___ /100 Reflection To learn my multiplication facts last week, I [space provided] How well did this plan work for me? terrible not well o.k. pretty well great Why did or didn't my plan work? Prediction I want to try to score ___ /100 next week. This week to help me improve I will use these strategies: [space provided] This face shows how I feel about multiplication: (draw a face below) [space provided]</p>

Figure 13: Goals, Plan, Action and Reflection sheet in original and revised version (Brookhart et al., 2004, pp. 216–217)

6. Perspectives

This report is intended to give an overview of the current state of the art in formative and summative assessment in IBE in STM. Instruments for the summative and formative assessment of IBE are described for each subject as far as they have been found by the different search strategies, as far as they exist and as far as they have been investigated. The results of this literature review are limited by the chosen keywords and search strategies. For example, IBE is not a common approach in mathematics education. This might be the reason why there are only few publications in mathematics education. Another reason might be that the common approach of problem-solving is not included as a keyword in the list of relevant keywords. This is a serious restriction which has to be made.

Nevertheless, the literature review reveals some subject-specific emphases, especially in science education. For this subject, half of the publications found report the use of multiple-choice items. Constructed-response and open-ended items are used by half of the empirical studies. However, in both cases, the only purpose of the methods is summative assessment. All other assessment instruments are only used in science education research quite rarely. Subject-specific instruments are mapping techniques like concept mapping.

In technology education, as well as in mathematics education, the emphases lay on constructed-response and open-ended items. In technology education, portfolios were also used. They play an important role in assessing constructing processes.

In view of the assessment type, the emphasis lies on summative assessment. Compared to summative assessment, formative assessment is an aspect that is only investigated in a few studies. All in all, there is not much variation observed with respect to the employed assessment instruments.

In a certain way, there is also not much variation observed in view of IBE. In order to make this result visible, a network for each subject was created with R (R Core Team, 2013) and the *igraph* package (Csardi & Nepusz, 2006). Figure 14, Figure 15 and Figure 16 show the relations between several aspects of IBE. The size of the circles thereby represents the number of publications investigating a certain aspect of IBE. The figures thus allow for the identification of the so-called 'hot spots' of inquiry for each subject. Obviously, the aspect 'constructing and critiquing arguments or explanations, argumentation, reasoning, and using evidence' is the aspect that is most often focused on or investigated in the field of IBE. In science education, it is followed by 'debating with peers and communication', 'collecting and interpreting data', 'planning investigations', 'diagnosing problems and identifying questions', 'evaluating results' and 'formulating hypotheses'. Thus, these are the core aspects of scientific inquiry whereas 'considering alternatives' is less significant.

In technology education, IBE covers fewer aspects. The considered ones are much more knotted than in science education because the net looks much more regular and has not a single dominating node. In mathematics education, 'searching for generaliza-

tions', 'creating mental representations' and 'evaluating results' are the most prominent aspects of IBE.

Furthermore, the results of the literature review and the three figures indicate that there are 'blind spots'. These are aspects of IBE or methods of formative and summative assessment that are more or less not assessed at all or they are assessment methods that are used very seldom.

However, because the specific focus of the ASSIST-ME project is on the relation between aspects of inquiry and assessment methods, further research within the project is necessary to investigate these 'blind spots'. The three figures give a first impression of the content of the prospective recommendation report. The forthcoming report D 2.7 will – on the basis of all previous reports of WP 2 – emphasize this issue by answering the following questions: Do aspects of inquiry exist that should be preferably assessed by a specific assessment method? Or, vice versa, are certain assessment methods particularly suited for assessing certain aspects of inquiry? Thus, D 2.7 will present the connections between aspects of IBE in STM and formative and summative assessment methods.

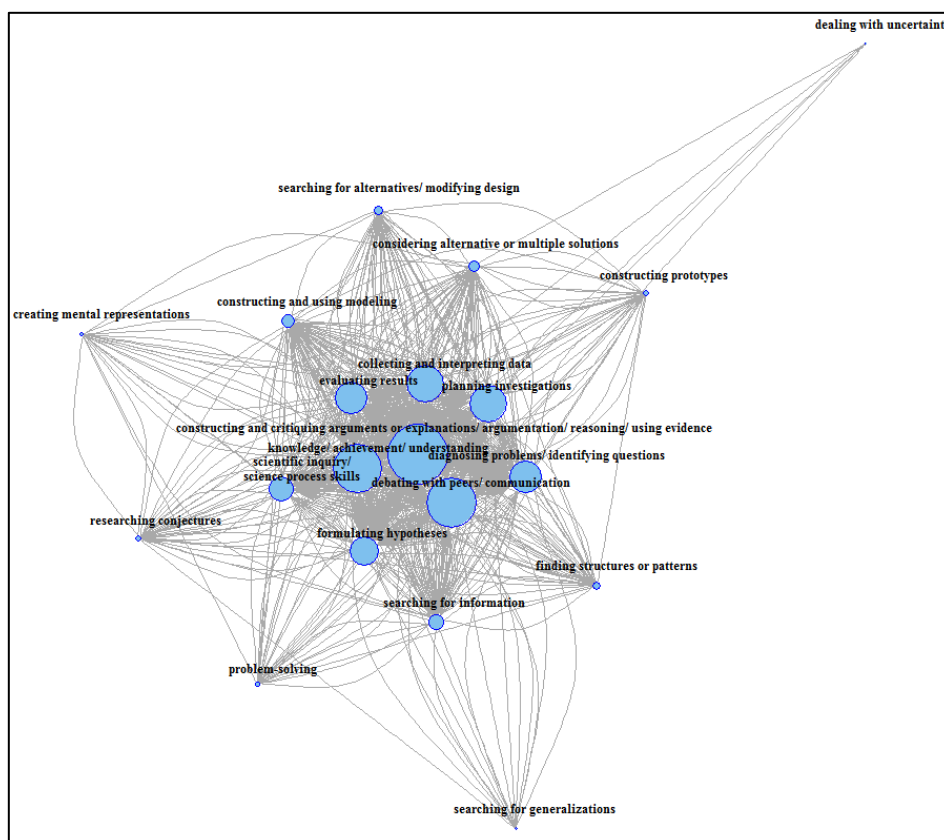


Figure 14: 'hot spots' of inquiry in science education

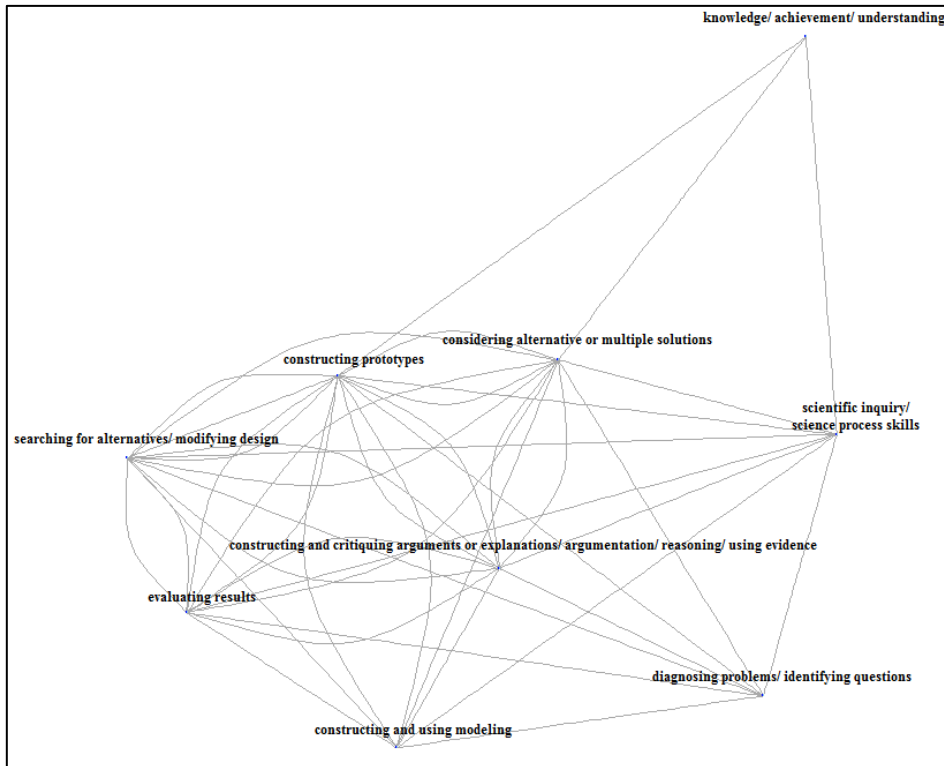


Figure 15: 'hot spots' of inquiry in technology education

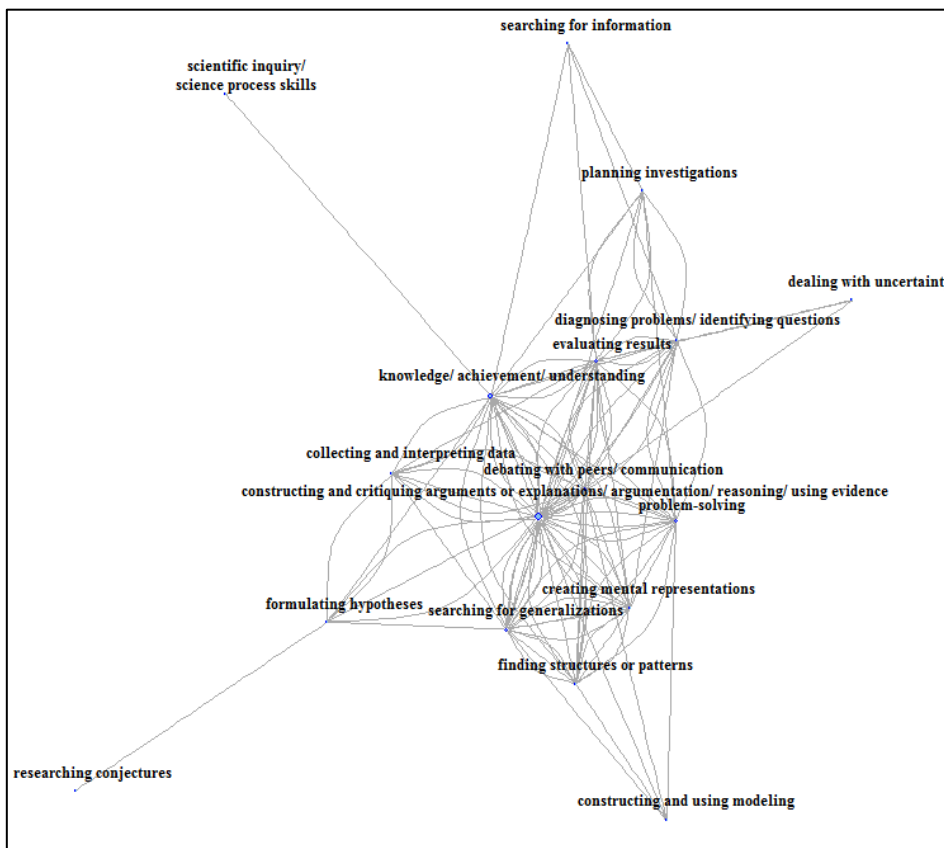


Figure 16: 'hot spots' of inquiry in mathematics education

7. Appendix

7.1 Frameworks of inquiry competences and/or assessment

- Brown, N. J. S., Furtak, E. M., Timms, M., Nagashima, S. O., & Wilson, M. (2010). The Evidence-Based Reasoning Framework: Assessing Scientific Reasoning. *Educational Assessment*, 15(3-4), 123–141.
- Brown, N. J. S., Nagashima, S. O., Fu, A., Timms, M., & Wilson, M. (2010). A Framework for Analysing Scientific Reasoning in Assessments. *Educational Assessment*, 15(3-4), 142–174.
- Champagne, A. B., Kouba, V. L., & Hurley, M. (2000). Assessing inquiry. In J. Minstrell & E. H. van Zee (Eds.), *Inquiring into Inquiry Learning and Teaching in Science* (pp. 447–470). Washington, DC: American Association for the Advancement of Science.
- Garden, R. A. (1999). Development of TIMSS performance assessment tasks. *Studies in Educational Evaluation*, 25(3), 217–241.
- Gitomer, D. H., & Duschl, R. A. (1995). Moving toward a portfolio culture in science education. In S. M. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 299–326). Mahwah: Erlbaum.
- Heritage, M., & Niemi, D. (2006). Toward a Framework for Using Student Mathematical Representations as Formative Assessments. *Educational Assessment*, 11(3-4), 265–282.
- Hickey, D. T., Taasobshirazi, G., & Cross, D. (2012). Assessment as learning: Enhancing discourse, understanding, and achievement in innovative science curricula. *Journal of Research in Science Teaching*, 49(10), 1240–1270.
- Johnson, R. S., Mims-Cox, J. S., & Doyle-Nichols, A. (op. 2006). *Developing portfolios in education: A guide to reflection, inquiry, and assessment*. Thousand Oaks: Sage Publications Ltd.
- Lane, S. (1993). The Conceptual Framework for the Development of a Mathematics Performance Assessment Instrument. *Educational Measurement: Issues and Practice*, 12(2), 16–23.
- Lawson, A. E. (2010). Basic inferences of scientific reasoning, argumentation, and discovery. *Science Education*, 94(2), 336–364.
- Lederman, N., Wade, P., & Bell, R. L. (1998). Assessing understanding of the nature of science: A historical perspective. In W. F. McComas (Ed.), *The nature of science in science education* (pp. 331–350). Dordrecht: Kluwer Academic Publishers.
- Lewis, T. (2005). Creativity – A Framework for the Design/Problem Solving Discourse in Technology Education. *Journal of Technology Education*, 17(1), 35–52.
- McComas, W. F. (Ed.). (1998). *The nature of science in science education*. Dordrecht: Kluwer Academic Publishers.
- Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative Discourse Idealized and Realized: Accountable Talk in the Classroom and in Civic Life. *Studies in Philosophy and Education*, 27(4), 283–297.
- Minstrell, J. (2000). Student thinking and related assessment: Creating a facet-based learning environment. In N. Raju, J. Pellegrino, M. Bertenthal, K. Mitchell, & L. Jones

- (Eds.), *Grading the nation's report card. Research from the evaluation of NAEP* (pp. 44–73). Washington, D.C: National Academy Press.
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, 25(4), 6–20.
- Nichols, P. D., Meyers, J. L., & Burling, K. S. (2009). A Framework for Evaluating and Planning Assessments Intended to Improve Student Achievement. *Educational Measurement: Issues and Practice*, 28(3), 14–23.
- Osborne, J., & Patterson, A. (2012). Authors' response to “For whom is argument and explanation a necessary distinction? A response to Osborne and Patterson” by Berland and McNeill. *Science Education*, 96(5), 814–817.
- Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education*, 95(4), 627–638.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. E. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: National Academies Press.
- Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, D.C: National Academy Press.
- Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and Testing. *Science*, 323, 75–79.
- Quellmalz, E. S., Timms, M. J., & Buckley, B. (2010). The promise of simulation-based science assessment: the Calipers project. *International Journal of Learning Technology*, 5(3), 243–263.
- Ruiz-Primo, M. A. (2011). Informal formative assessment: The role of instructional dialogues in assessing students' learning. *Studies in Educational Evaluation*, 37(1), 15–24.
- Ruiz-Primo, M. A. & Shavelson, R. J. (1997). *Concept-Map based assessment: On possible sources of sampling viability*. Los Angeles. Retrieved from http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_ERICExtSearch_SearchValue_0=ED422403&ERICExtSearch_SearchType_0=no&accno=ED422403
- Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education*, 92(3), 499–525.
- Ryve, A. (2011). Discourse research in mathematics education: a critical evaluation of 108 journal articles. *Journal for Research in Mathematics Education*, 42(2), 167–199.
- Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education*, 92(3), 447–472.
- Scardamalia, M., Bransford, J. D., Kozma, B., & Quellmalz, E. S. (2012). New Assessments and Environments for Knowledge Building. In P. E. Griffin, B. McGaw, & E.

Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 231–300). Dordrecht, New York: Springer.

Wilson, M., & Sloane, K. (2000). From Principles to Practice: An Embedded Assessment System. *Applied Measurement in Education*, 13(2), 181–208.

7.2 Computer-supported inquiry learning environments and computer-based assessment tools

Name	Description	Reference(s)
Web of Inquiry (WOI)	Selection of web inquiry projects (WIPs); no special focus on assessment	Herrenkohl, Tasker, & White, 2011; Molebash, no date
Web-based Inquiry Science Environment (WISE)	e.g. provides electronic student notebooks; learners are asked at several points to think about questions that challenge them to reflect more deeply, to see things from another perspective, or to apply knowledge built in the preceding section; the student answers about the project are saved in the notebook and can be reviewed as a whole at any time by the student or by the teacher for assessment purposes; includes different assessment tools (pre/post, embedded) to assess interpreting and constructing graphs, reasoning using data/evidence, explaining, and experimentation strategy (using log files); empirical study showed large, significant gains for WISE students	Bell, Urhahne, Schanze, & Ploetzner, 2010; Linn, Clark, & Slotta, 2003; McElhaney & Linn, 2008; University of Berkeley, 2013
Modeling Across the Curriculum (MAC)	e.g. BioLogica, a hypermodel, interactive environment for learning genetics; traces of students' actions and responses to computer-based tasks are electronically collected (log files) and systematically analysed	Buckley et al., 2004
Collaborative Laboratories across Europe (Co-Lab)	e.g. self-evaluation by process displays/prompts; reflective notebooks; long instructional Co-Lab units allow teachers to evaluate the inquiry process skills of individual students more effectively	van Joolingen, Jong, Lazonder, Savelsbergh, & Manlove, 2005; Urhahne, Schanze, Bell, Mansfield, & Holmes, 2010
	Overview of computer-supported learning environments	Bell et al., 2010
ThinkerTools Curriculum	inquiry curriculum centres around a metacognitive model of research, called the Inquiry Cycle, and a metacognitive process, called Reflective Assessment, in which students reflect on their own and each other's inquiry	White & Frederiksen, 1998

DIAGNOSER	analyses facets of students' thinking; description of facets can be used as scoring guide	Pellegrino, Baxter, & Glaser, 1999; Pellegrino, Chudowsky, & Glaser, 2001
SimScientist	simulation-based science assessments designed to serve formative purposes during a unit and to provide summative evidence of end-of-unit proficiencies; evidence-centred assessment design and model-based learning shaped assessments; IRT analyses demonstrated the high psychometric quality (reliability and validity) of the assessments and their discrimination between content knowledge and inquiry practices. Students performed better in the interactive, simulation-based assessments than in static, conventional items in a post-test. Importantly, gaps between the performance of the general population and English language learners and the students with disabilities were considerably smaller in the simulation-based assessments than in the post-tests	Quellmalz & Pellegrino, 2009; Quellmalz, Timms, Silbergliitt, & Buckley, 2012
Calipers project: Using Simulations to Assess Complex Science Learning	developed assessment designs and prototypes that can take advantage of technology to bring high-quality assessments of complex performances into science tests with either accountability or formative goals	Quellmalz et al., 2007; Quellmalz, Timms, & Buckley, 2010
	Role of games and simulations in science assessments; description of several interactive environments, e.g. SimScientist, Calipers II, IMMEX (Interactive Multimedia Exercises), River City, Crystal Island	Honey & Hilton, 2011

Viten	e.g. provides electronic student notebooks; learners are asked at several points to think about questions that challenge them to reflect more deeply, to see things from another perspective, or to apply knowledge built in the preceding section. The student answers about the project are saved in the notebook and can be reviewed as a whole at any time by the student or by the teacher for assessment purposes; allows teachers to give electronic feedback to students via an assessment tool judged helpful by teachers and students; students are asked to show communication/argumentation skills by a role-play debate in a TV discussion programme; communication data is logged thus offering teachers the possibility to look it up later for coaching or assessment purposes	Bell, Urhahne, Schanze, & Ploetzner, 2010; Jorde, Strømme, Sorborg, Erlie, & Mork, 2003
Multi-User Virtual Environment (MUVE) River City	In this environment, middle school students collaboratively solve problems about disease in a virtual town called River City; results indicate that students were able to conduct inquiry in virtual worlds and were motivated by that process; however, results from assessments vary depending on the assessment strategy employed; also assessment of student engagement and influence of student self-efficacy on inquiry	e.g. Ketelhut, Nelson, Clarke, & Dede, 2010; Ketelhut & Nelson, 2010; Ketelhut, 2007
ASSISTments	ASSISTments is a free online platform that allows teachers to write and select questions, students to get immediate and useful tutoring, and teachers to receive instant reports to help inform their classroom instruction	Worcester Polytechnic Institute, 2013
	validity of computer-automated scoring	Clauser, Kane, & Swanson, 2002

	intelligent argumentation assessment system for computer-supported cooperative learning; is effective in classifying and improving students' argumentation level and assisting the students in learning the core concepts at primary school	Huang et al., 2011
Berkeley Evaluation and Assessment research (BEAR) – assessment system		Wilson & Scalise, 2003; Wilson & Sloane, 2000
Formative Assessment in Science Teaching (FAST) homepage	Hosts output from the FAST project, e.g. case studies, resources, and investigative tools (e.g. feedback coding scheme, assessment experience questionnaire)	Brown, 2008; The Open University & Sheffield Hallam University, 2008
Principled Assessment Designs for Inquiry (PADI) homepage	Uses evidence-centred design framework; aims to provide a practical, theory-based approach to developing quality assessments of science inquiry by combining developments in cognitive psychology and research on science inquiry with advances in measurement theory and technology	SRI International, 2007



7.3 Assessment instruments

Name	Description	Reference(s)
Measuring up. Prototypes for mathematics assessment.	Collection of assessment tasks that bring standards to life and thus offer children opportunities to demonstrate the full range of their mathematical power, including such important facets as communication, problem solving, inventiveness, persistence, and curiosity; focuses on grade 4	Mathematical Sciences Education Board & National Research Council, 1993
	Instruments to assess technology literacy	Garmire & Pearson, 2006
Discovery Inquiry Test in Science (DIT)	consists of released NAEP items that measure students' abilities to analyse and interpret data, to extrapolate from one situation to another, and to utilize conceptual understanding; was, e.g., used in study to assess impact of effective teaching	Johnson, Kahle, & Fargo, 2007; Program in Education, no date
Competence Scale for Learning Science	Questionnaire assessing competence scale for learning science regarding competencies in scientific inquiry and communication; 29 self-report, Likert-type items	Chang et al., 2011
Number Knowledge Test	test to assess mathematical understanding of whole numbers	Griffin, 2005
Indicators and Instruments in the Context of Inquiry-based Science Education	Instruments to assess IBST identified within the EU project S-TEAM	Heinz, 2012
Practical Tests Assessment Inventory	Instrument to assess inquiry practical examinations in biology	Tamir, Nussinovitz, & Friedler, 1982
McGill Inventory of Student Inquiry Outcomes (MISIO)	23-item, criterion-referenced; student outcomes include knowledge and skills, intrinsic motivation, and development of expertise	Saunders-Stewart, Gyles, & Shore, 2012

Assessment of inquiry or science process skills		
Test of the Integrated Science Process Skills	Develop a reliable and valid instrument to measure integrated science process skills	Dillashaw & Okey, 1980
Test of Inquiry Process Skills (TIPS II)	Provides a reliable instrument for measuring the process skill achievement of middle and high school students	Burns, Okey, & Wise, 1985
Test of Science Process Skills		Molitor & George, 1976
Test of science processes		Tannenbaum, 1971
	Test items for four integrated science processes	McLeod, Berkheimer, Fyffe, & Robison, 1975
	questionnaire with 15 constructed-response (CR) type items and one hands-on task to assess science process skills; grade 9	Temiz, Taşar, & Tan, 2006
Test of enquiry skills	Development and validation of a content free test of enquiry skills	Fraser, 1980
Processes of biological investigations test	Easily administered, reliable p&p test for high school biology students that measures the science process skills developing hypotheses, making predictions, identifying assumptions, analysing data, and formulating conclusions	Germann, 1989



Assessment of reasoning		
Evidence-Based Reasoning in Science Classroom Discourse	Instrument is intended to provide a means for measuring the quality of evidence-based reasoning in whole-class discussions, capturing teachers' and students' co-constructed reasoning about scientific phenomena; coding system for assessing argumentation in science classroom discourse is developed	Furtak, Hardy, Beinbrech, Shavelson, & Shemwell, 2010
Raven's Progressive matrices	measures general mental ability and offers information about someone's capacity for analysing and solving problems, abstract reasoning, and the ability to learn; an earlier version (Raven's progressive test of non-verbal reasoning) used to assess scientific reasoning	Mercer, Dawes, Wegerif, & Sams, 2004
Assessment of attitudes and affect		
Views of Nature of Science (VNOS)	Questionnaire for NOS	Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002
Views of Scientific Inquiry (VOSI)		Schwartz, Lederman, & Lederman, 2008
Views of Scientific Inquiry – primary school (VOSI-P)		Program in Education, no date
Test of Science Related Attitudes (TOSRA)		Fraser, 1981; Fraser & Butts, 1982; Program in Education, no date
“Learning how to learn”-project	A Project of the ESRC Teaching and Learning Research Program; presents e.g. self-evaluation questionnaires	Learning how to Learn Project, 2002
	Questionnaire for assessing students' motivation	Nolen, 2003; Osborne et al., 2013

	Questionnaire for assessing students' attitudes towards science in grades 1-5	Pell & Jarvis, 2001; Osborne et al., 2013
	Questionnaire for assessing four dimensions of epistemic beliefs (source, certainty, development, justification) in primary school	Conley, Pintrich, Vekiri, & Harrison, 2004; Osborne et al., 2013
	MC test to assess development of epistemological understanding (absolutist, multiplist, evaluativist)	Kuhn, Cheney, & Weinstock, 2000; Osborne et al., 2013
	Overview of existing instruments to assess affective measures in mathematics	Chamberlin, 2010
Attitudes towards mathematics inventory (short version)		Lim & Chapman, 2013
Assessment of assessment literacy		
Teacher assessment literacy questionnaire	psychometric properties of the teacher assessment literacy questionnaire	Alkharusi, 2011
Classroom assessment literacy inventory	35 items related to the seven Standards for Teacher Competence in the Educational Assessment of Students; Some of the items are intended to measure general concepts related to testing and assessment; other items are related to knowledge of standardized testing and the remaining items are related to classroom assessment	Mertler, no date

References

- Abi-El-Mona, I., & Abd-El-Khalick, F. (2006). Argumentative Discourse in a High School Chemistry Classroom. *School Science and Mathematics*, 106(8), 349–361.*
- Acar, B., & Tarhan, L. (2007). Effect of Cooperative Learning Strategies on Students' Understanding of Concepts in Electrochemistry. *International Journal of Science and Mathematics Education*, 5(2), 349–373.*
- Aguiar, O. G., Mortimer, E. F., & Scott, P. (2010). Learning From and Responding to Students' Questions: The Authoritative and Dialogic Tension. *Journal of Research in Science Teaching*, 47(2), 174–193.*
- Ai, X. (2002). *District Mathematics Plan Evaluation: 2001-2002 Evaluation Report*. Retrieved from <http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED472491>*
- Akerson, V., & Donnelly, L. A. (2010). Teaching Nature of Science to K-2 Students: What Understandings Can They Attain? *International Journal of Science Education*, 32(1), 97–124.*
- Alexopoulou, E., & Driver, R. (1996). Small-group discussion in physics: Peer interaction modes in pairs and fours. *Journal of Research in Science Teaching*, 33(10), 1099–1114.
- Alkharusi, H. (2011). Psychometric properties of the teacher assessment literacy questionnaire for preservice teachers in Oman. *Procedia – Social and Behavioral Sciences*, 29, 1614–1624.
- American Association for the Advancement of Science (1998). *Blueprints for Reform - Project 2061: Chapter 8: Assessment*. Retrieved from <http://www.project2061.org/publications/bfr/online/blpintro.htm>
- American Association for the Advancement of Science (2009). *Benchmarks for Science Literacy*. Retrieved from <http://www.project2061.org/publications/bsl/online/index.php>
- American Federation of Teachers, National Council on Measurement in Education, & National Education Association (1990). *Standards for teacher competence in educational assessment of students*. Washington, DC: National Council on Measurement in Education.
- Anderson, C. W. (2003). *Teaching science for motivation and understanding*. Unpublished manuscript. Retrieved from <https://www.msu.edu/~tuckey1/presentations/VIPP/TSMU.pdf>
- Anderson, K. J. (2012). Science education and test-based accountability: Reviewing their relationship and exploring implications for future policy. *Science Education*, 96(1), 104–129.
- Anderson, R. D. (2002). Reforming Science Teaching: What Research Says About Inquiry. *Journal of Science Teacher Education*, 13(1), 1–12.
- Artigue, M., & Baptist, P. (2012). *Inquiry in Mathematics Education* (Resources for Implementing Inquiry in Science and in Mathematics at School). Retrieved from <http://www.fibonacci-project.eu/>

- Artigue, M., Dillon, J., Harlen, W., & Léna, P. (2012). *Learning through inquiry* (Resources for Implementing Inquiry in Science and in Mathematics at School). Retrieved from <http://www.fibonacci-project.eu/resources>
- Aschbacher, P., & Alonzo, A. (2006). Examining the Utility of Elementary Science Notebooks for Formative Assessment Purposes. *Educational Assessment, 11*(3&4), 179–203.*
- Ash, D. (2008). Thematic continuities: Talking and thinking about adaptation in a socially complex classroom. *Journal of Research in Science Teaching, 45*(1), 1–30.*
- Ayala, C. C., Shavelson, R. J., Ruiz-Primo, M. A., Brandon, P. R., Yin, Y., Furtak, E. M., Young, D. B., & Tomita, M. K. (2008). From Formal Embedded Assessments to Reflective Lessons: The Development of Formative Assessment Studies. *Applied Measurement in Education, 21*(4), 315–334.
- Baker, D. R., Lewis, E. B., Purzer, S., Watts, N. B., Perkins, G., Uysal, S., Wong, S., Beard, R., & Lang, M. (2009). The Communication in Science Inquiry Project (CISIP): A Project to Enhance Scientific Literacy through the Creation of Science Classroom Discourse Communities. *International Journal of Environmental and Science Education, 4*(3), 259–274.*
- Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research, 61*(2), 213–238.
- Barak, M., & Doppelt, Y. (2000). Using portfolios to enhance creative thinking. *Journal of Technology Studies, 26*(2), 16–24.*
- Barron, B. & Darling-Hammond, L. (2008). Teaching for meaningful learning: A review of research on inquiry-based and cooperative learning. In L. Darling-Hammond, B. Barron, P. D. Pearson, A. H. Schoenfeld, E. K. Stage, T. D. Zimmermann, G. N. Cervetti, & J. Tilson (Eds.), *Powerful Learning. What we know about teaching for understanding*. San Francisco: Jossey-Bass. Retrieved from <http://www.edutopia.org/pdfs/edutopia-teaching-for-meaningful-learning.pdf>
- Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of Procedure-Based Scoring for Hands-On Science Assessment. *Journal of Educational Measurement, 29*(1), 1–17.*
- Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education, 85*(5), 536–553.
- Bell, P., & Linn, M. C. (2000). Scientific arguments as learning artifacts: Designing for learning from the web with KIE. *International Journal of Science Education, 22*(8), 797–817.
- Bell, T., Urhahne, D., Schanze, S., & Ploetzner, R. (2010). Collaborative Inquiry Learning: Models, tools, and challenges. *International Journal of Science Education, 32*(3), 349–377.
- Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25.

- Berland, L. K. (2011). Explaining Variation in How Classroom Communities Adapt the Practice of Scientific Argumentation. *Journal of the Learning Sciences*, 20(4), 625–664.*
- Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. [References]. *Science Education*, 93(1), 26–55.*
- Bernholt, S., Neumann, K. & Nentwig, P. (2012). *Making it tangible – Learning outcomes in science education*. Münster: Waxmann.
- Bielaczyc, K., & Blake, P. (2006). *Shifting epistemologies: examining student understanding of new models of knowledge and learning*. Retrieved from http://portal.acm.org/ft_gateway.cfm?id=1150042&type=pdf&coll=&dl=ACM&CFID=52035040&CFTOKEN=66842494
- Binkley, M., Erstad, O., Herman, J. L., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. E. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Dordrecht, New York: Springer.
- Birchfield, D., & Megowan-Romanowicz, C. (2009). Earth Science Learning in SMAL-Lab: A Design Experiment for Mixed Reality. *International Journal of Computer-supported Collaborative Learning*, 4(4), 403–421.*
- Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., Wiesemes, R. (Ed.), & Nickmans, G. (Ed.) (2006). A learning integrated assessment system. *Educational Research Review*, 1, 61–67.
- Black, P., Harrison, C., & Hodgen, J. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice*, 17(2), 215–232.
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the Black Box: Assessment for Learning in the Classroom. *Phi Delta Kappan*, 86(1), 8–21.
- Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.
- Blanchard, M. R., Southerland, S. A., Osborne, J. W., Sampson, V. D., Annetta, L. A., & Granger, E. M. (2010). Is inquiry possible in light of accountability? A quantitative comparison of the relative effectiveness of guided inquiry and verification laboratory instruction. *Science Education*, 94(4), 577–616.*
- Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.), *National Society for the Study of Education Yearbook: 68 (2). Educational evaluation: New roles, new means* (pp. 26–50). Chicago: University of Chicago Press.
- Boaler, J. (1998). Open and closed mathematics: student experiences and understandings. *Journal for Research in Mathematics Education*, 29(1), 41–62.*
- Boesen, J., Lithner, J., & Palm, T. (2010). The relation between types of assessment tasks and the mathematical reasoning students use. *Educational Studies in Mathematics*, 75(1), 89–105.*

- Bouck, E. C., & Kulkarni, G. (2009). Middle-School Mathematics Curricula and Students with Learning Disabilities: Is One Curriculum Better? *Learning Disability Quarterly*, 32(4), 228–244.*
- Brandstädter, K., Harms, U., & Großschedl, J. (2012). Assessing System Thinking Through Different Concept-Mapping Practices. *International Journal of Science Education*, 34(14), 2147–2170.*
- Britt, M. S., & Irwin, K. C. (2008). Algebraic thinking with and without algebraic representation: a three-year longitudinal study. *ZDM*, 40(1), 39–53.*
- Brookhart, S. M. (2011). Educational Assessment Knowledge and Skills for Teachers. *Educational Measurement: Issues and Practice*, 30(1), 3–12.
- Brookhart, S. M., Andolina, M., Zuza, M., & Furman, R. (2004). Minute math: An action research study of student self-assessment. *Educational Studies in Mathematics*, 57(2), 213–227.*
- Brousseau, G., & Balacheff, N. (1997). *Theory of didactical situations in mathematics: Didactique des mathématiques, 1970-1990*. Dordrecht: Kluwer Academic Publishers.
- Brown, E. (2008). *Removing the grade from a formative assessment*. Retrieved from <http://www.open.ac.uk/fast/pdfs/Brown%20-AEQ.pdf>
- Brown, N. J. S., Nagashima, S. O., Fu, A., Timms, M., & Wilson, M. (2010). A Framework for Analysing Scientific Reasoning in Assessments. *Educational Assessment*, 15(3-4), 142–174.*
- Buckley, B. C., Gobert, J. D., Kindfield, A. C. H., Horwitz, P., Tinker, R. F., Gerlits, B., Wilensky, U., Dede, C., & Willett, J. (2004). Model-based teaching and learning with BioLogica: What do they learn? How do they learn? How do we know? *Journal of Science Education and Technology*, 13(1), 23–41.
- Burghardt, M. D., Hecht, D., Russo, M., Lauckhardt, J., & Hacker, M. (2010). A Study of Mathematics Infusion in Middle School Technology Education Classes. *Journal of Technology Education*, 22(1), 58–74.*
- Burns, J. C., Okey, J. R., & Wise, K. C. (1985). Development of an integrated process skill test: TIPS II. *Journal of Research in Science Teaching*, 22(2), 169–177.*
- Butler, K. A., & Lumpe, A. (2008). Student Use of Scaffolding Software: Relationships with Motivation and Conceptual Understanding. *Journal of Science Education and Technology*, 17(5), 427–436.*
- Carruthers, R., & Berg, K. de (2010). The Use of Magnets for Introducing Primary School Students to Some Properties of Forces through Small-Group Pedagogy. *Teaching Science*, 56(2), 13–17.*
- Cavagnetto, A., Hand, B. M., & Norton-Meier, L. (2010). The Nature of Elementary Student Science Discourse in the Context of the Science Writing Heuristic Approach. *International Journal of Science Education*, 32(4), 427–449.*
- Chamberlin, S. A. (2010). A review of Instruments Created to Assess Affect in Mathematics. *Journal of Mathematics Education*, 3(1), 167–182.

- Chang, H.-P., Chen, C.-C., Guo, G.-J., Cheng, Y.-J., Lin, C.-Y., & Jen, T.-H. (2011). The development of a competence scale for learning science: Inquiry and communication. *International Journal of Science and Mathematics Education*, 9(5), 1213–1233.*
- Chang, K.-E., Wu, L.-J., Weng, S.-E., & Sung, Y.-T. (2012). Embedding game-based problem-solving phase into problem-posing system for mathematics learning. *Computers & Education*, 58(2), 775–786.*
- Chen, W., & Looi, C.-K. (2011). Active Classroom Participation in a Group Scribbles Primary Science Classroom. *British Journal of Educational Technology*, 42(4), 676–686.*
- Chen, Z., & Klahr, D. (1999). All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. *Child Development*, 70(5), 1098–1120.*
- Chin, C., & Osborne, J. (2010). Students' Questions and Discursive Interaction: Their Impact on Argumentation during Collaborative Group Discussions in Science. *Journal of Research in Science Teaching*, 47(7), 883–908.*
- Chin, C., & Teou, L.-Y. (2009). Using Concept Cartoons in Formative Assessment: Scaffolding Students' Argumentation. *International Journal of Science Education*, 31(10), 1307–1332.*
- Chiu, M. M. (2008). Effects of argumentation on group micro-creativity: Statistical discourse analyses of algebra students' collaborative problem solving. *Contemporary Educational Psychology*, 33(3), 382–402.*
- Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessments that support learning: what will it take? *Theory into Practice*, 42(1), 75–83.
- Cizek, G. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20, 19–28.
- Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity Issues for Performance-Based Tests Scored With Computer-Automated Scoring Systems. *Applied Measurement in Education*, 15(4), 413–432.
- Cobb, P., Wood, T., Yackel, E., Nicholls, J., Wheatley, G., Trigatti, B., & Perlwitz, M. (1991). Assessment of a Problem-Centered Second-Grade Mathematics Project. *Journal for Research in Mathematics Education*, 22(1), 3–29.
- Cobb, P., Wood, T., Yackel, E., & McNeal, B. (1992). Characteristics of Classroom Mathematics Traditions: An Interactional Analysis. *American Educational Research Journal*, 29(3), 573–604.
- Cobern, W. W., Schuster, D., Adams, B., Applegate, B., Skjold, B., Undreiu, A., Loving, C. C., Gobert, J. D. (2010). Experimental comparison of inquiry and direct instruction in science. *Research in Science & Technological Education*, 28(1).81–96.*
- Coffey, J. E., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching*, 48(10), 1109–1136.
- Collis, K. F., Romberg, T. A., Jurdak, M. E. (1986). A technique for assessing mathematical problem-solving ability. *Journal for Research in Mathematics Education*, 17(3), 206–221.

- Conley, A. M., Pintrich, P. R., Vekiri, I., & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. *Contemporary Educational Psychology*, 29(2), 186–204.
- Cross, D., Taasoobshirazi, G., Hendricks, S., & Hickey, D. T. (2008). Argumentation: A Strategy for Improving Achievement and Revealing Scientific Identities. *International Journal of Science Education*, 30(6), 837–861.*
- Cross, D. I. (2009). Creating Optimal Mathematics Learning Environments: Combining Argumentation and Writing to Enhance Achievement. *International Journal of Science and Mathematics Education*, 7(5), 905–930.*
- Csardi, G. & Nepusz T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695. Retrieved from <http://igraph.sf.net>
- Davis, R. S., Ginns, I. S., & McRobbie, C. J. (2002). Elementary School Students' Understandings of Technology Concepts. *Journal of Technology Education*, 14(1), 35–50.*
- Dawson, V., & Venville, G. J. (2009). High-School Students' Informal Reasoning and Argumentation about Biotechnology: An Indicator of Scientific Literacy? *International Journal of Science Education*, 31(11), 1421–1445.*
- Delandshere, G. (2002). Assessment as Inquiry. *Teachers College Record*, 104(7), 1461–1484.
- Dillashaw, F. G., & Okey, J. R. (1980). Test of the integrated science process skills for secondary science students. *Science Education*, 64(5), 601–608.
- Ding, N., & Harskamp, E. G. (2011). Collaboration and Peer Tutoring in Chemistry Laboratory Education. *International Journal of Science Education*, 33(6), 839–863.*
- Dolin, J. (2012). *Assess Inquiry in Science, Technology and Mathematics Education: ASSIST-ME proposal*.
- Doppelt, Y. (2003). Implementation and assessment of project-based learning in a flexible environment. *International Journal of Technology and Design Education*, 13(3), 255–272.*
- Doppelt, Y. (2005). Assessment of Project-Based Learning in a MECHATRONICS Context. *Journal of Technology Education*, 16(2), 7–24.
- Doppelt, Y. (2009). Assessing creative thinking in design-based learning. *International Journal of Technology and Design Education*, 19(1), 55–65.*
- Dori, Y. J. (2003). From nationwide standardized testing to school-based alternative embedded assessment in Israel: Students' performance in the matriculation 2000 project. *Journal of Research in Science Teaching*, 40(1), 34–52.*
- Dori, Y. J., & Herscovitz, O. (1999). Question-posing capability as an alternative evaluation method: Analysis of an environmental case study. *Journal of Research in Science Teaching*, 36(4), 411–430.*
- Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, 84(3), 287–312.
- Dunn, K. E., & Mulvenon, S. W. (2009). A Critical Review of Research on Formative Assessment: The Limited Scientific Evidence of the Impact of Formative Assessment in Education. *Practical Assessment, Research and Evaluation*, 14(7), 1–11.

- Duschl, R. (1990). *Restructuring Science Education: The Importance of Theories and Their Development*. New York: Teacher's College Press.
- Duschl, R. (2000). Making the nature of science explicit. In R. Millar, Leech, J., & J. Osborne (Eds.), *Improving Science Education: The contribution of research* (pp. 187–206). Philadelphia: Open University Press.
- Ebenezer, J., Kaya, O. N., & Ebenezer, D. L. (2011). Engaging students in environmental research projects: Perceptions of fluency with innovative technologies and levels of scientific inquiry abilities. *Journal of Research in Science Teaching*, 48(1), 94–116.*
- Elia, I., Gagatsis, A., Panaoura, A., Zachariades, T., & Zoulinaki, F. (2009). Geometric and Algebraic Approaches in the Concept of "Limit" and the Impact of the "Didactic Contract". *International Journal of Science and Mathematics Education*, 7(4), 765–790.
- Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's Argument Pattern for studying science discourse. *Science Education*, 88(6), 915–933.*
- ESTABLISH project. (2011). *Report on how IBSE is implemented and assessed in participating countries: Deliverable 2.1*.
- European Commission. (2004). *Increasing human resources for science and technology in Europe: Report of the High Level Group on Human Resources for Science and Technology in Europe, chaired by Prof. José Mariano Gago*. Luxembourg: Office for Official Publications of the European Communities.
- European Commission. (2007). *Science education now: A renewed pedagogy for the future of Europe*. Luxembourg: Office for Official Publications of the European Communities.
- European Parliament, C. (2006). *Key competences for lifelong learning: Summary of the recommendation 2006/962/EC of the European Parliament and of the Council of 18 December 2006 on key competences for lifelong learning*. Retrieved from http://europa.eu/legislation_summaries/education_training_youth/lifelong_learning/c_11090_en.htm
- Fibonacci project. (no date). *Disseminating inquiry-based science and mathematics education in Europe: Principles*. Retrieved from <http://www.fibonacci-project.eu/project/principles>
- Fox-Turnbull, W. (2006). The influences of teacher knowledge and authentic formative assessment on student learning in technology education. *International Journal of Technology and Design Education*, 16(1), 53–77.*
- Fraser, B. J. (1980). Development and validation of a test of enquiry skills. *Journal of Research in Science Teaching*, 17(1), 7–16.
- Fraser, B. J. (1981). *Test of Science-Related Attitudes (TOSRA)*. Melbourne: Australian Council for Educational Research.
- Fraser, B. J., & Butts, W. L. (1982). Relationship between perceived levels of classroom individualization and science-related attitudes. *Journal of Research in Science Teaching*, 19(2), 143–154.

- Freudenthal, H. (1973). *Mathematics as an educational task*. Dordrecht: Kluwer Academic Publishers.
- Furtak, E. M., Hardy, I., Beinbrech, C., Shavelson, R. J., & Shemwell, J. T. (2010). A Framework for Analyzing Evidence-Based Reasoning in Science Classroom Discourse. *Educational Assessment*, 15(3-4), 175–196.
- Furtak, E. M., & Ruiz-Primo, M. A. (2008). Making students' thinking explicit in writing and discussion: An analysis of formative assessment prompts. *Science Education*, 92(5), 799–824.*
- Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P. R., Shavelson, R. J., & Yin, Y. (2008). On the Fidelity of Implementing Embedded Formative Assessments and Its Relation to Student Learning. *Applied Measurement in Education*, 21(4), 360–389.*
- Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and Quasi-Experimental Studies of Inquiry-Based Science Teaching: A Meta-Analysis. *Review of Educational Research*, 82(3), 300–329.
- Furtak, E. M., Shavelson, R. J., Shemwell, J. T., & Figueroa, M. (2012). To teach or not to teach through inquiry: Is that the question? In S. M. Carver & J. Shrager (Eds.), *The journey from child to scientist. Integrating cognitive development and the education sciences* (1st ed., pp. 227–244). Washington, D.C.: American Psychological Association.
- Gallin, P. (2012). Dialogic learning - from an educational concept to daily classroom teaching. In P. Baptist & D. Raab (Eds.), *Resources for Implementing Inquiry in Science and in Mathematics at School. Implementing Inquiry in Mathematics Education* (pp. 23–33). Retrieved from <http://www.fibonacci-project.eu/resources>
- Gardner, J., Harlen, W., Hayward, L., Stobart, G., & Montgomery, M. (2010). *Developing teacher assessment*. Maidenhead: Open University Press.
- Garmire, E., & Pearson, G. (2006). *Tech tally: Approaches to assessing technological literacy*. Washington, DC: National Academies Press.
- Geier, R., Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. *Journal of Research in Science Teaching*, 45(8), 922–939.*
- Genter, D., & Stevens, A. L. (1983). *Mental models*. Hillsdale, London: Lawrence Erlbaum.
- Gerard, L. F., Spitulnik, M., & Linn, M. C. (2010). Teacher use of evidence to customize inquiry science instruction. *Journal of Research in Science Teaching*, 47(9), 1037–1063.*
- Germann, P. J. (1989). The processes of biological investigations test. *Journal of Research in Science Teaching*, 26(7), 609–625.
- Gibson, H. L., & Chase, C. (2002). Longitudinal impact of an inquiry-based science program on middle school students' attitudes toward science. *Science Education*, 86(5), 693–705.*

- Gijlers, H., & Jong, T. de. (2005). The relation between prior knowledge and students' collaborative discovery learning processes. *Journal of Research in Science Teaching*, 42(3), 264–282.*
- Gitomer, D. H., & Duschl, R. A. (1995). Moving toward a portfolio culture in science education. In S. M. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 299–326). Mahwah: Erlbaum.
- Gobert, J. D., Pallant, A. R., & Daniels, J. T. (2010). Unpacking inquiry skills from content knowledge in geoscience: a research and development study with implications for assessment design. *International Journal of Learning Technology*, 5(3), 310–334.*
- Goodnough, K., & Long, R. (2006). Mind mapping as a flexible assessment tool. In M. McMahon, P. Simmons, R. Sommers, D. DeBeats, & F. Crawley (Eds.), *Assessment in science: Practical experiences and education research* (pp. 219–228). Arlington: NSTA Press.*
- Gotwals, A. W., & Songer, N. B. (2009). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education*, 94(2), 2010, 259–281.*
- Griffin, S. (2005). Fostering the development of whole-number sense: Teaching mathematics in the primary grades. In S. Donovan & J. Bransford (Eds.), *How students learn. History, mathematics, and science in the classroom* (pp. 257–308). Washington, D.C: National Academies Press.
- Gustafson, B., MacDonald, D., & Gentilini, S. (2007). Using Talking and Drawing to Design: Elementary Children Collaborating With University Industrial Design Students. *Journal of Technology Education*, 19(1), 19–34.*
- Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview Procedures for Validating Science Assessments. *Applied Measurement in Education*, 10(2), 181–200.*
- Harlen, W. (2007). *The Quality of Learning: Assessment Alternatives for Primary Education* (Primary Review Research Survey No. 3/4). Retrieved from <http://image.guardian.co.uk/sysfiles/Education/documents/2007/11/01/assessment.pdf>
- Harlen, W. (2009). Teaching and learning science for a better future. *School Science Review*, 90(333), 33–41.
- Harlen, W., & James, M. (1997). Assessment and Learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, 4(3), 365–379.
- Harris, C. J., McNeill, K. L., Lizotte, D. J., Marx, R. W., & Krajcik, J. (2006). Usable assessments for teaching science content and inquiry standards. In M. McMahon, P. Simmons, R. Sommers, D. DeBeats, & F. Crawley (Eds.), *Assessment in science: Practical experiences and education research* (pp. 67–87). Arlington: NSTA Press.*
- Harskamp, E., Ding, N., & Suhre, C. (2008). Group Composition and Its Effect on Female and Male Problem-Solving in Science Education. *Educational Research*, 50(4), 307–318.*

- Hatano, G., & Inagaki, K. (1991). Sharing cognition through collective comprehension activity. In B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 331–348). Washington, D.C.: APA.
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112.
- Heinz, J. (2012). *Indicators and instruments in the context of inquiry-based science education*. Münster: Waxmann.
- Heinze, A., Cheng, Y.-H., Ufer, S., Lin, F.-L., & Reiss, K. (2008). Strategies to foster students' competencies in constructing multi-steps geometric proofs: teaching experiments in Taiwan and Germany. *International Journal of Mathematics Education*, 40(3), 443–453.*
- Heritage, M., Kim, J., Vendlinski, T. P., & Herman, J. L. (2009). From Evidence to Action: A Seamless Process in Formative Assessment? *Educational Measurement: Issues and Practice*, 28(3), 24–31.
- Herman, J. L., Osmundson, E., & Silver, D. (2010). *Capturing quality in formative assessment practice: Measurement challenges: CRESST Report 770*. Los Angeles.
- Herrenkohl, L., Palincsar, A., DeWater, L., & Kawasaki, K. (1999). Developing scientific communities in classrooms: A sociocognitive approach. *The Journal of the Learning Sciences*, 8(3-4), 451–493.*
- Herrenkohl, L. R., Tasker, T., & White, B. Y. (2011). Pedagogical practices to support classroom cultures of scientific inquiry. *Cognition and Instruction*, 29(1), 1-44.*
- Hickey, D. T., Taasoobshirazi, G., & Cross, D. (2012). Assessment as learning: Enhancing discourse, understanding, and achievement in innovative science curricula. *Journal of Research in Science Teaching*, 49(10), 1240–1270.*
- Hickey, D. T., & Zuiker, S. J. (2012). Multilevel Assessment for Discourse, Understanding, and Achievement. *Journal of the Learning Sciences*, 21(4), 522–582.*
- Hmelo, C. E., Holton, D. L., & Kolodner, J. L. (2000). Designing to Learn About Complex Systems. *Journal of the Learning Sciences*, 9(3), 247–298.*
- Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and Achievement in Problem-Based and Inquiry Learning: A Response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, 42(2), 99–107.
- Hofstein, A., Navon, O., Kipnis, M., & Mamlok-Naaman, R. (2005). Developing students' ability to ask more and better questions resulting from inquiry-type chemistry laboratories. *Journal of Research in Science Teaching*, 42(7), 791–806.*
- Hogan, K., Nastasi, B. K., & Pressley, M. (1999). Discourse patterns and collaborative scientific reasoning in peer and teacher-guided discussions. *Cognition and Instruction*, 17(4), 379–432.
- Honey, M., & Hilton, M. L. (2011). *Learning science through computer games and simulations*. Washington, D.C: National Academies Press.
- Hong, J.-C., Yu, K.-C., & Chen, M.-Y. (2011). Collaborative learning in technological project design. *International Journal of Technology and Design Education*, 21(3), 335–347.*

- Huang, C. J., Wang, Y. W., Huang, T. H., Chen, Y. C., Chen, H. M., & Chang, S. C. (2011). Performance evaluation of an online argumentation learning assistance agent. *Computers & Education*, 57(1), 1270–1280.*
- Hume, A., & Coll, R. K. (2010). Authentic student inquiry: The mismatch between the intended curriculum and the student-experienced curriculum. *Research in Science & Technological Education*, 28(1), 43–62.
- Hunter, R., & Anthony, G. (2011). Forging Mathematical Relationships in Inquiry-Based Classrooms With Pasifika Students. *Journal of Urban Mathematics Education*, 4(1), 98–119.
- Ingerman, Å., & Collier-Reed, B. (2011). Technological literacy reconsidered: a model for enactment. *International Journal of Technology and Design Education*, 21(2), 137–148.
- INQUIRE project. (2010). *Taking IBSE into secondary education: Report on the conference*. York, UK. Retrieved from <http://www.inquirebotany.org/en/news/taking-ibse-into-secondary-education-188.html>.
- International Technology Education Association. (1996). *Technology for all Americans: A Rationale and Structure for the Study of Technology*. Retrieved from http://www.iteea.org/TAA/PDFs/Taa_RandS.pdf
- Jang, S.-J. (2010). The Impact on Incorporating Collaborative Concept Mapping with Coteaching Techniques in Elementary Science Classes. *School Science and Mathematics*, 110(2), 86–97.*
- Jimenez-Aleixandre, M. P., Rodriguez, A. B., & Duschl, A. R. (2000). 'Doing the Lesson' or 'Doing Science': Argument in high school genetics. *Science Education*, 84(6), 757–792.
- Johnson, C. C., Kahle, J. B., & Fargo, J. D. (2007). Effective teaching results in increased science achievement for all students. *Science Education*, 91(3), 371–383.
- Johnson, S. D., & Daugherty, J. (2008). Quality and Characteristics of Recent Research in Technology Education. *Journal of Technology Education*, 20(1), 16–31.
- Jorde, D., Strømme, A., Sorborg, Ø., Erlien, W., & Mork, S. M. (2003). *Virtual Environments in Science: Viten.no* (Viten reports No. 17). Retrieved from http://www.ituarkiv.no/filearchive/fil_ITU_Rapport_17.pdf
- Kaberman, Z., & Dori, Y. J. (2009). Question Posing, Inquiry, and Modeling Skills of Chemistry Students in the Case-based Computerized Laboratory Environment. *International Journal of Science and Mathematics Education*, 7(3), 597–625.*
- Kelly, G., & Green, J. (1998). The social nature of knowing: Toward a sociocultural perspective on conceptual change and knowledge construction. In B. Guzzetti & C. Hynd (Eds.), *Perspectives on conceptual change* (pp. 145–182). Mahwah, NJ: Erlbaum.
- Kelly, G. J., Druker, S., & Chen, C. (1998). Students' reasoning about electricity: combining performance assessments with argumentation analysis. *International Journal of Science Education*, 20(7), 849–871.*

- Kessler, J. H., & Galvan, P. M. (2007). *Inquiry in Action: Investigating Matter through Inquiry*. A project of the American Chemical Society Education Division, Office of K–8 Science. American Chemical Society. Retrieved from <http://www.inquiry-inaction.org/download/>
- Ketelhut, D., Nelson, B., Clarke, J., & Dede, C. (2010). A Multi-user virtual environment for building higher order inquiry skills in science. *British Journal of Educational Technology, 41*(1), 56–68.
- Ketelhut, D. J. (2007). The Impact of Student Self-efficacy on Scientific Inquiry Skills: An Exploratory Investigation in River City, a Multi-user Virtual Environment. *Journal of Science Education and Technology, 16*(1), 99–111.
- Ketelhut, D. J., & Nelson, B. C. (2010). Designing for real-world scientific inquiry in virtual environments. *Educational Research, 52*(2), 151–167.*
- Khishfe, R. (2008). The Development of Seventh Graders' Views of Nature of Science. *Journal of Research in Science Teaching, 45*(4), 470–496.*
- Kim, H., & Song, J. (2006). The Features of Peer Argumentation in Middle School Students' Scientific Inquiry. *Research in Science Education, 36*(3), 211–233.*
- Kim, K. H., VanTassel-Baska, J., Bracken, B. A., Feng, A., Stambaugh, T., & Bland, L. (2012). Project Clarion: Three Years of Science Instruction in Title I Schools among K-Third Grade Students. *Research in Science Education, 42*(5), 813–829.*
- Kingston, N., & Nash, B. (2011). Formative Assessment: A Meta-Analysis and a Call for Research. *Educational Measurement: Issues and Practice, 30*(4), 28–37.
- Klahr, D., & Dunbar, K. (1988). Dual Space Searching During Scientific Reasoning. *Cognitive Science, 12*, 1–48.
- Klahr, D., Triona, L. M., & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *Journal of Research in Science Teaching, 44*(1), 183–203.*
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.
- Knuth, E. J., Alibali, M. W., McNeil, N. M., Weinberg, A., & Stephens, A. C. (2005). Middle school students' understanding of core algebraic concepts: Equivalence & Variable. *International Journal of Mathematics Education, 37*(1), 68–76.*
- Koedinger, K. R. (1992). *Emergent properties and structural constraints: Advantages of diagrammatic representations for reasoning and learning*. In: AAAI Technical Report SS-92-02, AAAI (pp. 151–156). Retrieved from <https://www.aaai.org/Papers/Symposia/Spring/1992/SS-92-02/SS92-02-031.pdf>
- Koretz, D. (1998). Large-scale Portfolio Assessments in the US: evidence pertaining to the quality of measurement. *Assessment in Education: Principles, Policy & Practice, 5*(3), 309–334.*
- Krajcik, J. S., McNeill, K. L., & Reiser, B. J. (2008). Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Science Education, 92*(1), 1–32.

- Kubasko, D., Jones, M. G., Tretter, T., & Andre, T. (2008). Is it live or is it memorex? Students' synchronous and asynchronous communication with scientists. *International Journal of Science Education*, 30(4), 495–514.*
- Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive Development*, 15, 309–328.
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, London: The University of Chicago Press.
- Kwon, O. N., Park, J. H., & Park, J. S. (2006). Cultivating divergent thinking in mathematics through an open-ended approach. *Asia Pacific Educational Review*, 7(1), 51–61.*
- Kyza, E. A. (2009). Middle-School Students' Reasoning about Alternative Hypotheses in a Scaffolded, Software-Based Inquiry Investigation. *Cognition and Instruction*, 27(4), 277–311.*
- Larkin, J. H., & Simon, H. A. (1987). Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, 11(1), 65–100.
- Latour, B. (1980). Is it possible to reconstruct the research process? Sociology of a brain peptide. In K. D. Knorr, R. Krohn, & R. Whitley (Eds.), 4. *The social process of scientific investigation*. Dordrecht: D. Reidel.
- Lavoie, D. R. (1999). Effects of emphasizing hypothetico-predictive reasoning within the science learning cycle on high school student's process skills and conceptual understandings in biology. *Journal of Research in Science Teaching*, 36(10), 1127–1147.*
- Learning how to Learn Project. (2002). *Learning how to learn Homepage*. Retrieved from <http://www.learntolearn.ac.uk>
- Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (2002). Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of Research in Science Teaching*, 39(6), 497–521.
- Lee, H.-S., & Liu, O. L. (2010). Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective. *Science Education*, 94(4), 665–688.*
- Lee, S. J., Brown, R. E., & Orrill, C. H. (2011). Mathematics Teachers' Reasoning about Fractions and Decimals Using Drawn Representations. *Mathematical Thinking and Learning: An International Journal*, 13(3), 198–220.*
- Liedtke, W. W. (1999). Teacher-Centered Projects: Confidence, Risk Taking and Flexible Thinking (Mathematics). Full text at Web site: <http://www.educ.uvic.ca/connections>. Retrieved from <http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED442612>*
- Lim, S. Y., & Chapman, E. (2013). Development of a short form of the attitudes toward mathematics inventory. *Educational Studies in Mathematics*, 82(1), 145–164.
- Lin, F.-L., Yang, K.-L., & Chen, C.-Y. (2004). The Features and Relationships of Reasoning, Proving and Understanding Proof in Number Patterns. *International Journal of Science and Mathematics Education*, 2(2), 227–256.*

- Lin, S.-S., & Mintzes, J. J. (2010). Learning Argumentation Skills through Instruction in Socioscientific Issues: The Effect of Ability Level. *International Journal of Science and Mathematics Education*, 8(6), 993–1017.*
- Linn, M. C. (2006). Inquiry Learning: Teaching and Assessing Knowledge Integration in Science. *Science*, 313(5790), 1049–1050.*
- Linn, M. C., Clark, D., & Slotta, J. D. (2003). WISE design for knowledge integration. *Science Education*, 87(4), 517–538.
- Linn, M. C., Davis, E. A., & Bell, P. (Eds.). (2004). *Internet environments for science education*. Mahwah: Lawrence Erlbaum Associates Publishers.
- Linn, M. C., Songer, N. B., & Eylon, B. S. (1996). Shifts and convergences in science learning and instruction. In R. Calfee & D. Berliner (Eds.), *Handbook of educational psychology* (pp. 438–490). Riverside, NJ: Macmillan.
- Linn, R., Burton, E., DeStefano, L., & Hanson, M. (1995). *Generalizability of New Standards Project 1993 pilot study tasks in mathematics: CSE Technical Report 392*. Los Angeles.*
- Liu, O. L., Lee, H. S., & Linn, M. C. (2011). Measuring knowledge integration: Validation of four-year assessments. *Journal of Research in Science Teaching*, 48(9), 1079–1107.*
- Liu, O. L., Lee, H.-S., & Linn, M. C. (2010a). An Investigation of Teacher Impact on Student Inquiry Science Performance Using a Hierarchical Linear Model. *Journal of Research in Science Teaching*, 47(7), 807–819.*
- Liu, O. L., Lee, H.-S., & Linn, M. C. (2010b). Multifaceted Assessment of Inquiry-Based Science Learning. *Educational Assessment*, 15(2), 69–86.*
- Looney, J. W. (2011). *Integrating Formative and Summative Assessment: Progress Toward a Seamless System?* (OECD Education Working Papers No. 58).
- Lorenzo, M. (2005). The Development, Implementation, and Evaluation of a Problem Solving Heuristic. *International Journal of Science and Mathematics Education*, 3(1), 33–58.*
- Lubben, F., Sadeck, M., Scholtz, Z., & Braund, M. (2010). Gauging Students' Untutored Ability in Argumentation about Experimental Data: A South African Case Study. *International Journal of Science Education*, 32(16), 2143–2166.*
- Lyon, E. G., Bunch, G. C., & Shaw, J. M. (2012). Navigating the language demands of an inquiry-based science performance assessment: Classroom challenges and opportunities for English learners. *Science Education*, 96(4), 631–651.*
- MacDonald, D., & Gustafson, B. (2004). The Role of Design Drawing Among Children Engaged in a Parachute Building Activity. *Journal of Technology Education*, 16(1), 55–71.*
- Martin, T. S., McCrone, S. M. S., Bower, M. L. W., & Dindyal, J. (2005). The Interplay of Teacher and Student Actions in the Teaching and Learning of Geometric Proof. *Educational Studies in Mathematics*, 60(1), 95–124.*
- Mason, L. (2001). Introducing talk and writing for conceptual change: a classroom study. *Learning and Instruction*, 11(4-5), 305–329.*

- Mathematical Sciences Education Board, & National Research Council. (1993). *Measuring up: Prototypes for mathematics assessment. Perspectives on school mathematics*. Washington, DC: National Academy Press.
- Mathematical Sciences Education Board, N. R. C. (1990). *Reshaping School Mathematics: A Philosophy and Framework for Curriculum*: The National Academies Press. Retrieved from http://www.nap.edu/openbook.php?record_id=1498
- Mattheis, F. E. & Nakayama, G. (1988). *Effects of a Laboratory-Centered Inquiry Program on Laboratory Skills, Science Process Skills, and Understanding of Science Knowledge in Middle Grades Students* (Reports - research/technical). Retrieved from <http://www.eric.ed.gov/PDFS/ED307148.pdf>*
- McElhaney, K. W., & Linn, M. C. (2008). Impacts of students' experimentation using a dynamic visualization on their understanding of motion. In P. A. Kirschner, J. J. G. van Merriënboer, & T. de Jong (Eds.), *Creating a learning world. Proceedings of the 8th International Conference for the Learning Sciences* (Vol. 2, pp. 51–58). International Society of the Learning Sciences 2008. Retrieved from <http://dl.acm.org/citation.cfm?id=1599878>*
- McElhaney, K. W., & Linn, M. C. (2011). Investigations of a Complex, Realistic Task: Intentional, Unsystematic, and Exhaustive Experimenters. *Journal of Research in Science Teaching*, 48(7), 745–770.*
- McLeod, R. J., Berkheimer, G. D., Fyffe, D. W., & Robison, R. W. (1975). The development of criterion-validated test items for four integrated science processes. *Journal of Research in Science Teaching*, 12(4), 415–421.
- McNeill, K. L. (2009). Teachers' use of curriculum to support students in writing scientific arguments to explain phenomena. *Science Education*, 93(2), 233–268.*
- McNeill, K. L. (2011). Elementary Students' Views of Explanation, Argumentation, and Evidence, and Their Abilities to Construct Arguments over the School Year. *Journal of Research in Science Teaching*, 48(7), 793–823.*
- McNeill, K. L., & Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In M. Lovett & P. Shah (Eds.), *Thinking with data: the proceedings of the 33rd Carnegie Symposium on Cognition*. Mahwah: Lawrence Erlbaum Associates Publishers.*
- Mercer, N., Dawes, L., Wegerif, R., & Sams, C. (2004). Reasoning as a scientist: ways of helping children to use language to learn science. *British Educational Research Journal*, 30(3), 359–377.
- Merrill, C., Custer, R. L., Daugherty, J., Westrick, M., & Zeng, Y. (2008). Delivering Core Engineering Concepts to Secondary Level Students. *Journal of Technology Education*, 20(1), 48–64.*
- Mertler, C. A. (no date). *Classroom Assessment Literacy Inventory*. Retrieved from <http://pareonline.net/htm/v8n22/cali.htm>
- Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative Discourse Idealized and Realized: Accountable Talk in the Classroom and in Civic Life. *Studies in Philosophy and Education*, 27(4), 283–297.

- Mioduser, D., & Betzer, N. (2007). The contribution of Project-based-learning to high-achievers' acquisition of technological knowledge and skills. *International Journal of Technology and Design Education*, 18(1), 59–77.*
- Miranda, M. A. de. (2004). The grounding of a discipline: Cognition and instruction in technology education. *International Journal of Technology and Design Education*, 14(1), 61–77.
- Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G. D., Hafter, Amy, Hamel, Larry, Kennedy, Kathleen, Long, Kathy, Morrison, Alissa L., Murphy, Robert, Pena, Patricia, Quellmalz, Edys S., Rosenquist, Anders, Butler Songer, Nancy, Schank, Patricia, Wenk, Amelia, & Wilson, Mark (2003). *Design Patterns for Assessing Science Inquiry: Principled Assessment Designs for Inquiry (PADI Technical Report 1)*. Menlo Park: SRI International, Center for Technology in Learning. Retrieved from http://padi.sri.com/downloads/TR1_Design_Patterns.pdf
- Mislevy, R. J., Steinberg, L. S., Almond, R. G., Haertel, G. D., & Penuel, W. R. (2001). *Levering point for improving educational assessment* (CSE Technical Report No. 534). Los Angeles. Retrieved from <http://www.cse.ucla.edu/products/reports/newTR534.pdf>
- Mistler Jackson, M., & Songer, N. B. (2000). Student motivation and internet technology: Are students empowered to learn science? *Journal of Research in Science Teaching*, 37(5), 459–479.*
- Molebash, P. (no date). *Web of Inquiry (WOI)*. Retrieved from <http://www.webof-inquiry.org>
- Molitor, L. L., & George, K. D. (1976). Development of a test of science process skills. *Journal of Research in Science Teaching*, 13(5), 405–412.
- Moore, K. & Carlson, M. P. (2012). Students' Images of Problem Contexts when Solving Applied Problems. *The Journal of Mathematical Behavior*, 31(1), 48–59.
- Nantawanit, N., Panijpan, B., & Ruenwongsa, P. (2012). Promoting Students' Conceptual Understanding of Plant Defense Responses Using the Fighting Plant Learning Unit (FPLU). *International Journal of Science and Mathematics Education*, 10(4), 827–864.*
- National Research Council. (1996). *National Science Education Standards*. Washington, D.C.: The National Academies Press.
- National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, D.C.: The National Academies Press.
- Newton, P., Driver, R., & Osborne, J. (1999). The place of argumentation in the pedagogy of school science. *International Journal of Science Education*, 21(5), 553–576.*
- Nichols, S., Glass, G. V., & Berliner, D. (2006). *High-stakes testing and student achievement: Does accountability pressure increase student learning?* (Education Policy Analysis Archives No. 14(1)). Retrieved from <http://epaa.asu.edu/ojs/article/view/72>

- Nielsen, J. A. (2012). Co-opting Science: A preliminary study of how students invoke science in value-laden discussions. *International Journal of Science Education*, 34(2), 275–299.*
- Nohda, N. (2000). Teaching by Open-Approach Method in Japanese Mathematics Classroom. *Proceedings of the Conference of the International Group for the Psychology of Mathematics Education (PME)*, 1, 39–53.
- Nolen, S. B. (2003). Learning environment, motivation, and achievement in high school science. *Journal of Research in Science Teaching*, 40(4), 347–368.
- OECD. (2005). *Formative Assessment: Improving Learning in Secondary Classrooms*. Paris: OECD Publishing and Centre for Educational Research and Innovation.
- Ogborn, J., Kress, G., Martins, I., & McGillicuddy, K. (1996). *Explaining science in the classroom*. Buckingham, Philadelphia: Open University Press.
- Oh, E. Y. Y., Treagust, D. F., Koh, T. S., Phang, W. L., Ng, S. L., Sim, G., & Chandrasegaran, A. L. (2012). Using Visualisations in Secondary School Physics Teaching and Learning: Evaluating the Efficacy of an Instructional Program to Facilitate Understanding of Gas and Liquid Pressure Concepts. *Teaching Science*, 58(4), 34–42.*
- Okada, A., & Shum, S. B. (2008). Evidence-Based Dialogue Maps as a Research Tool to Investigate the Quality of School Pupils' Scientific Argumentation. *International Journal of Research & Method in Education*, 31(3), 291–315.*
- Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the Quality of Argumentation in School Science. *Journal of Research in Science Teaching*, 41(10), 994–1020.*
- Osborne, J., Simon, S., Christodoulou, A., Howell-Richardson, C., & Richardson, K. (2013). Learning to argue: A study of four schools and their attempt to develop the use of argumentation as a common instructional practice and its impact on students. *Journal of Research in Science Teaching*, 50(3), 315–347.*
- Pedder, D. (2006). Organizational conditions that foster successful classroom promotion of Learning How to Learn. *Research Papers in Education*, 21(2), 171–200.
- Pell, T., & Jarvis, T. (2001). Developing attitude to science scales for use with children of ages from five to eleven years. *International Journal of Science Education*, 23(8), 847–862.
- Pellegrino, J. W., Baxter, G. P., & Glaser, R. E. (1999). Chapter 9: Addressing the "Two Disciplines" Problem: Linking Theories of Cognition and Learning With Assessment and Instructional Practice. *Review of Research in Education*, 24(1), 307–353.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. E. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: National Academies Press.
- Phelan, J. C., Choi, K., Niemi, D., Vendlinski, T. P., Baker, E. L., & Herman, J. L. (2012). The effects of POWERSOURCE © assessments on middle-school students' math performance. *Assessment in Education: Principles, Policy & Practice*, 19(2), 211–230.*

- Pifarre T. M. (2010). Inquiry Web-based Learning to Enhance Knowledge Construction in Science: A Study in Secondary Education. In B. A. Morris & G. M. Ferguson (Eds.), *Education in a Competitive and Globalizing World. Computer-Assisted Teaching: New Developments* (pp. 63–92).*
- Pijls, M., Dekker, R., & van Hout-Wolters, B. (2007). Reconstruction of a collaborative mathematical learning process. *Educational Studies in Mathematics*, 65(3), 309–329.*
- Pine, J., Aschbacher, P., Roth, E., Jones, M., McPhee, C., Martin, C., Phelps, S., Kyle, T., & Foley, B. (2006). Fifth Graders' Science Inquiry Abilities: A Comparative Study of Students in Hands-On and Textbook Curricula. *Journal of Research in Science Teaching*, 43(5), 467–484.*
- Polya, G. (1957). *How to Solve It*. Princeton, N.J: Princeton University Press.
- PRIMAS project. (2010). *Promoting inquiry in science and mathematics education across Europe: What does inquiry-based learning mean?* Retrieved from <http://www.primas-project.eu/artikel/en/1302/What+exactly+does+inquiry-based+learning+mean/view.do?lang=en>
- Program in Education (no date_a). *Discovery Inquiry Test in Science (DIT)* (Assessment tools in informal science). Retrieved from <http://www.pearweb.org/atis/tools/4>
- Program in Education, (no date_b). *Test of Science Related Attitudes (TOSRA)* (Assessment tools in informal science). Retrieved from <http://www.pearweb.org/atis/tools/13>
- Program in Education, (no date_c). *Views of Scientific Inquiry, Primary School Version (VOSI-P)* (Assessment tools in informal science). Retrieved from <http://www.pearweb.org/atis/tools/22>
- Quellmalz, E., DeBarger, A., Haertel, G., Schank, P., Buckley, B., Gobert, J., Horwitz, P., & Ayala, C. (2007). *Exploring the Role of Technology-Based Simulations in Science Assessment: The Calipers Project*. Paper presented at the American Educational Research Association (AERA), Chicago.
- Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and Testing. *Science*, 323, 75–79.
- Quellmalz, E. S., Timms, M. J., & Buckley, B. (2010). The promise of simulation-based science assessment: the Calipers project. *International Journal of Learning Technology*, 5(3), 243–263.
- Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: Integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching*, 49(3), 363–393.
- R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org>
- Reiss, K. M., Heinze, A., Renkl, A., & Groß, C. (2008). Reasoning and proof in geometry: effects of a learning environment based on heuristic worked-out examples. *International Journal of Mathematics Education*, 40(3), 455–467.*

- Repenning, A., Ioannidou, A., Luhn, L., Daetwyler, C., & Repenning, N. (2010). Mr. Vetro: Assessing a Collective Simulation Framework. *Journal of Interactive Learning Research*, 21(4), 515–537.*
- Reyes, I. (2008). English Language Learners' Discourse Strategies in Science Instruction. *Bilingual Research Journal*, 31(1), 95–114.*
- Reys, R., Reys, B., Lapan, R., Holiday, G., & Wasman, D. (2003). Assessing the impact of standards-based middle grades mathematics curriculum materials on student achievement. *Journal for Research in Mathematics Education*, 34(1), 74–95.*
- Rivet, A. E., & Kastens, K. A. (2012). Developing a construct-based assessment to examine students' analogical reasoning around physical models in Earth Science. *Journal of Research in Science Teaching*, 49(6), 713–743.*
- Rivet, A. E., & Krajcik, J. S. (2004). Achieving Standards in Urban Systemic Reform: An Example of a Sixth Grade Project-Based Science Curriculum. *Journal of Research in Science Teaching*, 41(7), 669–692.*
- Rodríguez, E., Bosch, M. & Gascón, J. (2008). A networking method to compare theories: metacognition in problem solving reformulated within the Anthropological Theory of the Didactic. *ZDM*, 40(2), 287–301.
- Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student Self-Evaluation in Grade 5-6 Mathematics Effects on Problem- Solving Achievement. *Educational Assessment*, 8(1), 43–58.*
- Rossouw, A., Hacker, M., & Vries, M. J. de. (2011). Concepts and contexts in engineering and technology education: an international and interdisciplinary Delphi study. *International Journal of Technology and Design Education*, 21(4), 409–424.
- Rubel, L. H. (2007). Middle school and high school students' probabilistic reasoning on coin tasks. *Journal for Research in Mathematics Education*, 38(5), 531–556.*
- Ruiz-Primo, M. A., & Furtak, E. M. (2006). Informal Formative Assessment and Scientific Inquiry: Exploring Teachers' Practices and Student Learning. *Educational Assessment*, 11(3-4), 205–235.*
- Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring Teachers' Informal Formative Assessment Practices and Students' Understanding in the Context of Scientific Inquiry. *Journal of Research in Science Teaching*, 44(1), 57–84.*
- Ruiz-Primo, M. A., Li, M., Ayala, C., & Shavelson, R. J. (2004). Evaluating students' science notebooks as an assessment tool. *International Journal of Science Education*, 26(12), 1477–1506.*
- Ruiz-Primo, M. A., Li, M., Tsai, S.-P., & Schneider, J. (2010). Testing one premise of scientific inquiry in science classrooms: Examining students' scientific explanations and student learning. [References]. *Journal of Research in Science Teaching*, 47(5), 583–608.*
- Ruiz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M.-C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching*, 49(6), 691–712.*

- Ruiz-Primo, M. A. & Shavelson, R. J. (1997). *Concept-Map based assessment: On possible sources of sampling variability*. Los Angeles. Retrieved from http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_ERICExtSearch_SearchValue_0=ED422403&ERICExtSearch_SearchType_0=no&accno=ED422403
- Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, 39(5), 369–393.*
- Ryu, S., & Sandoval, W. A. (2012). Improvements to Elementary Children's Epistemic Understanding from Sustained Argumentation. *Science Education*, 96(3), 488–526.*
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144.
- Sadler, D. R. (1998). Formative Assessment: revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, 5(1), 77–84.
- Samarapungavan, A., Mantzicopoulos, P., & Patrick, H. (2008). Learning science through inquiry in kindergarten. *Science Education*, 92(5), 868–908.*
- Samarapungavan, A., Patrick, H., & Mantzicopoulos, P. (2011). What kindergarten students learn in inquiry-based science classrooms. *Cognition and Instruction*, 29(4), 416–470.*
- Sampson, V., Grooms, J., & Walker, J. P. (2011). Argument-Driven Inquiry as a Way to Help Students Learn How to Participate in Scientific Argumentation and Craft Written Arguments: An Exploratory Study. *Science Education*, 95(2), 217–257.*
- Santau, A. O., Maerten-Rivera, J. L., & Huggins, A. C. (2011). Science achievement of English language learners in urban elementary schools: Fourth-grade student achievement results from a professional development intervention. *Science Education*, 95(5), 771–793.*
- Saunders-Stewart, K. S., Gyles, P. D. T., & Shore, B. M. (2012). Student Outcomes in Inquiry Instruction: A Literature-Derived Inventory. *Journal of Advanced Academics*, 23(1), 5–31.
- Scardamalia, M., & Bereiter, C. (1994). The CSILE project: Trying to bring the classroom into world 3. In K. McGilly (Ed.), *Classroom Lessons: Integrating Cognitive Theory and Classroom Practice*. Cambridge, MA: MIT Press/Bradford Books.
- Schaal, S., Bogner, F. X., & Girwidz, R. (2010). Concept Mapping Assessment of Media Assisted Learning in Interdisciplinary Science Education. *Research in Science Education*, 40(3), 339–352.*
- Schneider, R. M., Krajcik, J., Marx, R. W., & Soloway, E. (2002). Performance of students in project-based science classrooms on a national measure of science achievement. *Journal of Research in Science Teaching*, 39(5), 410–422.*
- Schnittka, C., & Bell, R. (2011). Engineering Design and Conceptual Change in Science: Addressing thermal energy and heat transfer in eighth grade. *International Journal of Science Education*, 33(13), 1861–1887.*
- Schoenfeld, A. H. (1985). *Mathematical problem solving*. San Diego: Academic Press.

- Schukajlow, S., Leiss, D., Pekrun, R., Blum, W., Müller, M., & Messner, R. (2012). Teaching methods for modelling problems and students' task-specific enjoyment, value, interest and self-efficacy expectations. *Educational Studies in Mathematics*, 79(2), 215–237.*
- Schwartz, R. S., Lederman, N. G., & Lederman, J. S. (2008). *An Instrument To Assess Views Of Scientific Inquiry: The VOSI Questionnaire: Paper presented at the annual meeting of the National Association for Research in Science Teaching Teaching*, Baltimore. Retrieved from <http://homepages.wmich.edu/~rschwart/docs/VOSInarst08.pdf>
- Schwarz, C. V., & White, B. Y. (2005). Metamodeling Knowledge: Developing Students' Understanding of Scientific Modeling. *Cognition and Instruction*, 23(2), 165–205.*
- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Monograph Series on Educational Evaluation: Vol. 1. Perspectives of curriculum evaluation* (pp. 39–83). Chicago: Rand McNally.
- Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance Assessment in Science. *Applied Measurement in Education*, 4(4), 347–362.*
- Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., Tomita, M. K., & Yin, Y. (2008). On the Impact of Curriculum-Embedded Formative Assessment on Learning: A Collaboration between Curriculum and Assessment Developers. *Applied Measurement in Education*, 21(4), 295–314.*
- Shemwell, J. T., & Furtak, E. M. (2010). Science Classroom Discussion as Scientific Argumentation: A Study of Conceptually Rich (and Poor) Student Talk. *Educational Assessment*, 15(3), 222–250.*
- Shepard, L. A. (2000). The Role of Assessment in a Learning Culture. *Educational Researcher*, 29(7), 4–14.
- Shepard, L. A. (2003). Reconsidering Large-Scale Assessment to Heighten Its Relevance to Learning. In J. M. Atkin & J. E. Coffey (Eds.), *Science Educators' Essay Collection. Everyday Assessment in the Science Classroom* (pp. 121–146). Arlington: NSTA Press.
- Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, 78(1), 153–189.
- Shymansky, J. A., Yore, L. D., & Anderson, J. O. (2004). Impact of a School District's Science Reform Effort on the Achievement and Attitudes of Third- and Fourth-Grade Students. *Journal of Research in Science Teaching*, 41(8), 771–790.*
- Siegel, M. A., Hynds, P., Siciliano, M., & Nagle, B. (2006). Using rubrics to foster meaningful learning. In M. McMahon, P. Simmons, R. Sommers, D. DeBeats, & F. Crawley (Eds.), *Assessment in science: Practical experiences and education research* (pp. 89–106). Arlington: NSTA Press.*
- Silk, E. M., Schunn, C. D., & Cary, M. S. (2009). The Impact of an Engineering Design Curriculum on Science Reasoning in an Urban Setting. *Journal of Science Education and Technology*, 18(3), 209–223.*

- Simons, K. D., & Klein, J. D. (2007). The impact of scaffolding and student achievement levels in a problem-based learning environment. *Instructional Science*, 35(1), 41–72.*
- Smith, E. L. (1991). A conceptual change model of learning science. In S. M. Glynn, R. H. Yeany, & B. K. Britton (Eds.), *The psychology of learning science* (pp. 43–63). Hillsdale, NJ: Erlbaum.
- So, W. W.-M. (2003). Learning Science through investigations: An experience with Hong Kong primary school children. *International Journal of Science and Mathematics Education*, 1(2), 175–200.*
- Southerland, S., Kittleson, J., Settlage, J., & Lanier, K. (2005). Individual and Group Meaning-Making in an Urban Third Grade Classroom: Red Fog, Cold Cans, and Seeping Vapor. *Journal of Research in Science Teaching*, 42(9), 1032–1061.*
- Spires, H. A., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011). Problem Solving and Game-Based Learning: Effects of Middle Grade Students' Hypothesis Testing Strategies on Learning Outcomes. *Journal of Educational Computing Research*, 44(4), 453–472.*
- SRI International. (2007). *Principled Assessment Designs for Inquiry (PADI): advancing evidence-centered assessment design*. Retrieved from <http://padi.sri.com/index.html>
- Stecher, B. M., Klein, S. P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R. J., & Haertel, E. (2000). The Effects of Content, Format, and Inquiry Level on Science Performance Assessment Scores. *Applied Measurement in Education*, 13(2), 139–160.*
- Steinberg, R. N., Cormier, S., & Fernandez, A. (2009). Probing Student Understanding of Scientific Thinking in the Context of Introductory Astrophysics. *Physical Review Special Topics - Physics Education Research*, 5(2), 020104-1–020104-10.*
- Stieff, M. (2011). Improving representational competence using molecular simulations embedded in inquiry activities. *Journal of Research in Science Teaching*, 48(10), 1137–1158.
- Strike, K. A., & Posner, G. J. (1985). A conceptual change view of learning and understanding. In West, L. H. T. & A. Pines (Eds.), *Cognitive Structure and Conceptual Change* (pp. 211–231). New York: Academic Press.
- Taasoobshirazi, G., & Hickey, D. T. (2005). Promoting Argumentative Discourse: A Design-Based Implementation and Refinement of an Astronomy Multimedia Curriculum, Assessment Model, and Learning Environment. *Astronomy Education Review*, 4(1), 53–70.*
- Taasoobshirazi, G., Zuiker, S. J., Anderson, K. T., & Hickey, D. T. (2006). Enhancing Inquiry, Understanding, and Achievement in an Astronomy Multimedia Learning Environment. *Journal of Science Education and Technology*, 15(5-6), 383–395.*
- Tamir, P., Nussinovitz, R., & Friedler, Y. (1982). The design and use of a Practical Tests Assessment Inventory. *Journal of Biological Education*, 16(1), 42–50.
- Tannenbaum, R. S. (1971). The development of the test of science processes. *Journal of Research in Science Teaching*, 8(2), 123–136.

- Temiz, B. K., Taşar, M., & Tan, F. (2006). Development and validation of a multiple format test of science process skills. *International Education Journal*, 7(7), 1007–1027.
- The Open University & Sheffield Hallam University. (2008). *FAST Website*. Retrieved from <http://www.open.ac.uk/fast/>
- Thomson Reuters (2012). *About Journal Citation Reports*. Retrieved from http://admin-apps.webofknowledge.com/JCR/help/h_jcrabout.htm
- Thomson Reuters (2013). *Web of knowledge – Journal citation reports*. Retrieved from <http://admin-apps.webofknowledge.com/JCR/JCR?PointOfEntry=Home&SID=3F36dpJCKemKLP7aK2p>
- Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). “Mapping to know”: The effects of representational guidance and reflective assessment on scientific inquiry. *Science Education*, 86(2), 264–286.*
- Toulmin, S. E. (1972). *Human Understanding: The Collective Use and Evolution of Concepts*. Princeton, NJ: Princeton University Press.
- Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.
- Trefil, J. (2008). *Why Science?* New York: Teachers College Press.
- Tsai, P.-S., Hwang, G.-J., Tsai, C.-C., Hung, C.-M., & Huang, I. (2012). An Electronic Library-based Learning Environment for Supporting Web-based Problem-Solving Activities. *Educational Technology and Society*, 15(4), 252–264.*
- Tytler, R., Haslam, F., Prain, V., & Hubber, P. (2009). An Explicit Representational Focus for Teaching and Learning about Animals in the Environment. *Teaching Science*, 55(4), 21–27.*
- Tzur, R. (2007). Fine grain assessment of students’ mathematical understanding: participatory and anticipatory stages in learning a new mathematical conception. *Educational Studies in Mathematics*, 66(3), 273–291.*
- University of Berkeley. (2013). *WISE web-based inquiry science environment*. Retrieved from <http://wise.berkeley.edu/webapp/index.html>
- Urhahne, D., Schanze, S., Bell, T., Mansfield, A., & Holmes, J. (2010). Role of the Teacher in Computer-supported Collaborative Inquiry Learning. *International Journal of Science Education*, 32(2), 221–243.
- Valanides, N., & Angeli, C. (2008). Distributed Cognition in a Sixth-Grade Classroom: An Attempt to Overcome Alternative Conceptions about Light and Color. *Journal of Research on Technology in Education*, 40(3), 309–336.*
- van Aalst, J., & Mya Sioux Truong. (2011). Promoting Knowledge Creation Discourse in an Asian Primary Five Classroom: Results from an inquiry into life cycles. *International Journal of Science Education*, 33(4), 487–515.*
- van Joolingen, W., Jong, T. de, Lazonder, A., Savelsbergh, E., & Manlove, S. (2005). Co-Lab: Research and development of an online learning environment for collaborative scientific discovery learning. *Computers in Human Behavior*, 21(4), 671–688.

- van Niekerk, E., Piet Ankievicz, & Swardt, E. de. (2010). A process-based assessment framework for technology education: a case study. *International Journal of Technology and Design Education*, 20(2), 191–215.*
- Vasconcelos, C. (2012). Teaching Environmental Education through PBL: Evaluation of a Teaching Intervention Program. *Research in Science Education*, 42(2), 219–232.*
- Veal, W. R., & Chandler, A. T. (2008). Science Sampler: The Use of Stations to Develop Inquiry Skills and Content for Rock Hounds. *Science Scope*, 32(1), 54–57.*
- Vellom, R. P., & Anderson, C. W. (1999). Reasoning about data in middle school science. *Journal of Research in Science Teaching*, 36(2), 179–199.*
- Verschaffel, L., Corte, E. de, Vierstraete, H. (1999). Upper elementary school pupils' difficulties in modeling and solving nonstandard additive word problems involving ordinal numbers. *Journal for Research in Mathematics Education*, 30(3), 265–285.
- Vries, M. J. de, & Mottier, I. (Eds.). (2006). *International Handbook of Technology Education: Reviewing the past twenty years*. Rotterdam: Sense Publishers.
- Waddington, D., Nentwig, P., & Schanze, S. (2007). *Making it comparable. Standards in science education*. Münster: Waxmann.
- Watson, A. (2006). Some difficulties in informal assessment in mathematics. *Assessment in Education: Principles, Policy & Practice*, 13(3), 289–303.
- Webb, N. M., Nemer, K. M., & Ing, M. (2006). Small-group reflections: Parallels between teacher discourse and student Behavior in peer-directed groups. *Journal of the Learning Sciences*, 15(1), 63–119.*
- White, B. Y., & Frederiksen, J. R. (1998). Inquiry, Modeling, and Metacognition: Making Science Accessible to All Students. *Cognition and Instruction*, 16(1), 3–118.*
- Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment*, 11(3-4), 283–289.
- Wiliam, D. (2007). Keeping Learning on Track. Classroom Assessment and the Regulation of Learning. In F. K. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 1053-1098). Charlotte, NC: Information Age Publishing.
- Wiliam, D. (2008). International comparisons and sensitivity to instruction. *Assessment in Education: Principles, Policy & Practice*, 15(3), 253–257.
- Williams, J., & Ryan, J. (2000). National Testing and the Improvement of Classroom Teaching: Can they coexist? *British Educational Research Journal*, 26(1), 49–73.
- Williams, P. J. (2012). Investigating the Feasibility of Using Digital Representations of Work for Performance Assessment in Engineering. *International Journal of Technology and Design Education*, 22(2), 187–203.*
- Wilson, C. D., Taylor, J. A., Kowalski, S. M., & Carlson, J. (2010). The relative effects and equity of inquiry-based and commonplace science teaching on students' knowledge, reasoning, and argumentation. *Journal of Research in Science Teaching*, 47(3), 276–301.*

- Wilson, M., & Scalise, K. (2003). Reporting Progress to Parents and Others: Beyond Grades. In J. M. Atkin & J. E. Coffey (Eds.), *Science Educators' Essay Collection. Everyday Assessment in the Science Classroom* (pp. 89–108). Arlington: NSTA Press.
- Wilson, M., & Sloane, K. (2000). From Principles to Practice: An Embedded Assessment System. *Applied Measurement in Education*, 13(2), 181–208.*
- Winters, F. I., & Alexander, P. A. (2011). Peer collaboration: the relation of regulatory behaviors to learning with hypermedia. *Instructional Science*, 39(4), 407–427.*
- Wirth, J., & Klieme, E. (2003). Computer-based Assessment of Problem Solving Competence. *Assessment in Education: Principles, Policy & Practice*, 10(3), 329–345.*
- Wong, K. K. H., & Day, J. R. (2009). A Comparative Study of Problem-Based and Lecture-Based Learning in Junior Secondary School Science. *Research in Science Education*, 39(5), 625–642.*
- Wood, T., & Sellers, P. (1997). Deepening the analysis: Longitudinal assessment of a problem-centered mathematics program. *Journal for Research in Mathematics Education*, 28(2), 163–186.*
- Woods, T., Williams, G., & McNeal, B. (2006). Children's mathematical thinking in different classroom cultures. *Journal for Research in Mathematics Education*, 37(3), 222–255.*
- Worcester Polytechnic Institute. (2013). *ASSISTments: Formative assessment that exists*. Retrieved from <https://www.assistments.org/>
- Yin, Y., Vanides, J., Ruiz-Primo, M. A., Ayala, C. C., & Shavelson, R. J. (2005). Comparison of two concept-mapping techniques: Implications for scoring, interpretation, and use. *Journal of Research in Science Teaching*, 42(2), 166–184.*
- Yoon, C. H. (2009). Self-regulated learning and instructional factors in the scientific inquiry of scientifically gifted Korean middle school students. *Gifted Child Quarterly*, 53(3), 203–216.*
- Young, B. J., & Lee, S. K. (2005). The effects of a kit-based science curriculum and intensive science professional development on elementary student science achievement. *Journal of Science Education and Technology*, 14(5-6), 471–481.*
- Zhang, J., & Sun, Y. (2011). Reading for Idea Advancement in a Grade 4 Knowledge Building Community. *Instructional Science: An International Journal of the Learning Sciences*, 39(4), 429–452.*
- Zhang, L., Wilson, L., & Manon, J. (1999). An Analysis of Gender Differences on Performance Assessment in Mathematics – A Follow-Up Study. Retrieved from <http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED431791>*
- Zion, M., Michalsky, T., & Mevarech, Z. R. (2005). The effects of metacognitive instruction embedded within an asynchronous learning network on scientific inquiry skills. *International Journal of Science Education*, 27(8), 957–983.*

Note: Not all of the 191 publications found within the literature review are cited in the reference list. Publications from the review are indicated with an asterisk.

Figures

Figure 1: A sample gravity problem from a physics test (White & Frederiksen, 1998, p. 60)	62
Figure 2: Formative assessment item on dominance relationships (Hickey & Zuiker, 2012, p. 24)	63
Figure 3: Given concepts and linking words for the construction of a concept map in biology (Brandstädter et al., 2012, p. 2167)	64
Figure 4: Activity-oriented quiz (Hickey et al., 2012, p. 1247).....	66
Figure 5: Feedback conversation guidelines (Hickey et al., 2012, p. 1248).....	67
Figure 6: Examples of questions for a semi-structured interview (Dawson & Venville, 2009, p. 1445).....	69
Figure 7: Assessment rubric for self-assessment (van Niekerk, Piet Ankiewicz, & Swardt, 2010, p. 213).....	70
Figure 8: Help me peel task and photo (Fox-Turnbull, 2006, p. 59).....	76
Figure 9: Hands-on and virtual mousetraps (Klahr et al., 2007, pp. 188–189).....	77
Figure 10: The items of the pre-test (Heinze et al., 2008, p. 448).....	79
Figure 11: Using the concept of mathematical equivalence (Knuth et al., 2005, p. 70).....	79
Figure 12: “Dressed up” world problem “football pitch” (Schukajlow et al., 2012, p. 225)	79
Figure 13: Goals, Plan, Action and Reflection sheet in original and revised version (Brookhart et al., 2004, pp. 216–217).....	80
Figure 14: ‘hot spots’ of inquiry in science education	82
Figure 15: ‘hot spots’ of inquiry in technology education	83
Figure 16: ‘hot spots’ of inquiry in mathematics education	83

Tables

Table 1: Aspects of IBE in STM	10
Table 2: Starting point for the identification of possible connections between IBE and formative assessment.....	20
Table 3: Keywords for searches in data bases.....	24
Table 4: Results of the searches in data bases.....	26
Table 5: Relevant journals and their impact factors.....	27
Table 6: Results of the searches in the issues of relevant journals by subject	28
Table 7: Categorization of literature	29
Table 8: Final extract for the literature review	30
Table 9: Scheme for the evaluation of the literature	31
Table 10: Number of studies investigating ‘diagnosing problems/ identifying questions’	39
Table 11: Number of studies investigating ‘searching for information’	40
Table 12: Number of studies investigating ‘considering alternative or multiple solutions/ searching for alternatives/ modifying designs’	42
Table 13: Number of studies investigating ‘creating mental representations’	43
Table 14: Number of studies investigating ‘constructing and using models’	44
Table 15: Number of studies investigating ‘formulating hypotheses/ researching conjectures’	46
Table 16: Number of studies investigating ‘planning investigations’	47
Table 17: Number of studies investigating ‘constructing prototypes’	47
Table 18: Number of studies investigating ‘finding structures or patterns’	49
Table 19: Number of studies investigating ‘collecting and interpreting data/ evaluating results’	51
Table 20: Number of studies investigating ‘constructing and critiquing arguments or explanations, argumentation, reasoning, and using evidence’	54
Table 21: Number of studies investigating ‘communication/ debating with peers’.....	55
Table 22: Number of studies investigating ‘searching for generalizations’	56
Table 23: Number of studies investigating ‘dealing with uncertainty’	56
Table 24: Number of studies investigating ‘problem solving’	57
Table 25: Number of studies investigating ‘IBE and inquiry process skills in general’..	59
Table 26: Number of studies investigating ‘knowledge/ achievement/ understanding..	60
Table 27: Assessment practices by subject	61
Table 28: Character of the assessment	61
Table 29: Holistic concept mapping scoring rubric (Nantawanit et al., 2012).....	64
Table 30: Frequency of assessment methods in the studies from the field of science education.....	71
Table 31: Frequency of assessment methods in the studies from the field of technology education.....	75
Table 32: Frequency of assessment methods in the studies from the field of mathematics education	78

Report from the FP7 project:

Assess Inquiry in Science, Technology and Mathematics Education



ASSISTME

Report on current state of the art in formative and summative assessment in IBE in STM

– Part II –

Contribution from Pearson Education International

A literature review to inform the development of digital
assessments which are relevant to the aims of the ASSIST-ME
project

Sarah Hughes, Clare Green, & Vanessa Greene

Delivery date	31.07.2013
Deliverable number	D2.4
Lead participant	Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany
Contact person	Rose Clesham, rose.clesham@pearson.com – PEI
Dissemination level	PU

Table of Contents

1. BACKGROUND AND CONTEXT	4
1.1 The ASSIST-ME project	4
1.2 Work Package 2 – the literature review.....	4
1.3 The purpose and objectives relating to this report	5
2. METHODOLOGY	6
2.1 E-assessment	6
2.2 Sources	7
2.3 Search terms.....	8
3. FINDINGS	9
3.1 Aspects of e-assessment tasks/items	13
3.2 Assessment/teaching link	20
3.3 Inquiry based education	24
3.4 Competency-based learning	25
3.5 Formative assessment.....	26
3.6 Summative Assessment.....	30
3.7 Formative / summative assessment link	31
3.8 Feedback.....	32
3.9 Effects of e-assessment on the learner	37
3.10 Quality	39
3.11 Implications for the implementation of e-assessment	42
3.12 Implications for evaluation of assessments.....	45
3.13 Exemplars	46

4. CONCLUSIONS AND IMPLICATIONS.....	54
4.1 Reflections on the methodology and the process of searching	54
4.2 Conclusions specifically related to objectives	55
4.3 E-assessment and formative assessment.....	61
4.4 Last words.....	61
REFERENCES.....	63
APPENDICES	69
Appendix 1. Search terms used in literature review.....	69
Appendix 2. Journals searched	70
Appendix 3. Sources viewed with no relevant content	71

1. Background and context

1.1 The ASSIST-ME project

The “Assess Inquiry in Science, Technology and Mathematics Education (ASSIST-ME)” project is an EU funded Europe-wide project which aims to investigate formative and summative assessment methods to support and to improve inquiry-based approaches in European science, technology and mathematics (STM) education.

A number of work packages will constitute the project which began in early 2013 and will be complete in 2017.

Work package	Title
1	Project Management
2	Synthesize existing research
3	Characterise Educational Systems
4	Design Assessment Methods
5	Trial Implementation of Assessment Methods
6	Transform Results into National Contexts
7	Promotion of Guidelines and Results

Based on an initial analysis of the literature (WP2) to identify what is known about summative and formative assessment of knowledge, skills and attitudes related to key STM competences and an analysis of European educational systems (WP3), the project will design a range of assessment methods (WP4). These methods will be tested as part of the project in primary and secondary schools in different educational cultures in Europe (WP5). This will enable an analysis of the conditions that support or undermine the uptake of formative assessment related to inquiry processes (WP6).

Reflections on the development and trialling of the assessments will enable the formulation of guidelines and recommendations for policy makers, curriculum developers, teacher trainers and other stakeholders in the different European educational systems (WP7).

1.2 Work Package 2 – the literature review

A number of work packages (WPs) make up the project. WP2 comprises a literature review which aims to analyse existing research on how summative and formative assessment of knowledge, competences and attitudes in STM can be coupled with inquiry-based teaching. Pearson Education’s role in that literature review is to review the use of *e-assessment* in the formative and summative assessment of STM subjects at primary and secondary levels with a focus on inquiry-based and competence-based learning.

Pearson Education's work on WP2 is an independent forerunner for Pearson's work on WPs 4 and 5 (the development and trialling of assessments). This independence will ensure that the best practice identified in the literature review is reported and implications for the development of e-assessments are described, without reference to what might be possible or desirable, from Pearson's point of view, to propose as a design for the ASSIST-ME e-assessments.

1.3 The purpose and objectives relating to this report

This report is the output of Pearson's work in WP2.

The purpose of this report is to provide a literature review that will inform the development of digital assessments which are relevant to the aims of ASSIST-ME, that is, it will:

1. enable both formative and summative assessment
2. cover STM subjects
3. focus on inquiry-based education
4. focus on competence-based learning
5. be relevant to primary and/or secondary education.

To enable us to fulfil that purpose, our objectives were to

1. through the literature, identify theories and models which are relevant to the development of such digital assessments,
2. identify strategies used in the evaluation of the models which could inform good practice,
3. identify existing relevant digital assessments,
4. identify implications for the development of the digital assessments relevant to the aims of ASSIST-ME.

This report contains sections on the methodology employed followed by the findings, described under themed heading including implications for work packages 4 and 5: the development of e-assessments for ASSIST-ME. The conclusions are structured around the four objectives listed above.

2. Methodology

2.1 E-assessment

The concepts of competence, inquiry-based STM education and assessment are all described in the ASSIST-ME project proposal, where definitions of terms are shared. We would like to clarify what we mean by e-assessment in the context of the ASSIST-ME project because this impacts on the scope and implications of the work. Beevers and Winkley (2011) provide two definitions of e-assessment, one focuses on the e-administration of tests and the other one relates to providing automation in the pedagogic process:

- A. e-assessment occurs when there is an automated marking/response to student input on-screen in a test, informing on the process of answering a question and providing feedback to learners and their teachers through well-crafted advice and reports.

Alternatively,

- B. e-assessment occurs when there is use of technology in testing which encompasses the on-screen computer-marked assessments of (A) above but also includes on-screen human marking of tests, electronic management and presentation of results, moderation and awarding processes with awarding bodies, anti-plagiarism software, tools which enable collaboration on the assessment and feedback processes, voting systems/clickers and e-portfolios.

This difference in scope is not helpful as it confuses those who are not 'into e-assessment' and even allows experts to talk at crossed purposes at times.

JISC (2007) give an alternative definition of e-assessment: The range of activities in which digital technologies are used in assessment – designing and delivering assessments, marking, processes of reporting, storing and transferring data. More recently, Broadfoot et al. (2013a) explored technology-enhanced assessment which refers to the wide range of ways in which technology can be used to support assessment and feedback. It includes on-screen assessment, often called e-assessment.

For the purpose of this literature review e-assessment was taken to include:

- the onscreen presentation of tasks and tests,
- delivery of assessments,
- automated marking,
- automated feedback to students,
- students' onscreen and digital responses,
- creation, management and manipulation of data for teachers and
- tools which enable collaboration on the assessment and feedback processes.

2.2 Sources

Initial searches led to a list of productive journals. These, and others, were searched using the terms described in section 0 and Appendix 1, covering at least the last 10 years. Searches were not confined geographically; all countries were included, but only those sources which were available in the English language were considered.

Priority journals:

1. Educational Technology, Research and Assessment
2. Journal of Technology, Learning, and Assessment
3. British Journal of Educational Technology

Other journals:

4. Computer-Based Testing
5. Computers and Education
6. Education and Information Technologies
7. European Journal of Education: special issue – ICT and Education
8. Frontiers in Artificial Intelligence and Information and Communication Technologies
9. International Encyclopaedia of Education (Technology and Learning - assessment)
10. International Journal of Computer-Supported Collaborative Learning
11. International Journal of E-assessment (journal of the e-assessment association)
12. International Journal of Educational Research
13. Journal of Applied Testing Technology
14. Journal of Computer Assisted Learning
15. Journal of Information Technology in Teacher Education
16. Journal of Research on Computing in Education
17. Journal of Science Education and Technology
18. Learning, Media and Technology
19. Research in Learning Technology (Journal of the Association of Learning Technology)

Other types of sources that were searched included:

- Organisations (e.g. NFER, BECTA, OECD)
- Government websites (e.g. Ofqual, Office for Official Publications of the European Communities)
- Specific assessment projects and examples of online assessments (e.g. Operation ARIES!)
- Conferences (e.g. Computer Aided Learning Conference, International Conference on Intelligent Tutoring Systems)
- University departments (e.g. The Centre for Mathematics, Science and Computer Education, Rutgers University)

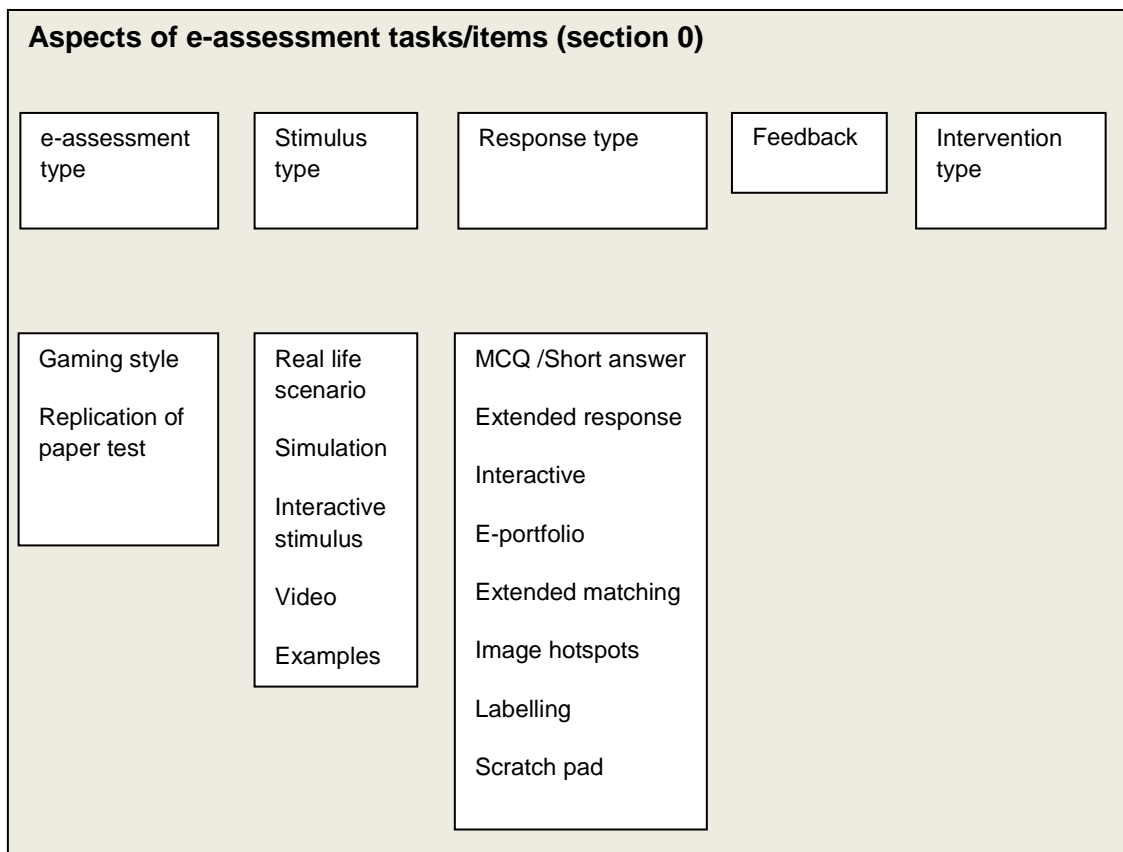
2.3 Search terms

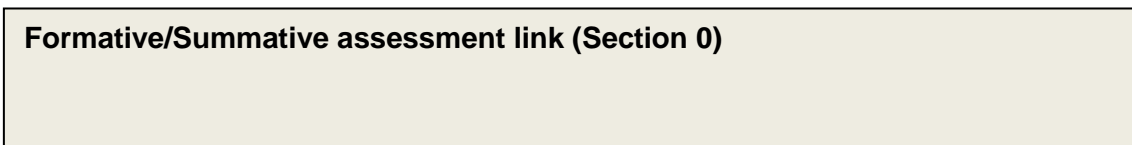
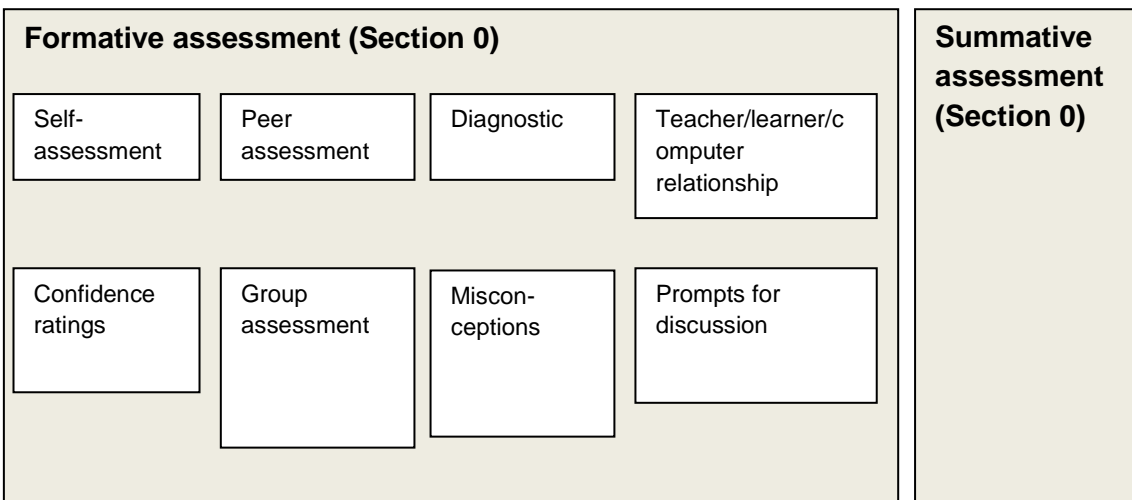
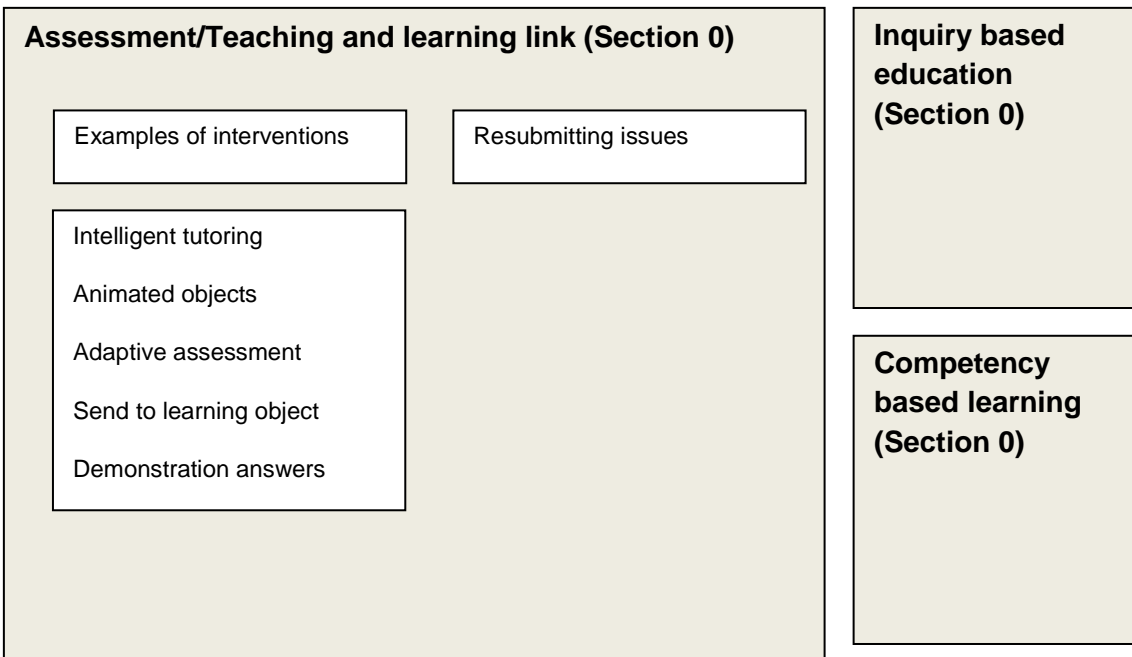
All searches targeted e-assessment (which incorporated e-learning). We searched for instances where e-assessment occurred as a key word alongside these other keywords or their alternatives shown in Appendix 1:

- formative assessment
- summative assessment
- inquiry based education
- competencies
- mathematics
- science
- technology

3. Findings

Having applied the methods, approximately eighty articles and sources were included in this literature review. Emerging from the literature was a number of interrelated themes relevant to the use of e-assessment in the ASSIST-ME project. A visual representation of these is presented below, followed by a description of findings for each theme and sub-theme. This visual representation does not do justice to the extent of overlap between the themes. Given this overlap between themes, it is possible that the findings could have been organised in a number of different ways, but these themes and their order were chosen to best fit with the aims and characteristics of the ASSIST-ME assessments.





Feedback (Section 0)

Output to teacher

Feedback to learner

Learning analytics

Feedback level

Characteristics of feedback:

Type Level of detail

Timing Frequency

Effects of e-assessment on learner (Section 0)

Motivation

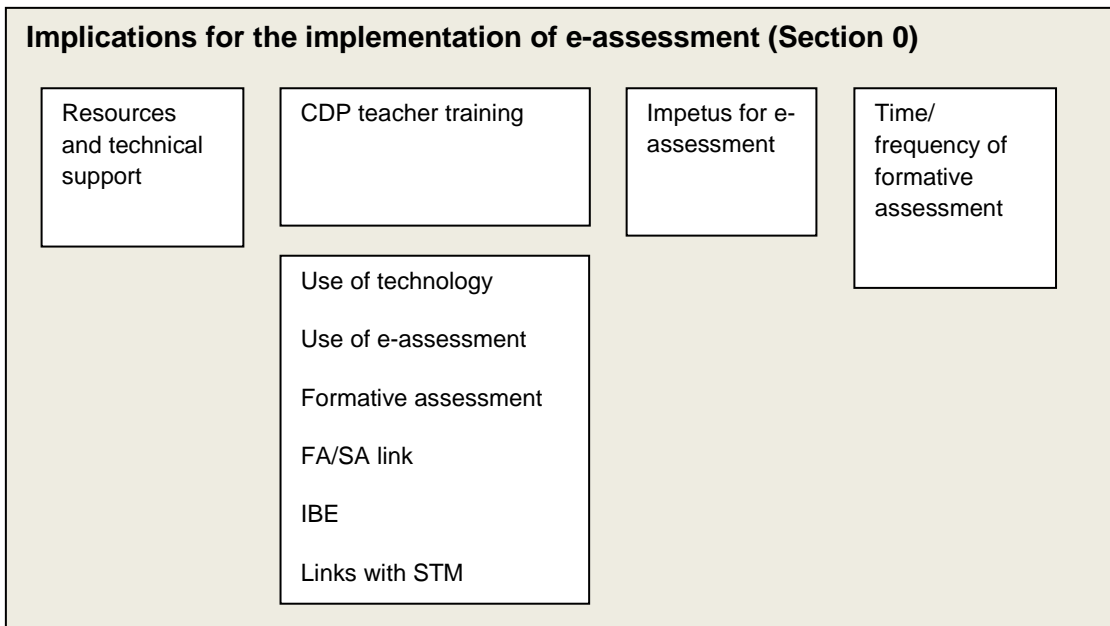
Learning
gains

Confidence

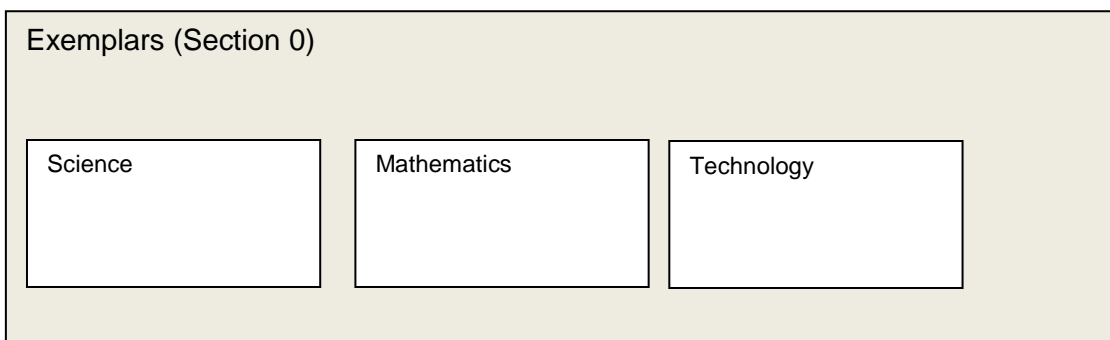
Quality (Section 3.10)

Value added paper to e-
assessment

Validity
Reliability



Implications for evaluation of assessments (Section 0)



The themes under the heading 'Aspects of e-assessment tasks/items' provide an overview of e-assessment and its component parts. This is given as a background and framework on which the reader can hang the detail contained in the rest of the findings.

3.1 Aspects of e-assessment tasks/items

An e-assessment (like a paper-based assessment) comprises a stimulus, the student response, some form of feedback and, potentially, an intervention. A description of these aspects, and examples taken from e-assessments identified in the literature follow.

3.1.1 E-Assessment type

Replication of paper test

Hughes et al. (2011) in a comparison of paper and onscreen mathematics assessment questions found no questions that were more accessible onscreen than on paper. It was predicted that questions in which the response mode was ordering or involved dragging and dropping objects would favour computer-delivery and performance would be higher onscreen, but this was not the case. Two particular question characteristics which were more accessible on paper were questions which required annotation and questions of a visual nature. It was also the case, although less strongly, that graphical questions and dense questions performed differently on paper and onscreen, with a bias towards paper. This is likely to be because annotating and working with diagrams is more difficult when working on screen, as there were no resources to allow students to interact with the visuals and diagrams as those working on paper could.

Gaming style assessments

Gee and Schaffer (2010) promoted the use of gaming for problem solving: 'Video games are good for learning because games can create virtual worlds where players solve simulations of real-world problems and in the process learn real-world skills, knowledge and values.' They added that the choices pupils make while problem solving can tell teachers a great deal about their ability to learn new material later on. Games involve continual diagnostic assessment of strengths and weaknesses in thinking, giving a portrait of problem solving decisions over time, so feedback to customize learning can be provided. They did highlight an interesting issue about the priorities when designing games compared to assessments. They observed that games are constructed in the opposite way to education: games consider first how to test and challenge a player and then design the learning, whereas in educational contexts we are more used to designing the learning and then the assessment to reflect the learning goals. Kennewell (2008) reported that the success of an e-assessment system may be influenced by the extent to which the software pre-specifies course activity. This suggests that a shift to the gaming approach, whereby the tools available and their affordances lead e-assessment content and processes, can lead to better e-assessments.

Games focus on problem solving with a mix of practice and guidance, complex concepts are introduced when needed and when a player's position in the game suggests that they would be most beneficial. Typically in gaming, players spend a lot of time on task, and are motivated because they are presented with a sequence of activities gradually increasing in difficulty, which means that players are constantly working at the edge of their abilities (Gee and Schaffer, 2010).

Johnson (2013) described an emerging field where computer game development and educational assessment are coming together, in which games are able to capture valid and reliable evidence. Both seek to engage learners by placing them in situations where they face challenges. Games then provide feedback in response to the choices made while pushing capabilities to the limit, rendering games similar to adaptive tests moving along different paths depending on skills. Good e-assessment extracts data not only from the results of test takers, but also from the processes they used to achieve those results (e.g. the use of gaming technology can provide evidence of what actions the learner takes during an activity, what they learn). As is good practice for developing any assessment, in whatever mode, developers should focus on the kinds of knowledge and skills they want the learners to display in an assessment.

Operation ARA is an example of a gaming style assessment (Halperna et al. 2012) and is described in section 0.

3.1.2 Stimulus type

Real life scenario

Koenig (2011) described Operation ARIES!, a scenario based e-learning tool for science. Real life examples are used to help students transfer what they have learnt in one context to another scenario-based assessment, and this was argued to be useful where there was a need to apply knowledge to practical situations. Koenig also described aspects of Operation ARIES! where students were expected to apply what they have learnt in the previous modules, for example, students were presented with inaccurate science information through the medium of newspaper headlines and television news channels and had to ask questions to ascertain the truth.

Halperna et al. (2012) furthered the work on Operation ARIES! (renamed Operation ARA), retaining the valued scenario-based assessments, in which students applied their understanding of scientific concepts to determine whether a described research case was reliable or flawed.

Simulation

Simulations enable students to interact with and control variables using technology. Simulations can function as independent learning tools, but are also valuable for assessment purposes (Neumann, 2010). Simulations can superimpose multiple representations and permit manipulation of structures and patterns that otherwise might not be visible, they can probe knowledge of how components of a system interact, as well as encourage learners to investigate the impact of varying multiple variables simultaneously (Quellmalz, 2009).

Examples of simulation use were found in all three STM subjects:

Science Clesham (BERA, 2009) developed and trialled secondary science, computer-based simulations to teach and assess scientific enquiry skills. Interactive investigations were developed by storyboarding investigative processes. It included experimentation using interactive simulators (modelling trialling and data collection) and questions involving manipulation of on-screen tools.

Pellegrino and Quellmalz's (2010) paper on 'SimScientists' an online environment for teaching and learning in science, illustrates ways that assessment tasks can take advantage of the benefits of simulations to represent generalisable, progressively complex models of science systems and such innovative items were included in the 2009 NAEP science administration.

Mathematics Neumann (2010) used a statistics simulation (followed by multiple choice questions) for summative assessment in HE.

Technology The new 2014 Technology and Engineering Literacy Framework for NAEP will be entirely computer administered and will include specifications for interactive, simulation-based tasks involving problem solving, communication, and collaboration related to technology and society, design and systems, and information communications technology (Pellegrino and Quellmalz 2010). In the UK, a national test in ICT for 14-year olds was piloted, which involved the simulation of a desktop with a suite of programmes and email software (Boyle 2006).

Interactive

Interactivity often occurs alongside the use of simulations in technology enhanced learning and assessment.

Neumann (2010) developed an interactive statistics simulation used for summative assessment. Neumann described how the affordances of the technology were exploited in the design of the tasks: 'The crucial aspect of each simulation was that it was interactive. Students were able to change data values, simulate events, and see what effects their changes had. It was this interactive nature that was exploited in the assessment approach.' (Neumann 2010)

Beevers et al. (2011) also valued the use of interactivity; through the CALM project lessons were learned including how to design assessments to give learners more autonomy through interactivity. Another example is Operation ARIES! which was designed to assess and teach critical thinking about science (Koenig 2011). It uses intelligent tutoring and makes use of animated characters. Students receive feedback and tutorage from two characters from the program throughout. Here the interactive element is how the learner is able to interact with virtual peers and tutors.

Video

Many e-assessments use multimedia stimuli followed by objective questions, for example, Operation ARIES! which aimed to engage students by using multimedia; it was designed to assess and teach critical thinking about science. It uses intelligent tutoring and makes use of animated characters. Students watch videos and receive communications through email and text message (Koenig, 2011).

Examples

Animation has also been used to improve examples given at the beginning of a test or set of items to show students how to answer questions. Direct observation in class suggests that many pupils rush straight into answering the questions in a worksheet without reading examples. Animating the examples makes them more interesting and

encourages learners to engage with them. In some e-assessments, engagement of the learner with the example answers is encouraged by the example remaining on screen for a set time, with the learner not being able to move on from that screen until this time is up. Animated examples may also hold the attention of the learner more than static ones (onscreen or on paper). Some systems now also ask students how well they have understood the example before proceeding on to the questions. In principle, their answers could inform future navigation.

3.1.3 Response type

MCQ and short answer question

There were concerns raised within the literature as to whether multiple choice questions (MCQs) are suitable for computer based formative assessment (e.g. Velan et al., 2008). Although Velan et al. found that MCQs used in formative assessments did have a positive impact of the learning of medical undergraduates, they recognised that this was a different outcome to most other research.

The literature described two key ways in which MCQs have been used productively in formative assessments:

1. MCQs or other objective questions have been presented to students after they have worked on a simulation or interactive stimulus.
2. Crisp and Ward (2008) used objective item types following a simulation of a real life situation (in the context of teacher training). Feedback gave the correct answer(s) with reasons and, where necessary, explained why the student's choice was not correct. E.g. within Operation ARIES! Koenig (2011) and the subsequent tool Operation ARA (Halperna et al. 2012) reported that during a phase of e-learning, learners read an e-book and after each chapter they were quizzed with multiple choice-type questions

MCQs can also provide useful outputs when responses, both correct and erroneous, are codified to allow the individualisation of feedback. After the MCQs are attempted in Operation ARIES! Students participate in dialog discussions with avatars where the understanding of the material from the chapter is clarified and reinforced.

Wylie and Dolan (2013) raised the issue that when using MCQs in technology enhanced formative assessment, the challenge is in supporting teachers to use outcomes to stimulate discussion and move understanding forward. Wylie and Dolan used MCQs specifically to identify any mathematical misconceptions held by secondary school students and output these to the teacher. Wylie suggested particular implications for how to construct the MCQs and distractors when aiming to identify commonly held misconceptions.

Extended response

It was rare to see examples of extended response. Crisp and Ward (2008) reported on the use of essay questions following the presentation of a real life scenario on screen. These extended responses would either 1) require marking by the tutor or 2) feedback was given by the tool by either providing model answers or the use of multiple-response question 'Which of the following points did you include?'

The use of extended response questions raises issues again about utilising the affordances of the technology; there are tools available that would automatically mark extended writing (e.g. Pearson's Intelligent Essay Assessor is used by classroom teachers as a learning aid. The software gives students immediate feedback to improve their writing, which they can revise and resubmit). Work by Shermis reported in the New York Times (2012) compared human graders and software designed to score student essays, Shermis reported virtually identical levels of accuracy, with the software in some cases proving to be more reliable. But nonetheless such automated marking systems still raise suspicions that key aspects of writing, including style and appreciation of poetry may be lost.

Interactive response

E-assessment provides opportunities for the *stimulus* of a task to be interactive as well as the *response* to be interactive. This section refers to the rarer cases of where the response, not just the stimulus, requires interaction between the learner and the technology. Hughes (2006) reported on a trial of computer mediated mathematics questions for 11 and 14-year olds. A question was considered to require an 'interactive' response if it had these features:

1. There was an animated element to the question which was more than just an illustration, but with which the pupils needed to engage by moving or controlling some part of it
2. The interaction with the question changed the appearance of the animation and so gave immediate visual feedback. This means that the interaction is two-way, i.e. the pupil interacts with the question and the software allows some response to the pupil, usually in the form of some visual feedback and
3. The pupil had some control over the animation or objects in the question.

For example, one question aimed to assess the understanding of properties of an isosceles, right-angled triangle. Pupils dragged one vertex of a triangle to make it right-angled and isosceles. In a question assessing understating of ratio learners chose what size to make a grid in order to show a given ratio of red squares to grey squares.

Hughes found that it was with the interactive questions that there was most evidence of the affordances of the technology affecting pupil behaviour.

The NRich website provides resources for school mathematicians www.nrich.ths.org; these include interactive tasks like those described in Hughes' work above which could be used for formative or self-assessment.

Implicit versus explicit feedback

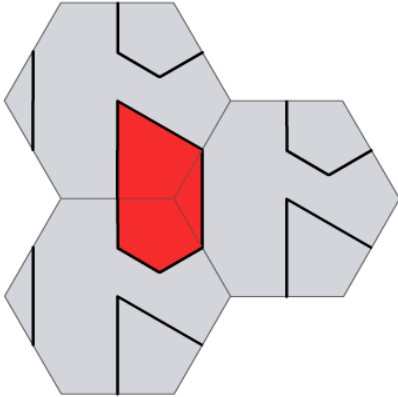
Questions which require an interactive response typically give implicit feedback. Boyle's 2006 paper highlighted a distinction between two types of feedback

Extrinsic feedback is given following the completion of an activity and states whether the attempt was right or wrong, this type of feedback requires that the technology apply some algorithm to translate the student's action or response into feedback, e.g. by presenting a score or the appearance of a 'shiny star'.

Intrinsic feedback, does not occur after the learner has submitted an answer, but is intrinsic to the act of working on the task. E.g. the question below (Hughes 2006) gave immediate intrinsic feedback in response to the student's action; the change in the red shape provided immediate visual feedback as to what shape the student had created. Then they would interpret this intrinsic feedback to decide if they had reached their goal of creating a pentagon with one line of symmetry. This type of intrinsic feedback allows the student to modify their behaviour and actions before submitting their answers, a strategy which is more akin with IBE than traditional approaches.

Rotate the tiles to make the red shape in the middle a pentagon with one line of symmetry.

Click on any tile to rotate it.



Similarly, the ARIES! Science online tool (Koenig 2011) provided implicit feedback as the learner received feedback that was not in the form of a score or reward; learners received implicit feedback as they communicate with a student avatar and as they are party to conversations between two animated characters.

In Boyle (2006) we saw warnings that when designing feedback we need to be wary that e-feedback can lead to what was called 'superficial behaviour'. For example, in the example above taken from Hughes 2006 disengaged students were observed to be carrying out what could be called 'mindless clicking'; they were continually clicking on the hexagons receiving some kind of visual 'reward' for clicking on interactive objects, though not engaging with the goal of the task. Cook and Crabb (2002) asked how computer-based learning could be designed to maximise cognitive engagement and stimulate thinking rather than what they called 'random button pressing'.

Kennewell (2008) reflected that when teachers first adopted ICT as part of their practice, there was a tendency for interactivity to be superficial and authoritative, which the hexagon question above could be described as. Kennewell argued that it was only when technology was embedded in teachers' pedagogical knowledge did the technology contribute to deeper, more dialogic interaction amongst students.

E-portfolio

Broadfoot et al. (2013a) described e-portfolios as an information repository, a personal development record which provides a structure for the organisation of learning and collaboration. An e-portfolio is a collection of digital objects showing evidence of a student's work. The technology provides an online system to manage the sharing of this work and to communicate feedback to students (Kimbell et al., 2009). When using e-portfolios, learners need personal online space for recording and evidencing attainment in e-portfolio

JISC (2007) proposed that e-portfolios were useful in promoting 21st C skills as when using e-portfolios, students are required to demonstrate skills of command of software, use web technologies and digital images, communicate electronically, solve problems and present and collaborate.

Other response types

Other response types seen or described in the literature included extended matching, repositioning objects by dragging and labelling diagrams.

Scratch pads

Onscreen scratch pads provide a means of catching learner's rough notes or sketches through touch screen technology. Learners are able to use either freehand 'writing' onscreen, and their device will capture the images and store them, or they can input via a keyboard. Hughes et al. (2011) reported on the development of a tool (at Pearson) named 'Override' which aimed to bring the experience of answering a mathematics questions onscreen closer to the familiar experience of working on paper. Students could, via the keyboard and mouse, make jottings and annotations onto objects on the screen and save notes. This tool also aided assessment by collecting information on the processes students used to answer questions. One concern raised in relation to tools like this is that they are translating paper tasks to screen for the sake of comparability of the experience, rather than recognising that technology offers many affordances which can enhance learners' experiences, rather than trying to replicate what is done on paper (Kennewell, 2001)

3.1.4 Feedback

A longer section relating to the range of types of and audiences for feedback appears in section 0, but here we want to highlight that the ease of provision of immediate and detailed feedback is one of the most valuable affordances of the technology (Kennewell 2001). This supports the learner and can then lead to the provision of hints and appropriate learning activities, including Integrated Learning Systems.

3.1.5 Intervention type

One benefit of using technology is that learners' responses can be analysed and codified to provide not only useful feedback, but pointers to the students about how to progress. These pointers can include interventions which link the assessment outcome to assessment, so truly supporting formative assessment.

Crisp and Ward (2008) provided learners with references to further reading and links to learning resources available on the web. Some other forms of help included the option

to read a short explanation of relevant ideas before moving on from the stimulus to the assessment questions. Other types and examples of interventions are described in section 0 below.

3.2 Assessment/teaching link

The link between assessment and teaching is key, and central to the process of formative assessment, and this also applies to the relationship between e-assessment and e-learning.

Bennett (2002) stated that technology would change assessments dramatically. He proposed that electronic test development would evolve through three generations.

1. He described the then current (2002) generation of CAAs as 'migrational', whereby on-screen test were simply a migration of existing paper tests onto the screen without reconceptualising the process or the content. This migration of tests, he argued, failed to 'realize the dramatic improvements that the innovations could allow' (Bennett, 2002). As such he argued that these first generation tests didn't utilise the functions of the technology to change the test for the better.
2. He predicted that the second generation of CAA would exploit the features of the technology, for example, by the use of colour, sound, animations, video and the integration of interaction between the test taker and the test.
3. Bennett predicted, finally, that 'Generation R' tests would evolve (the R standing for 'reinvent'). Generation R tests will be assessments so closely integrated into teaching and learning, that they will be indistinguishable from learning materials.

Formative assessment requires that the delineation of assessment and teaching is blurred, with assessment being a subset of teaching, suggesting that Bennett's 'Generation R' assessments would be akin to formative assessment.

Bennett's conceptualisation of Generation R assessments also brings up the issue of the compatibility of e-assessment with traditional teaching/learning methods; if technology is used to support teaching and learning, it follows that to ensure that an assessment is valid it also needs to be supported by the use of technology.

Kennewell (2008) recognised that some affordances of the technology relate to administrative rather than pedagogic concerns. However, in many CAA systems, Thomson Prometric reported in 2006, a complete electronic alternative to the existing assessment process is provided; the process can include the activities which can be more efficiently administered electronically than in paper-based systems:

1. Registering students and storing their details
2. Authoring questions
3. Pre-testing questions
4. Storing questions and the associated data
5. Delivering test to students
6. Delivering tests to markers
7. Marking
8. Storing completed papers
9. Storing outcomes
10. Converting outcomes into useful feedback (whether that be a pass/fail, mark, grade or report).

These may be benefits for administrators, but few would be recognised as benefits to learners. Learners benefit when assessment and teaching/learning are connected and relate to the same constructs. To be valid, an assessment must exploit the affordances that technology can bring to learners, not just administrators.

Neumann (2010) described the relationship between the use of technology in teaching and learning in Higher Education (HE) statistics as close. His work on using e-assessment with statistics students was possible because, in HE statistics courses, technology is central to teaching and learning, and hence can be more validly used in assessments and with less controversy than at lower levels of education.

Halperna et al. (2012) described the evidence-based design process behind a secondary school science e-learning programme with integrated assessment, 'Operation ARA' (previously Operation ARIES!). Developers identified good practice in e-learning and then integrated assessment into that programme.

Kennewell (2008) showed a specific focus on interactivity - his concern was for linking the concepts of interactive teaching and interactive technology, consequently championing *interactive* e-assessment. He argued that a shifting balance in the classroom towards dialogic would bring improvements to the learning process. His argument was that the nature of interactivity was more influential than the more general use of ICT, the latter of which could simply relate to administrative benefits of the technology. For example, JISC (2007) described how effective practice with e-assessment involves the linking of assessment and learning, with content available online via a learning platform and the contact time (lectures) between teachers and students being used to refine understanding rather than introduce a topic. This would require formative and/or diagnostic assessment to be used to identify what understanding to address during that contact time and at what level. In this example, the impact of technology on learning would be to increase the requirements for formative assessment.

3.2.1 Examples of interventions

Intervention is central to formative assessment; teachers use outcomes of formative assessment to select or devise interventions for learners to further their learning. Technology allows the process of translation of student response into an appropriate intervention to be speedy and based on evidence and good pedagogic principles. Broadfoot (2013d) defined learning analytics (further described in section 0) as the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs. Learning analytics are generally algorithms used by the technology resulting in decisions or suggestions about which type of interventions should be adopted.

Intelligent tutoring

Intelligent tutoring can refer to a huge variety of types of e-learning, including a simple clue or hint, the addition of scaffolding, feedback specific to particular responses, links to learning resources including animations, references to further reading or interactions with virtual or real tutors or peers.

Crisp and Ward (2008) incorporated in their feedback to students, references to further reading and links to learning resources available on the web. Other forms of help included the option to read a short explanation of relevant ideas before moving on.

Feng et al. (in Pellegrino and Quellmalz 2010) described the 'ASSISTment' system, which is a pseudo-tutor for middle school level mathematics. The system uses scaffolding questions, optional hints, and buggy messages (specific feedback given after student errors) for each item. Students must eventually reach the correct answer, and scaffolds/hints are limited to avoid giving away the answer.

Halperna et al. (2012) Operation ARA (previously ARIES!) described how students move through the interactive chapters of the system, they receive computer-generated tutoring that varies depending on how well the student responds. The type of tutoring that students receive following each chapter is determined by the number of questions they answer correctly about the chapter they have read.

Implications for the design of these types of interventions are that e-tutors need to both correctly gauge and adapt to the student's current level of understanding. A successful adaptive tutor chooses problems that specifically address the level of the student's prior knowledge and take previous test scores into consideration. In order to maintain engagement during vicarious learning (where the learner is party to a discussion between a virtual student and a virtual tutor or two virtual students) the learner is asked to respond to questions about the tutoring situation. For example, the virtual teacher might ask the human student whether the virtual student understands the concept or whether the virtual student's answer was correct.

Computer Adaptive Testing (CAT)

Quellmalz (2009) defined CAT as procedures in which items are selected based on the examinee's prior response history and an underlying model of proficiency, developed to reduce time testing and examinee burden. Such adaptive testing can integrate diagnosis of errors with student and teacher feedback. It can tie assessment more closely to process and contexts of learning and instruction.

In its simplest form CAT automates decisions about which questions to present in a sequence of questions as determined by the performance on the previous question. The complexity of the algorithm used to analyse previous performance can vary considerably.

DreamBox Learning' (available at www.dreambox.com) differentiates instruction and adjusts difficulty levels and the number and type of hints to give learners, based on tracking pupils' responses to several different questions. Many sources of evidence from many tasks relating to both processes and content can be used to input to decisions about what task to present to a student next. Quellmalz (2009) described how the computer's ability to capture pupils' inputs permits collection of evidence of problem-solving sequences and strategy use as reflected by information selected, numbers of attempts and time allocation. These can be combined using statistical and measurement algorithms for patterns associated with varying levels of expertise and then students can be directed as appropriate. This information can also be relevant for assessing against competencies.

Consideration needs to be made of which data it is useful to capture for the learner and for the teacher. The effective use of CAT requires reconceptualising assessment design and its use involves tying assessment more directly to the processes and contexts of learning and instruction.

Two concerns relating to CAT arose from the literature:

1. Use of item banks to randomly select questions to present to learners was seen by students as unfair (Voelkel, 2013). Voelkel found that different variants of computer marked questions could behave differently in terms of their level of demand and the construct being assessed.
2. Pachler et al. (2009) found widely differing theoretical emphases being applied in e-assessment development in the literature, as well as varying views of 'adaptivity' as a core component of e-assessment processes.

3.2.2 Re-submitting issues

Whether students are permitted to change their answers in an e-assessment usually depends on whether the assessment is high stakes or low stakes and whether it is formative or summative.

Neumann (2010) reported on a statistics simulation used for summative assessment in HE. Feedback was minimal because of the summative purpose of the assessment and the fact that students could attempt questions more than once. Voelkel (2013) described how feedback was tailored by providing more and different feedback after second or multiple attempts at the same questions. Scholar (www.scholar.hw.ac.uk)

used a method of being able to resubmit an answer before the correct answer is revealed, this allows a more iterative process to understanding. A number of repeated demonstrations, for example, in mathematics, allow different pupils to vary the amount of reinforcement received depending on their confidence and/or understanding.

This section brings up implications for giving feedback (or not) between attempts at a task. In high stakes assessments this can still provide useful feedback to the teacher, even if this is not shared with the student.

3.3 Inquiry based education

The ASSIST-ME proposal described Inquiry-based education (IBE) as an umbrella term, encompassing a wide range of teaching approaches that can enhance student motivation which have the potential for enhancing learning outcomes.

Inquiry-based STM education includes students' involvement in questioning, reasoning, searching for relevant documents, observing, conjecturing, data gathering and interpreting, investigative practical work and collaborative discussions, and working with problems from and applicable to real-life contexts (Anderson, 2002). Inquiry-based STM-education is not a new teaching method, but it is often used as a contrast to more traditional teaching approaches, such as those where the teacher presents results and methods which the students are then trained to apply. Giving students an active part in learning is in accordance with many teachers seeing the pedagogical principles of constructivism as the foundation for understanding and implementing inquiry-based learning (Llewellyn, 2007).

Only a few of the articles found here explicitly reported on the assessment of IBE.

Neumann (2010) assessed functional knowledge using MCQ questions following a simulation relevant to HE statistics and Feldman and Capobianco (2008) reported on the use of an electronic voting system in which assessment items were designed to be 'consistent with constructivist and active-learning pedagogies'. Feldman and Capobianco highlighted that, for teachers to successfully implement a conceptual learning approach (i.e. IBL) to physics teaching, they may need to make significant changes to their teaching methods. Scholar (www.scholar.hw.ac.uk) a secondary science and mathematics environment, including simulations, utilises animated graphs which could be used as stimuli for IBE.

Of the examples of assessment of IBE that we found we have selected examples of e-learning environments that were impressive in their scope and methods, for example in science Operation ARIES! (Koenig 2011) (which was subsequently acquired by Pearson and renamed Operation ARA) and SimScientists (Pellegrino and Quellmalz, 2010) were both e-learning environments which integrated the learning and assessment of IBE.

Operation ARIES! (Koenig 2011) includes an assessment model in which learners are presented with inaccurate science information through the medium of newspaper headlines and television news channels and must ask questions to ascertain the truth. Students engage in solving a problem through dialogue interactivity whereby students

learn by engaging in conversations and tutor groups. Halperna (2012) reported that students repeatedly practiced and applied concepts in different contexts and from different domains within science, and argued that this variability enhanced transfer of knowledge and skills.

Halperna (2012) also discussed relating to competencies and communication, students answer a number of multiple choice questions at the end of each chapter of an e-book and then participate in dialog discussions with avatars where the understanding of the material from the chapter is clarified and reinforced.

Pellegrino and Quellmalz's (2010) work on the SimScientists software illustrated ways that assessment tasks can take advantage of the affordances of simulations to represent challenging inquiry tasks, indeed many e-learning or e-assessment tools, which have claimed or attempted to assess IBE, have used simulation: in order to assess IBE there must be observable evidence of IBE, so assessments must provide opportunities for learners to engage in IBE. Pellegrino and Quellmalz (2010) described two national projects in which the use of interactive simulation tasks enabled the assessment of IBE:

- The 2009 NAEP science framework and specifications drew upon science simulations work (reported in Wylie and Dolan, 2013) in developing their rationale for the design and pilot testing of interactive computer tasks to test students' ability to engage in inquiry practices. Such innovative items were included in the 2009 NAEP science administration.
- Wylie and Dolan (2013) also reported that the new 2014 Technology and Engineering Literacy Framework for NAEP will be entirely computer administered and will include specifications for interactive, simulation-based tasks involving problem solving, communication, and collaboration related to technology and society, design and systems, and information communications technology.

Similarly, the 2015 PISA framework for scientific literacy (OECD, 2013, available at www.oecd.org/pisa) considers the possibility of assessing collaborative science problem solving skills by computer for science, in a summative high stakes assessment.

3.4 Competency-based learning

The ASSIST-ME proposal understands competence to mean a combination of skills, knowledge, characteristics, and traits that contribute to performances in particular domains. There is not a universal agreement on the terminology of competence. In this project we will use the word competence for both a *competence*, referring to the concept in general and a level of ability, and a *competency*, referring to a particular demand that a person may or may not be able to meet, and the plural form competences, to reflect an integration of understanding and attitude into the concept.

The more complex the learning goals, the more difficult they are to measure. The understanding of competences as the ability to cope with relatively complex challenges

in everyday life means that assessment methods necessarily have to be relatively advanced, flexible and process oriented. This suggests that in order to assess IBE related competencies, there is a need for learners to engage in IBE as part of their assessment.

Operation ARIES! (Koenig, 2011) is designed to assess and teach critical thinking about science, a core competency. It uses intelligent tutoring and makes use of animated characters to enable students to develop and exhibit relevant competencies, including inter personal skills. The way in which Operation ARIES! requires that students generate their own questions about abbreviated research descriptions in order to determine whether the research is flawed in also in line with CBL.

3.5 Formative assessment

The ASSIST-ME proposal describes formative and summative assessments as similar in that they involve the collection, interpretation and use of data for some purpose. They are mainly identified and distinguished from each other by the purpose of the assessment but often also in the way data is collected. Formative assessment has the *purpose* of assisting learning and for that reason is also called ‘assessment *for* learning’. It involves *processes* of “seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning and where they need to go and how best to get there” (Assessment Reform Group, 2002).

Under this theme of formative assessment, some key areas were identified in the literature which have relevance for the design of ASSIST-ME e-assessments:

- Self-assessment, incorporating confidence ratings
- Peer assessment, relating to both the assessment of collaborative and individual work
- Diagnostic assessment, including in relation to the diagnosis of misconceptions
- The iterative, personalised nature of some adaptive assessments and
- The interaction between and roles of the teacher, learner and computer, including reference to how e-assessment can support discussion.

Note the omission of feedback in the above list: this is an extremely important aspect of formative assessment. Whitelock (2006) identified one driver for the implementation of e-assessment being improving learning through faster feedback which he related to increases in student retention, flexibility, support in coping with large student numbers, providing objectivity in marking, the more effective use of VLEs, and more reflective learners who are more in control of their learning. Feedback is so important that rather than make it a subheading of formative assessment, it is discussed in section 0 which includes all feedback-related findings.

3.5.1 Self-assessment

Pachler et al. (2009) wrote that learner self-regulation was a core feature in assessment and is linked to motivation and emotional factors which affect learners' engagement with feedback.

Supporting students in judging their own learning or performance can help develop the skills of self-regulation. Pellegrino and Quellmalz (2010), reporting on the interactive game SimScientists, illustrated ways that assessment tasks could take advantage of the affordances of simulations to represent generalisable, progressively complex models of science systems which promote metacognitive skills through self-assessment.

Confidence ratings

A number of assessments found in the literature required that learners complete confidence ratings, in which they rate how confident they are in their answer.

JISC (2007) proposed that confidence based marking could promote a deeper level of learning by challenging learners to evaluate certainty in their answers so that they could address gaps that they discovered in their knowledge. A learner's confidence is affected by their self-efficacy, which is one's belief on one's ability to succeed in specific situations. One's sense of self efficacy can play a major role in how one approaches tasks, however the danger is, while students with a strong sense of efficacy are more likely to challenge themselves with difficult tasks and be intrinsically motivated, students with low self-efficacy believe that they cannot be successful and are thus less likely to make an extended effort and may consider challenging tasks as threats to be avoided.

Crisp and Ward (2008) also captured confidence ratings by asking users to indicate their confidence after answering each automatically scored question. At the end of an assessment an analysis of the student's metacognition was reported, based on the given ratings for use formative assessment. The ratings described by Crisp and Ward (2008) didn't influence a student's path through an assessment or the immediate feedback given, whereas Swithenby (2006) reported on how pupils were given options depending on how they rated their confidence in their answer at the point when they submitted their answer. These options included: submit; hints; show answer; review part; display mathematics; and give clues.

3.5.2 Peer assessment

Broadfoot et al. (2013c) stated that successful peer assessment required individual responsibility from students, interdependence on peers, and trust within the group. Practitioners should recognise that students can be anxious about the ability of their peers to assess learning, their own abilities to assess others' work and the overall validity of peer assessment.

Examples of peer assessment in the literature included

- Group peer assessment – in which each member of a group who had collaborated on a task judged each other's contributions and
- Individual peer assessment – in which a piece of work/performance/response carried out by one student was evaluated by one or more peers.

Group peer assessment

Electronic voting systems can be used to gather group peer assessments, for example, used an electronic voting system to evaluate previous students' practical work against specified marking criteria.

Digital technologies have the potential to support collaborative learning and assessment practices, such as undertaking knowledge building activities, co-evaluation and social interaction. A case study (Broadfoot et al., 2013c) using computer-supported collaborative learning (CSCL) limited the participation of students in some assignments and generally low-quality assessment reports.

Kennewell (2008) reported on software for group work incorporating features to capture contributions from different students. An example activity was to develop a concept map for photosynthesis.

Broadfoot et al. (2013b) described 'crowd-sourced grading' involving weekly peer evaluation of student blogs. McKinsey and Company (2013) described a process of evolving better answers through collaboration: Students were set assignments to write blogs for sharing test results of their designs and receiving comments from professors and classmates. Some argue (in Broadfoot et al. 2013c) that using tools like wikis or blogs in group assessments can further exclude some students by benefitting those who are already users of social media.

Individual peer assessment

A simulation of individual peer assessment is found within the Operation ARIES! science environment (Koenig 2011) where students receive feedback and tutorage from two characters from the program throughout. Two characters have conversations with each other in the presence of the learner, using virtual peer tutoring.

3.5.3 Diagnostic Assessment

The availability of measurable, detailed descriptions of the constructs and factors to be assessed is the essential prerequisite for the construction of diagnostic items, as well as the tests. Developing an online diagnostic assessment system for grades 1 to 6 CRLI (2009) described constructs being assessed in terms of misconceptions that students may have about the subject.

Misconceptions

We found examples of how technology can provide evidence of learners' misconception(s) and hence, use this to enable them to progress (e.g. Wylie and Dolan, 2013). Feedback to teachers can aid clarification of which misconceptions are held by students (Voelkel, 2013).

Furse (2009) described it as straightforward to incorporate misconception handling into an e-assessment system if the author knows of suitable misconceptions. Wylie & Dolan (2013) reported on the creation of a bank of items for High School mathematics and science teachers that drew on the misconception literature (Wylie and Ciofalo, 2008). These high school formative e-assessments were through multiple choice questions and each multiple-choice item that was developed drew on at least one previously identified student misconception. These question types have different implications for item development than those which are not misconception based (Wylie & Dolan 2013).

Progression levels

Also reported in Wylie and Dolan (2013) was a middle school mathematics project which focussed on the progression between levels of understanding, rather than just on categorising the learner according to in which level they sat. The project used two kinds of assessments:

- *locator* assessment: computer delivered and places student within three learning progressions and
- *incremental* tasks: which explicitly target a transition between levels, rather than the levels themselves.

3.5.4 Teacher/learner/computer relationship

Building on research principles of the Assessment Reform Group which firmly put an emphasis on Assessment for Learning and the relationship between the teacher and the student, the e-assessment association of the UK (Beevers et al., 2011) believe that software solutions designed for formative assessment should also follow the ten ARG principles:

1. be part of the effective planning of teaching and learning
2. focus on how students learn
3. be able to be central to classroom practice
4. promote professional skills for teachers
5. be sensitive and constructive, being aware of emotional impact
6. foster learner motivation
7. promote commitment to learning goals and assessment criteria
8. help learners to know how to improve
9. develop the learner's capacity for self-assessment and
10. recognise a range of educational achievement.

Feldman & Capobianco (2008) used whole class responses using an electronic voting system to prompt discussion where they reminded us that formative assessment was supported by classroom discussion and the technology in their study-aided classroom discussion.

To facilitate one to one discussions between learners and teachers there needs to be focus on the relationship between the teacher and the learner; Wylie and Dolan (2013)

placed this relationship at the centre of formative assessment. They warned that the use of external 'off the shelf' tools could distance this relationship: 'To carry out formative assessment, teachers must be proficient in developing their own evaluative tools as part of their instructional practice.' (Wylie and Dolan 2013 p1)

McKinsey and Company (2013) extended these relationships to include parents giving the example of Ultranet, a student-centred learning environment which allows students, teachers and parents to connect and collaborate to improve learning outcomes.

Prompts for discussion

Wylie and Dolan (2013) viewed the output of some formative assessment as a stimulus for teacher-led discussion. This requires that items are written to stimulate discussion, not to summatively assess. Discussion could be one to one, whole class, small groups etc. They stress that the tool collects the evidence and the responsibility is with the teacher to use the evidence appropriately.

The scenario-based assessments described by Crisp and Ward (2008) could be used to stimulate discussion amongst groups of learners but were also intended to be usable without the guidance of a facilitator. In the assessments, the computer provided some of the guidance, probing and directing mimicking the role that a discussion leader would normally provide in case methods.

3.6 Summative Assessment

Summative assessment has the purpose of summarising and reporting learning at a particular time and for that reason is also called 'assessment of learning'. It involves processes of summing up by reviewing learning over a period of time or checking-up by testing learning at a particular time.

The SimScientists game provides teachers with feedback on student and class progress both on general summative measures (e.g., time to completion, percentage correct) and on more specific knowledge components (Pellegrino and Quellmalz 2010). Neumann (2010) described using a simulation tool for summative assessment of statistics.

Concerns about using e-assessment for high stakes summative purposes include collusion, plagiarism, recognition of partial achievement, logistical problems of simultaneously allowing access to computers for a whole class (Swithenby, 2006) and security concerns (Dennick 2009). Summative assessment of collaborative work had added complications; Swithenby (2006) claimed that for group work it was easy to monitor the amount of time pupils spent on an activity or contribution, but difficult to judge the quality of it. One proposed solution was to use more peer assessment.

To mitigate against the risks of using e-assessment for summative purposes, Swithenby reported on how Warburton (2006) suggested a gradual, low risk strategy through quizzes and progress checks leading to one summative assessment (high stakes). Another means of using formative and summative assessments in the same platform was the repeated use of frequent formative assessments followed by a final summative assessment.

3.7 Formative / summative assessment link

There are a number of affordances offered by the technology (Kennewell, 2001) some of which relate to administration (e.g. reducing postal traffic, speeding up marking, etc.) and some to pedagogy (e.g. interactivity, simulation, giving students control, providing environments/experiences not possible on paper, giving detailed and immediate feedback). *Administrative benefits* of e-assessment have been utilised for high stakes summative assessments (e.g. the electronic marking of examination scripts), but there has been less take-up of the *pedagogic* affordances of technology for summative assessment.

As Boyle (2006) pointed out, there is a conflict when considering the use of e-assessment for high stakes summative assessment, because a high stakes environment is not conducive to innovation or risk taking. But he predicted that formative assessment would be the vehicle for innovation as there are more opportunities for risk taking in a formative assessment context.

3.7.1 E-assessment as a support for linking formative and summative assessment

E-assessment can support the link between formative and summative assessment. Pachler et al. (2009) reported that within e-assessment there is a tendency to conflate formative and summative assessment. Virtual learning environments bring together learning and assessment, and consequently formative and summative assessment. Bennett's 2002 vision of 'Generation R' e-assessment included assessments that are so closely integrated into teaching and learning, that they will be indistinguishable from learning materials. Broadfoot et al. (2013a) described how making assessment and instruction simultaneous would support the integration of formative and summative assessment.

Broadfoot et al. (2013b) proposed that integrating formative and summative assessment would be more meaningful for students; using an integrated assessment, learners could benefit from regular feedback which supports learning. They argued that the link could also contribute to an overall picture of learning, is more authentic, has the potential to track progress, aggregate data, create multi-media platforms for feedback and review, accumulate evidence and help learners understand the connections between learning and assessment.

3.7.2 Digital objects for both formative and summative assessment

There are many electronic tasks which can be effective as learning objects for formative assessment or for summative assessment. Neumann (2010) used simulations that were designed for use in summative assessment, but would be appropriate for use in formative assessment or for teaching and learning.

The evaluation of the e-portfolio system E-scape (Kimbell et al., 2009) showed that the e-portfolio can be used for formative assessment as all student activity is recorded and for summative assessment in which case students can edit and select which work to submit for judgement. This supported the argument that it is what is done with the

information that determines whether it is formative or summative assessment, and the same tasks could be used for either.

But some would argue that formative and summative assessments are different, e.g. Voelkel (2012) argued for the separation of summative and formative processes, based on a view that the use of the same assessment for formative and for summative purposes is not always beneficial for learning. Perhaps this depends on the type and quality of the assessment. Some examples of formative e-assessment can be argued to be serial summative assessment (e.g. Pachler et al. 2009). Formative assessment appears to be equated with 'low stakes' assessment, or 'practice' assessment in preparation or contributing towards high stakes summative outcomes. This does not necessarily reflect the principles of formative assessment as set out by the Assessment Reform Group (shown in section 0).

SimScientists is another example of an environment which uses assessments for both formative and summative purposes. The SimScientists (Quellmalz et al., 2012) includes assessments designed to supplement state science test evidence by providing science assessments that are

1. embedded within curriculum units that could serve formative assessment purposes by providing immediate feedback, monitoring progress, and informing needed adjustments to instruction and
2. administered at the end of a unit as summative measures of proficiency on the targeted science content and inquiry practices.

3.7.3 Games as a means of blurring the formative /summative boundaries

Games use actual learning as their basis for assessment: their assessments are built on problem solving and facing challenges (Gee and Schaffer 2010). Games:

- assess whether a player is ready for future challenge
- track information over time
- are designed in levels, and each level requires that students have mastered the previous level and that they learn new skills on the new level.

These three characteristics are compatible with both formative and summative assessment.

3.8 Feedback

One advantage of using e-assessment is the ability of the technology to provide quick and detailed feedback (Kennewell, 2003). Feedback supports the learner and can then lead to the provision of hints and appropriate learning activities.

It is clear from many sources that feedback is most effective when it is instant, differentiated and individualised (e.g. Swithenby, 2006, Pellegrino and Quellmalz, 2010).

A number of papers (e.g. Neumann, 2010, CALM (in Beevers 2011), Scholar, www.scholar.hw.ac.uk, and Voelkel, 2013) cited the impetus for using e-assessment as the opportunities that it brought for giving meaningful feedback to large numbers of students (which may be an issue particular to HE).

Boyle (2006) advised that in designing feedback we need to be aware of the audience for the feedback. He raised the questions: Is the audience the teacher and/or the student; Will one type of feedback be appropriate for both audiences?

The discussion of issues arising relating to feedback is most usefully structured by audience: section 0 describes what was found in relation to feedback that is provided to the teacher and section 0 describes findings in relation to feedback as it is given to learners.

3.8.1 Output to teacher

Key decisions need to be made in the assessment design process relating to what feedback the teacher would benefit from receiving. E-assessment generates rapid, reliable data on learners' progress and can indicate which learners are at risk and provide prompts for remedial action (JISC 2007). Through the CALM project, for example, (Beevers et al. 2011) lessons were learned including which details to record for the reporting process.

With such a variety and depth of data available, these decisions are not simple. Data goes way beyond just scores and performance data and can relate to processes and contexts in which assessments were tackled, for example the ALTA system (Adaptive Learning Teaching Assessment) (2009) includes the possibility to collect both pupil and class data and to collect longitudinal data and trends.

Learning analytics

Learning analytics use data about learners to optimise learning. Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs. It interrogates learner-based data interactions; techniques include predictive modelling, user profiling, adaptive learning and social network analysis (Broadfoot et al., 2013d).

Learning analytics can be used as a precursor to intelligent tutoring (Section 0) during which data is analysed to establish which type of learning or support should be tackled next. For example, McKinsey (2013) describes how through using mobile technology for learning and assessment, learning can be adapted to individual pupil's understanding and pace. This recognises the need to analyse patterns of pupil behaviour (not just performance) over a number of assessments. Unlike summative assessments which provide a snapshot of performance level, this enhanced type of feedback can be generated using learning analytics from numerous assessments over time.

Feedback level

Reports to teachers can be provided a number of levels including, for example, individual learner, group, class, year, whole school or cohort. Boyle (2006) concluded that formative reports should be threefold: one for the student audience, one for the teacher audience reporting individual student feedback and a third for teachers reporting on the whole class or a group of students.

Teacher reports are commonly presented via a dashboard (Broadfoot et al. 2013d) in which teachers can switch between individual, class or year group views. The E-scape portfolio system (Kimbell et al., 2009) used a timeline onto which all digital objects and student work were placed. This allowed both the teacher and student to see a process of learning and thinking, which served as an effective reflective tool for the student and evidence of process for the teacher making assessment judgements.

3.8.2 Feedback to student

Gipps (2005 in Crisp and Ward, 2008) described the facilitation of formative feedback to learners as one of the key reasons for the growing use of e-assessment in UK higher education, but stressed that only assessments providing adequate feedback would enhance learning. Gipps commented that 'the developments in automated, diagnostic feedback in short answer and multiple-choice tests are . . . potentially very valuable. If feedback from assessment could be automated, while maintaining quality in assessment, it could certainly be a powerful learning tool' (p. 175).

While dashboards can be used for teacher feedback, Broadfoot (2013d) also described how dashboard systems allow learners to monitor their own academic or behavioural activity, to access relevant strategies and support and to compare their performance to previous students/classes. However there are a number of challenges and ethical debates and concerns (e.g. demotivation of less able) to be considered here.

Gibbs and Dunbar Goddet (2007 in Voelkel, 2013) found that 'giving out clear goals and standards had little effect on learning, and that it was much more helpful when students received plenty of feedback' (Voelkel, 2013).

Characteristics of feedback

Shute (2008, in van der Kleij et al. 2012) suggested making a distinction between feedback *type* and feedback *timing*. Students value high quality, actionable feedback (Black and Wiliam, 2009).

The themes of type, timing and level of detail of feedback are developed below, along with some findings relating to how students action the feedback that they receive.

Type of feedback

Hattie and Timperley (2007 in van der Kleij et al. 2012), in their analysis of feedback in computer-based assessment for learning, distinguished four *levels* at which feedback could be aimed, which is an expansion of a previously developed model by Kluger and DeNisi (1996, in van der Kleij et al., 2012). The levels distinguished are the self, task, process, and regulation levels.

- Feedback at the *self* level is not related to the task performed but is aimed at characteristics of the learner. Praise is an example of feedback at the self level. Feedback at the self level is not seen as effective for learning because it does not provide the student with information regarding how to achieve the intended learning goals.
- Feedback at the *task* level is mainly intended to correct work and is focussed at a surface level of learning (e.g. knowledge or recognition); for example, the student is told whether the answer is correct or incorrect.
- Feedback at the *process* level relates to the process that was followed in order to finish the task. In this case, for example, a worked-out example is given.
- Feedback at the *regulation* level is related to processes in the mind of the learner, like self-assessment and willingness to receive feedback. In the ideal situation, the feedback is adapted to the current level of the learner.

Hattie and Timperley favoured feedback aimed at the process or regulation level in order to enhance learning. Van de Kleij et al. (2012) gave clear indications as to the type of feedback that learners perceived to be most useful for learning; learners' opinions gathered via questionnaires indicated that learners perceived immediate and delayed feedback to be more useful for learning than delayed knowledge of results only.

Timing of feedback

The timings with which the feedback from e-assessments is delivered to learners can vary and this variation can have an important effect on how useful it is to the learners. Boyle (2006) discussed how e-assessment designers need to be wary that feedback is given at the right time whether this be at the end of a question, a series of questions or a session.

Others differentiate differently timed feedback as 'immediate feedback' being feedback given immediately after completion of an item and 'delayed feedback' as feedback given directly after completion of all of the items in an assessment. Shute (2008 in Van der Kleij et al. 2012) also attempted to distinguish immediate and delayed feedback by claiming that immediate feedback is (usually) provided immediately after answering each item while the definition of 'delayed' is more difficult to make, since the degree of delay can vary. In some cases, the feedback is delayed until a block of items has been completed. Delayed feedback could also mean feedback being provided after the student has completed the entire assessment. However, feedback can be provided an entire day after completion of the assessment or even later. The nature of the assessment in terms of its summative or formative nature could affect which is more appropriate.

Van der Kleij et al. (2012) suggested that learners prefer immediate feedback to delayed feedback. A game, quiz or a simulation can give instant feedback. In these situations, immediate feedback is important as the task is acting as formative or self-assessment, for example, with the Scholar programme. Voelkel (2013) looked at science e-assessment in HE. Feedback was not given directly after a whole test, rather it was given at a lecture a week later; less than 60% of students reported that they

found this delayed feedback useful. Beevers et al. (2011) also highlighted the immediacy of feedback as an ingredient for successful formative e-assessment.

Frequency of feedback

Related to the timing of the feedback is the frequency of the feedback. Quellmalz et al. (2012) highlighted the comparisons between frequency and ownership of formative assessment, whereby formative assessment went beyond annual, high-stakes tests to multiple assessments over time and in time for teachers to tailor instruction.

Halperna et al. (2012) commenting on Operation ARIES! described how, as learners progressed through, their understanding of the concepts and principles of science were repeatedly tested. This required that learners demonstrated their learning consistently throughout the programme. Through this active engagement it was felt that more deep learning was achieved as the learners were actively engaged with the material. Through this model, Halperna et al. (2012) felt that the learners could become judges of their own performance and rely less and less on external knowledge of results than they would with constant feedback. Using this principle in the design of Operation ARIES!, as learners progressed through the programme, they received feedback that was increasingly less frequent and less detailed.

Level of detail of feedback

Scores alone do not provide the necessary information for learners to use them as effective feedback (Gipps, 2005 in Voelkel, 2013). Some believe that, for learners, knowing which answers were correct is just as important as knowing which answers were incorrect (Hattie and Timperley, 2007 in Voelkel, 2013). JISC (2007) reported how online mock tests were useful when immediate feedback was given for *correct and incorrect* answers.

However, generally it is agreed that while assessments that provide grades and scores tend to increase the tendency for learners to adopt performance, rather than mastery goals, these grades and scores can increase motivation in the short term. In the longer term, however, the effect appears to be detrimental to formative processes and to learning (Black and Wiliam, 1998 in Pachler et al., 2009). So for e-assessments, it should be that more detailed feedback than scores alone are provided for the learner.

Hattie and Timperley (2007) in Voelkel, S. (2013) emphasised that effective feedback needed to provide information that specifically related to the task, so that learners could develop self-regulation and error detection strategies and use the feedback to then tackle more challenging tasks. Furse (2009) felt that learners needed to have an explanation of the answer. It was not sufficient just to be told the correct answer especially as 'rushers' were likely to just want to get onto the next question rather than digest the answer. One solution was to leave the correct answer and its explanation up for 5 seconds before learners could proceed. A more sophisticated approach would be to use animated text. It is probably desirable for an e-system to also check whether the learners think they understand their mistakes.

Learners' use of feedback

This section arose from literature which showed a concern for how students use feedback. Questions were raised such as: Do learners use feedback?; What is the value of feedback and how can that be optimised?; How can we encourage students to use feedback effectively? What support or advice do they need to use feedback optimally?

Halperna et al. (2012) suggested that feedback was important in that it provided information to the learner about his or her own performance, but that the learner still had to derive meaning from it. It may be that the way learners interpret feedback is what determines when it will be beneficial. Beevers et al. (2011) reported that feedback from e-assessment encouraged learners to take responsibility for their own learning. This opportunity for learners to use their feedback for ipsative assessment essentially encouraged learners to become proactive self-critical learners rather than just using feedback for normative processes.

One key difference with the immediate feedback offered by e-assessments is that learners can act on it 'there and then'. Swithenby (2006), and Jordan (2009, in Voelkel 2013) found that interactive computer-assisted assessments allowed learners multiple attempts and with built-in feedback could engage learners in meaningful learning activities, as they were asked to act on it immediately.

3.9 Effects of e-assessment on the learner

This section discusses the impact that assessment can have on the learner. JISC (2007), Whitelock (2006) and Broadfoot et al. (2013) described some of the potential, positive effects of e-assessment on the learner:

- increasing the range of what is tested
- encouraging deeper learning
- fostering more effective learning for a wider diversity of learners
- presenting challenging yet stimulating ways to demonstrate understanding and skills
- more authentic experiences being offered, for example, through using simulations
- good quality, timely feedback
- linking to appropriate resources
- feedback including opportunities for further learning
- supporting personalisation: learners can progress at a pace and in a way appropriate to them, for example, e-portfolios helping learners to present themselves and their work in a more personalised manner
- allowing learners to realise their own potential
- on-demand summative assessments increasing motivation
- learners taking tests voluntarily if they are available anytime, anywhere which can in turn help to establish more regular patterns of study; learners have been more likely to test themselves more regularly than with pen and paper tests.

In the literature reviewed, three key effects that e-assessment could have on the learner were identified: motivation, learning gains and learner confidence.

3.9.1 Motivation

E-scape (Kimbell et al., 2009) showed how technology supported learning in itself was a significant motivator for 14-19 year olds and was in fact one of key tools in accomplishing higher levels of engagement and achievement with this age range. Broadfoot et al. (2013a) found that the tools that could be used to support e-assessment, for example, wikis, blogs, social networking activities, podcasting and e-portfolios, provided richer activities that lead to improved learner engagement.

Crisp & Ward (2008) found evidence that e-assessment increased motivation and performance. They also felt that we should make use of the motivational benefits of e-assessments by developing engaging, interactive formative assessments, which could be used either independently without teacher intervention, or in preparation for a classroom discussion or activity. Beevers et al. (2011) described the ingredients for successful formative e-assessment which included the immediacy of feedback effecting learner motivation.

Another example of an e-assessment tool motivating learners is badging, described by Broadfoot et al. (2013e). Badging is an alternative accreditation system arising from online communities as members validating each other's knowledge, skills or experience via the award of a visual icon. Another benefit of badging is that it can be used to help learners' visualise possible learning pathways. One specific study by Neumann (2010) found that when learners in HE evaluated a science simulation assessment, it was found that it engendered confidence and the author concluded that it held benefits for motivation.

Gee and Schaffer (2010) found that in gaming-style assessments, players were motivated because a sequence of activities gradually increased in difficulty so that players were constantly working at the edge of their abilities.

3.9.2 Learning gains

E-scape (Kimbell et al., 2009) reported on an advantage that came with more reliance on process driven activities for learners being more acquisition of soft skills with a verifiable collection of evidence generated by this type of activity. Broadfoot et al. (2013b) felt that questions answered by students using mobile devices or EVS (electronic voting system / clickers) promoted real-time feedback, collaborative interaction and reflection. While JISC (2007) discussed how e-assessment may illuminate skills of critical thinking, effective decision making, collaborative skills and practical problem solving.

3.9.3 Learner Confidence

Beevers et al. (2011) identified a further beneficial effect of e-assessment on the learner as being how feedback from a computer is non-judgemental, so the learner can explore knowledge and skills privately and comfortably.

3.10 Quality

This section contains descriptions of the issues around the quality of e-assessments. It includes a discussion of the value added by e-assessment, as compared to paper assessments, and of issues around reliability and validity.

3.10.1 Added value paper to e-assessment

Three papers were found that looked at what added value e-assessment gave over paper tests. Wylie & Dolan (2013) described how technology could provide quality data which could be categorised in terms of student learning, misconceptions and cognition to improve the information that learners receive. Furthermore, Boyle (2006) stated how e-assessment had the potential benefit over paper assessments of providing feedback tailored to the learner. Neumann (2010) felt that the impetus for using e-assessment was large class sizes and that e-assessment allowed tutors to track and record student activity.

Pachler et al. (2009) also considered what 'e' added to formative assessment, and his findings incorporated what Kennewell (2001) described as the 'affordances' of the technology. Pachler et al. found five main advantages of formative e-assessment over formative assessment: speed, storage capacity, processing, communication and construction, and representation:

Speed

- Speed of response is often important in enabling feedback to have an effect
- The ability to give feedback quickly means that the student's next problem solving iteration can begin more quickly.

Storage capacity

- The ability to access very large amounts of data (so appropriate feedback/additional work/illustrations can be identified).

Processing

- Automation - in some situations the e-assessment system can analyse responses automatically and provide appropriate feedback
- Scalability – can often be the result of some level of automation
- Adaptivity – systems can adapt to learners' needs / skills.

Communication

- Often the advantage of the 'e' is that it enables rapid communication of ideas across a range of audiences, and the technology allows this range to be controlled, it can be just one person, a group, a class or more
- Aspects of communication can be captured and given a degree of semi-permanence
- This semi-permanence supports the sharing of intellectual objects.

Construction and representation

- Representation – the ability to represent ideas in a variety of ways and to move and translate between these representations. E.g. for learners who are not highly literate the visual nature of the screen can increase motivation (Richardson et al. 2002)
- Technology can support learners in the representation of their own ideas
- Through representation, technology enables concepts to be shaped and this helps learners develop their meaning
- In representing their ideas in digital artefacts learners open up a window on their thinking.
- Mutability – shared objects are not fixed, they can change or be changed with ease.

Broadfoot et al. (2013e) outlined their perceived benefits of using digital technologies for e-assessment: they can provide opportunities for submitting evidence via a range of media; offer more personalised assessments (including prediction modelling); support the integration of summative assessments into learning activities in order to support learner reflection and development; and provide online simulations and environments that are more authentic and relevant.

3.10.2 Validity

Dennick (2009) discussed how assessment principles, such as reliability and validity, were just as important in e-assessment. He described how in fact e-assessment offered potential for new types of questions and formats which could be used to enhance reliability, validity and utility.

The short history of computer-based assessment (CRLI, 2009) shows that technology based testing does not always equate to results obtained for traditional paper tests. There are three main areas that are seen to be potential contributors to these differences.

1. solving problems displayed on the screen requires different cognitive processes from those required when working on paper
2. task types usually associated with paper tests may not always transfer to an electronic medium and
3. the ability of students to demonstrate ability in the assessed skills may be influenced by, or restricted to, their level of IT application skills.

Others that have commented on these issues around validity include Liu et al. (2001) who described evidence for increased validity as follows:

- assessments were more closely matched to the material being taught
- presentation of more than one medium of information seemed to aid the students' recall
- questions reflected real-world situations more accurately and
- students seemed to learn more in their assessments which helped them as they continued their studies.

Swithenby (2006) warned that 'learning experiences that are increasingly mediated through screen activities should be assessed using similar media.' In a study of on-screen mathematics assessments, Hughes et al. (2011) reported how, in interviews, learners talked about their preference for working on paper over the computer, and how this was 'familiar' and 'more natural'. In order to make e-assessments valid, there is a need to make sure that any onscreen formats are familiar to learners and to be aware of how learners show their working, for example, using an onscreen notepad like Overrite.

Dennick (2009) found that when using adaptive testing, questions varied from individual to individual, but if the range of these variables was within agreed boundaries, the *reliability* of the test should not be greatly compromised'. Dennick also commented on two specific types of validity:

- *Face validity*: Does it seem like a fair test to candidates? This is important with e-assessment as learners may be unfamiliar with its processes.
- *Content validity*: Can it be enhanced by using animations, video and sound, hotspot questions, dragging labels over pictures and simulations?

However, Threlfall (2007) argued that *construct validity* was most at risk when considering e-assessment and that the key considerations were how the assessment and the teaching were related and how the mode of assessment enabled the learner to demonstrate their understanding. For e-assessment to be valid we must accept that cognitive processes used when working onscreen and on paper may not be the same. Developers should not try and replicate paper testing, but each mode of assessment should exploit the affordances of that mode, be it paper or screen, to create an authentic experience for the learner. So, for example, the creation of an onscreen ruler which can be dragged over a line to 'measure' it is not a valid task to ask a student to do onscreen (that is replication of paper assessment for all the wrong reasons i.e. administrative not pedagogical); whereas using technology to create and manipulate a graph may well better represent the type of thinking that the student developed and used in the learning of that construct, resulting in a more valid task and valid interpretations of the assessment outcomes.

3.10.3 Reliability

Some affordances or benefits of e-assessment provide opportunities for increased reliability of assessment, particularly in relation to the marking of objective questions which can be automatically marked (Meadows and Billington, 2005). The possibility of human error is also removed when totalling scores.

3.11 Implications for the implementation of e-assessment

In this section, issues raised in the literature relating to what should be considered when putting e-assessment in place are outlined.

BECTA (2003) identified a number of obstacles to be overcome in order to improve the use of ICT in classrooms:

- lack of access to appropriate equipment
- lack of time for training
- lack of models of good practice
- negative attitudes
- computer anxiety
- fear of change
- unreliable equipment and lack of technical support.

The following sections relate to:

1. The source of the impetus for using e-assessment
2. Resources and technical support
3. Teacher training and CPD (including teacher orientation to technology and their pedagogical approach) and
4. Time and frequency of formative e-assessment.

3.11.1 The source of the impetus for e-assessment

Boyle et al. (2011) analysed three large scale UK e-assessment initiatives and concluded with advice for doing better e-assessments including these that are relevant to the ASSIST-ME project:

1. Where e-assessment is part of a policy initiative, make it a central part of the initiative, rather than an after-thought or peripheral concern.
2. Organisations should have a definite, positive reason for doing e-assessment; not just a vague sense that it may address the weaknesses of traditional, pencil and paper approaches. However, organisations should also be realistic and not go overboard with e-enthusiasm.
3. Organisations should find out their users' orientations to e-assessment; some may be conservative (such as schools) whereas others (for instance, employers) might see e-assessment as an essential tool to help them to implement education and training.

Feldman and Capobianco's (2008) literature review found that teachers choose to use technology (or not) in the classroom based on their own beliefs and confidence in using

the technology themselves. Kimbell et al.'s (2009) evaluation of E-scape e-portfolio system similarly found that success was often related to the enthusiasm and skills of individual teachers and that barriers included resistance to change, political complications of introducing new systems into established organisations and high teacher workloads.

So there is some evidence that the impetus for e-assessment take-up comes from school –level or class-level. Unfortunately this conflicts with advice coming from more than one source which recommends that top-down implementation of change, with senior management support and financial commitment is more likely to result in successful implementation, especially where large scale use of e-assessment is required (JISC 2007).

3.11.2 Resources and technical support

The implementation of e-assessment brings with it some specific issues that go beyond those relating to e-learning.

Much of the literature investigated originated in Higher Education (HE) or Further Education (FE). Resources and technical support in schools may be different to those in HE, so this should be kept in mind when interpreting these findings; resources and technical support are required for the implementation of e-learning and e-assessment.

Concerns were raised in the literature about the possible shortfalls in terms of resources and technical support required to enable the implementation of e-learning). An evaluation of the e-portfolio used in schools, E-scape, (Kimbell et al., 2009) identified barriers and enablers common to e-assessment/e-learning systems; enabling factors included minimal network disruption. The impact on the capacity and speed of the school network was a priority issue for centres, as was the non-uniform hard and software provision in schools across the UK (a problem which could be exacerbated when we consider Europe as a whole).

JISC (2007) described a need for technical support and resources including a programme of technical and pedagogic support for teaching staff, interoperability with other systems in institution and shared item banks. There are also potential problems to consider when using summative e-assessment including loss of data, verifying candidate's identity and training for e-invigilators.

Also specific to e-assessment is the need for authoring tools and support for e-assessment developers, especially if they are making the transition from paper assessments to the use of technology. Quellmalz (2009) described the use of tools to guide the process of item writing and item banks that enabled efficient development and assembly of items.

Dennick (2009) provided practical advice on how to implement e-assessment for a course from which we can learn lessons including identifying clear roles for all staff when scaling up e-assessment and considering the financial demands of implementing e-assessment.

3.11.3 Continual Professional Development (CPD) and teacher training

A number of themes arose in the literature relating to Continual Professional Development (CPD) that may be required for the implementation of the ASSIST-ME e-assessments:

- The use of technology
- The use of e-assessment
- The processes of formative assessment
- The links between formative and summative assessment
- The pedagogic approach of IBE and associated competencies
- And, importantly, the links between these characteristics for each of the STM subjects.

Assessing the need for CPD would be a complex task, as participants bring with them a wide variety of experience of each of these areas, as well as a variety of orientations and attitudes towards them. It was argued that e-assessment can enhance a learner's experience if assessment is closely aligned to the pedagogic approach used (JISC, 2007) suggesting that an understanding of teachers' orientation to technology, to formative assessment and IBE practices is necessary to support the use of an e-assessment which values these three aspects. Staff need support during the transitional phase to manage traditional and new methods simultaneously (JISC, 2007).

Wylie and Dolan (2013) reported that the use of items which reported misconceptions to the teacher and acted as a prompt for student/teacher discussion did not require significant change to practice; but they warned that teacher readiness and teacher development must be considered. Work on the E-scape e-portfolio project (Kimbell et al., 2009) reported that it was more effective to support and extend teachers' existing skills than impose radical change.

There was much advice to be found in the literature, the summary of the messages being that the availability of teacher professional development and the release from workload in order to take-up CPD is critical to success (e.g. Broadfoot et al. 2013b, Whitelock, 2006).

Use of technology

It was reported that teachers needed to feel that the use of computers in the class was manageable (ALTA, 2009). This suggests that personal orientation, preferences and having the skills and confidence to take ownership of and use technology varies considerably across teachers.

Use of e-assessment

Good CPD is needed to enable teachers to take advantage of the rich reporting capabilities of technology (Beevers et al., 2011). If teachers understand the affordances of the technology, then they have the power to make decisions about when it is valid to use technology and when to use traditional methods.

Formative assessment

The message came from Feldman and Capobianco (2008) that for teachers to really take on formative assessment and for it to make an impact, they need:

- time and opportunities to engage with the software and hardware
- to understand the items and their relationship with learning and pedagogic methods and
- the opportunities to collaborate with other interested teachers.

A core component around which there is much difference between e-assessment and paper-based assessment is the role of the teacher and to what extent their role in formative assessment includes adaptation of pedagogy (Pachler et al., 2009). E-scape (Kimbell et al., 2009) used existing hardware and software so that it was familiar and easy to follow. Teacher training was more focused on the collection of data and the analysis of it for formative purposes rather than on the use of the hardware and software.

IBE

Kennewell (2008) reported that, when teachers first adopted technology as part of their practice, there was a tendency for interactivity to be superficial and authoritative and only when technology was embedded in teachers' pedagogical knowledge did the technology contribute to deeper, more dialogic interaction amongst students. This suggests that teachers need to have an understanding of the technology and IBE and the relationship between them and what technology can and cannot bring to IBE.

3.11.4 Time and Frequency of formative e-assessment

Cheung (2011) reported that assessment programs that were used for more than 30 minutes a week had a bigger effect than those that were used for less than 30 minutes a week. This reflects Swithenby's (2006) concern that key issues for the success of assessment are 'the regularity and quality of student engagement, the timelines and quality of feedback and the student engagement with their feedback'.

3.12 Implications for evaluation of assessments

Halperna et al. (2012) reported on two evaluation studies of Operation ARIES!. These studies used experimental trials in which learning gains, as measured by performance on short answer questions, was compared across different types of e-learning and types and immediacy of feedback were investigated. The use of short-answer questions to measure performance gains could be criticised for a limited view of what benefits the learning types and feedback types could bring, for example, these performance gains seemed not to include IBE or skills.

Evaluations of the ALTA system (2009) on the other hand, considered wider issues including practical concerns like:

- the ease of use
- teacher training
- manageability for pupils and teachers,
- the ability to use computers in class

as well as educational issues including

- engagement from pupils,
- whether formative assessment was promoted
- how teaching and learning was supported
- whether mathematical skills were developed
- if there was any impact on pupils' ability and
- if self-assessment was enabled.

3.13 Exemplars

In this section examples of e-assessments are listed and described. Criteria for selection of which examples to include as exemplary are:

- e-assessments which are or have been in use and are well-established
- e-assessments which have an element of success i.e. provide some evidence of good practice.
- e-assessments which meet at least one of the ASSIST-ME aims for assessment:
 - support formative assessment
 - enable both formative and summative assessment
 - cover STM subjects
 - focus on inquiry-based education
 - focus on competencies related to inquiry-based education or
 - are relevant to primary and/or secondary education.

JISC (2007) made suggestions about the experiences and needs learners could acquire at different stages of learning, this gives an indication of the types of experiences that may appear in the exemplars from different levels of education:

5-11 years	11-14 years	14-16 years
web-based interactive multimedia learning resources and games, drill and skill quizzes, e-profiling of early years development, high quality, web-based AfL, SATs based on online assessments and exemplars	BBC mobile bite-size quizzes for learning on the move, gaming, online banks of SATS, multimedia resources for learning, innovative games-based assessments	online assessment materials for gifted and talented learners, virtual world simulations testing skills in context (ICT and science), opportunities to personalise their learning using on demand online testing

Exemplars are presented by subject, or combination of subjects.

3.13.1 Mathematics

HE statistics simulation

Neumann (2010) described simulations used for summative assessment of HE statistics. The use of technology for these simulations closely reflected how technology was used in the teaching and learning of statistics. The use of simulations enabled students to have control over their actions and make decisions which had impact.

NRich Mathematics

This team at the University of Cambridge works to enrich the experience of the mathematics curriculum for all learners aged 5 to 18 by offering online enrichment materials (problems, articles and games) to be integrated with every day practice. The main focus is on the development of mathematical thinking and problem-solving skills and many of the activities show rich mathematics in meaningful contexts.

Activities are accessed via the NRich website and some of them exploit the affordances of the technology, including digital learning objects such as simulations and interactive tasks. The resources are organised such that some activities are categorised by level of difficulty: warm up, try this next, think higher, explore further. This gives the learners support in choosing which way to proceed.

The interactive, animated tasks on the NRich website www.nrich.ths.org could be used for formative or self-assessment. Examples of interactive tasks include animations for tasks using peg boards, Cuisenaire rods, cogs, shapes, grids and Carroll diagrams; such tasks enable learners to attempt tasks again and give instant feedback. They could be used as stimuli for IBE, particularly where stimulations and interactive tasks are used.

Misconceptions in mathematics

Wylie and Dolan (2013) reported on the creation of a bank of items for high school mathematics and science teachers that drew on the misconception literature (Wylie and Ciofalo, 2008). Each multiple-choice item that was developed drew on at least one previously identified student misconception, so the formative e-assessments were

through MCQs, with some distractors relating to particular known misconceptions. The advantage over paper formative assessment was that the process of translating student responses into diagnoses of misconceptions was automated. These question types have different implications for item development than those which are not misconception based (Wylie and Dolan, 2013).

Progression assessment

Like the misconceptions-based assessments described above, Arieli-Attali et al. (2012) identified an assessment which aimed to provide quality data to support decision making to inform the next instructional steps and improve the information that learners received. This middle school mathematics project focussed on the progression between levels of understanding, rather than just on categorising the learner according to at which level they sat (CCSSO 2008). One interesting aspect of this tool is how it uses two kinds of assessments which could also feed into the gathering of information for formative and summative assessment purposes. These two kinds of assessments are: *locator* assessment (computer delivered and placing the learner within three learning progressions) and *incremental* tasks (which explicitly targeted a transition between levels, rather than the levels themselves).

The ASSISTment system

Pellegrino and Quellmalz (2010) looked at the ASSISTment technology-enabled assessments which use a pseudo-tutor for middle school level mathematics.

The system used scaffolding questions, optional hints, and buggy messages (specific feedback given after student errors) for each item. Students are eventually guided to reach the correct answer; scaffolds and hints are limited to avoid giving away the answers. Teachers receive feedback on student and class progress both on general summative measures (for example, percentage correct), on more specific knowledge components, and on formative aspects. This e-information available to teachers not only allows them to analyse individual and group performance, but the enhanced information, afforded by the technology, can feed into adapting teachers' pedagogy through the use of formative information.

Abacus Evolve

Pearson's (2013) Abacus Evolve mathematics programme provides online mathematics games. When originally introduced the materials included an interactive CD ROM, a Talk Maths CD ROM for pairs of children to use, a Solve The Problem CD ROM for pairs or groups and individual practice software. The materials designed for pairs and groups of children to work on together may compliment an IB approach and may also yield information when assessing competencies.

Centre for Mathematics, Science and Computer Education

The Centre for Mathematics, Science and Computer Education (whose purpose it is to improve mathematics, science, and computer education programs in the States) offers links to many interactive mathematics tools, including ideas for teaching and assessing mathematics with examples of online support for pupils. While these are not comprehensive systems like many of the other exemplars that we have described, the

electronic mathematics resources do lend themselves to use in the IB classroom. Four specific tools are described below.

1. The National Library of virtual manipulatives (http://nlvm.usu.edu/en/nav/category_g_2_t_1.html e.g.) Here digital learning objects such as abacus, fractions, number lines, bar charts and Venn diagrams can be found.
2. Online mathematics manipulatives http://www.ct4me.net/math_manipulatives.htm . Here learners can submit answers and, if they are wrong, they get instruction and examples of worked through answers. It covers a large range of mathematical content areas for students aged 5 to 16.
3. Visual mathematics learning <http://www.visualmathlearning.com/>. This has onscreen exercises for practice in mathematics, some of which have useful visual clues to support learners with answering questions.
4. Math cats <http://www.mathcats.com/>. This site also has interactive mathematics activities, some of which link to teaching in a more formative way and some of which link to more summative-style assessment. One feature allows learners to create graphs, <http://nces.ed.gov/nceskids/Graphing/> including bar charts, pie charts and line graphs.

Cognitive Tutor

Described as ‘adaptive curricula’, Cognitive Tutor software, <http://www.carnegielearning.com/specs/cognitive-tutor-overview/> , was developed around an artificial intelligence model that identifies weaknesses in each individual student's mastery of mathematical concepts. It then customises prompts to focus on areas where the learner is struggling and sends the learner to new problems that address those specific concepts.

Online activities include:

- multiple representations (these can be expressed numerically or display problems graphically)
- worksheet prompts to convert problems into mathematical expressions
- interactive examples (with step by step instructions for learners)
- flexible sequencing (for teachers or administrators to determine)
- pre and post-tests (a pre-test can be diagnostic and set the pace for further instruction)
- immediate feedback, including giving learners the opportunity for self-correction; the programme recognises the most common errors and misconceptions and responds appropriately and
- a ‘skillometer’ which indicates the journey to mastery for learners and teachers.

The flexible sequencing and pre and post-tests lend themselves to formative and summative assessment practices and, with the personalisation of the content, the system also allows for self-assessment.

ALTA

The ALTA (Adaptive Learning, Teaching and Assessment) system targets KS1 – 3 mathematics (age 5 to 16 years). It was designed to support and promote formative assessment, to inform self-assessment and to inform teaching through the collection of longitudinal records of pupil performance. It includes built-in information resources e.g. question banks mapped onto curricula, and all assessments are adaptive. Teachers can see pupil and class profiles; diagnostic analyses can show trends. The curriculum is represented on a 'STLC grid' which shows the subject, topic, level, criterion (unique learning objective) enabling questions across a range of difficulties. Developed and used in Northern Ireland, ALTA has been independently evaluated fifteen times over five years including five CCEA (Northern Ireland Curriculum Authority) evaluations which all report positively. This approach could be adapted for use with competency-based curriculum and the use of longitudinal data could feed into both formative and summative assessments, with the trends from the diagnostic analysis feeding into the teachers' adaptations of pedagogy.

Bioware and Atari computer game

In the Bioware and Atari computer game, 'Neverwinter Nights', players have to improve their literacy and numeracy skills in order to progress. Completed tasks are banked in an e-portfolio for assessment which was shown to significantly improve success rates in basic and key skills assessments (JISC, 2007). This is an illustration of how gaming-style e-assessment can motivate learners and series of formative assessments can feed into a summative result.

3.13.2 Science

SimScientists

SimScientists is a set of simulation-based science assessments used in middle school science classrooms. The system uses simulations to prompt curriculum-embedded formative assessment. The system identifies types of errors and follows up with increasing levels of feedback and coaching for learners, from identifying that an error has occurred and asking a student to try again, through explaining the concept, to demonstrating and explaining the correct answer (Quellmalz et al., 2012). This model supports IB learning.

SimScientists can also be used as a summative, benchmark assessment providing evidence of middle school students' understanding of ecosystems and inquiry practices (having completed a regular curriculum unit on ecosystems). It illustrates ways that assessment tasks can take advantage of simulations to represent generalisable, progressively complex models of science systems. It present significant, challenging inquiry tasks and provides individualised feedback and customised scaffolding. It claims to promote self-assessment and metacognitive skills which could be in line with IB-related science competencies.

Quellmalz et al. (2012) described how SimScientists linked the targets to be assessed with evidence of proficiency on them, and with tasks and items eliciting that evidence (Messick, 1994; Mislevy and Haertel, 2007 in Quellmalz et al. 2012). The process

begins by specifying a student model of the knowledge and skills to be assessed. The SimScientists assessments used the evidence-centered design method to align the science content and inquiry to be assessed, to scoring and reporting methods, and to the specification of the assessment tasks and items.

Operation ARIES!

Operation ARIES! and later named Operation ARA! (Koenig, 2011) was designed to assess and teach critical thinking about science which relates to specific science competencies. It uses intelligent tutoring and makes use of avatars. Students watch videos and receive communications through email and text message. The approaches involved in the modules are related to IB teaching and learning. It includes three types of module:

1. Interactive training: students read an e-book and after each chapter they are quizzed with multiple choice-type questions. Students receive feedback and tutoring from two avatars from the program throughout.
2. Case studies: students are expected to apply what they have learnt in the previous modules.
3. Interrogation: learners are presented with inaccurate science information through the medium of newspaper headlines and television news channels and must ask questions to ascertain the truth.

Through all three modules, key principals of learning are included, such as:

- Self-explanation- the learner communicates the material to another automated student
- Immediate feedback- through the tutoring system
- Multimedia effects- aiming to engage the student
- Active learning- students engage in solving a problem
- Dialog interactivity- students learn by engaging in conversations and tutor groups and
- Real life examples- intended to help students transfer what they have learnt in one context to another.

PISA and NAEP

PISA and NAEP are two examples of where e-assessment and its affordances are being introduced into formerly paper-based assessments. Pellegrino and Quellmalz (2010) reported on how the 2006 PISA pilot tested a computer-based assessment of science to test knowledge and inquiry processes not assessed in the paper-based booklets. The 2009 NAEP Science Framework and specifications drew upon ETS science simulations work (CCSSO, 2008) and other research to develop their rationale for the design and pilot testing of interactive computer tasks to test the students' ability to engage in inquiry practices. These innovative items were included in the 2009 NAEP science administration (Pellegrino and Quellmalz, 2010).

3.13.3 Mathematics and Science

Hungarian diagnostic assessment

An example of how e-assessment is being developed for national use can be found with the Center for Research on Learning and Instruction, (CRLI, 2009) in Hungary. Here a networked platform for diagnostic assessment is being developed from an online assessment system. The goal is to lay the foundation for a nationwide diagnostic assessment system for grades 1 through 6. The project will develop an item bank in nine dimensions (reading, mathematics and science in three domains).

Scholar

Scholar (www.scholar.hw.ac.uk) is one of the largest online learning programmes in the world with over 80,000 registered students in Scottish secondary schools. It provides online educational resources and a 'virtual college' support network. There are separate courses for biology, physics and chemistry and for mathematics. The science programmes contain some animated graphs with associated questions (some referred to as investigations) and the mathematics programme has step by step demonstrations, for example, on how to construct a pie chart.

It provides opportunities for independent learning plus formative assessment, with pupils working on repeated practice with immediate feedback. Item types include multiple choice questions, short answer and extended answer questions alongside a variety of materials including:

- sets of learning points for revision
- e-learning content for each topic area (which could be in the form of static text, diagrams and/or graphs) and
- revision planners for pupils.

Results of the end of topic tests go to teachers. Staff training is tailored to the school's needs.

SAM Learning

SAM learning (www.samlearning.com) is an online revision and test practice package for mathematics and science designed to mainly support individual revision which is often completed at home. The website claims that 10 hours on SAM Learning improves student achievement by 1 GCSE grade. Motivation for students comes from choosing avatars, 'playing' against their friends and seeing their friend's progress (with the aim of motivating them to do more revision). Activities viewed on the demonstration programme included onscreen versions of paper items (some with drop down menus), some drag and drop items and some fill in the blanks items. The interesting aspect of this is that it is a successful self-assessment package where no teacher input seems necessary.

3.13.4 Technology

National Curriculum Test in ICT for 14-year olds in the UK

Boyle et al. (2011) described the development and pilot of the KS3 ICT test. The test was radical in that it contained several novel features and required that learners solve problems in a virtual world. It assessed ICT capability although the initiative was criticised because the construct was not widely understood nor explicitly aligned to existing widely understood constructs of ICT competence. Finally the test was downgraded from a high stakes summative statutory to a formative test and redesigned as free-standing assessment tasks (QCDA, 2011).

NAEP technology and engineering

The new 2014 Technology and Engineering Literacy Framework for NAEP will be entirely computer administered and will include specifications for interactive, simulation-based tasks involving problem solving, communication, and collaboration related to technology and society, design and systems, and information communications technology (Pellegrino and Quellmalz, 2010).

E-scape

Kimbell et al. (2009) evaluated the use of e-portfolios for design and technology, geography and science. E-scape was designed to capture the process that learners go through, including capturing collaboration. It had the facility to capture a timeline of learner activities. Learners could record their own ideas alongside any justifications for their ideas or actions. The intelligent tutoring included teaching elements and the timeline could highlight points in the learner's development e.g. when designing a guitar, learners looked at videos and articles, it allowed students to comment on these articles or stimuli and the timeline captured how this affected the development of their design. The digital form of the e-portfolio meant that there were more options for students to upload material e.g. the use of film, photos, maps and web links.

BTEC in IT skills

JISC (2007) discussed the BTEC intermediate and advanced award in IT skills for learners in small businesses from Edexcel. The course is delivered via the internet with tutor support online. The skills-based course allows learners to progress at their own pace and has no formal examinations. On-going assessment is assimilated into the structure and content of the course. Learners complete tasks as evidence of achievement and self-assessment exercises at the end of each unit allow them to obtain formative feedback.

4. Conclusions and implications

The ASSIST-ME proposal states that 'ASSIST-ME will develop formative assessment methods that (1) fit into everyday classroom practice, (2) provide qualitatively oriented descriptions and monitoring of competence-oriented, inquiry-based learning processes, and (3) can be combined with existing summative assessment requirements and methods used in different educational systems. The assessment methods will be developed to capture both general competences and disciplinary process competences such as science investigations and authentic problem solving.

The development and design of these methods will be based on existing research on formative and summative assessment, on current research-based understandings of competences in STM, and on previous and on-going EU projects on inquiry-based education (IBE).'

Following some reflections on the application of the methodology and the analysis of the data, the conclusions are presented with reference to the objectives set out section 0.

1. Identify existing relevant digital assessments
2. Through the literature, identify theories and models which are relevant to the development of such digital assessments
3. Identify strategies used in the evaluation of the models which could inform good practice
4. Identify implications for the development of the digital assessments relevant to the aims of ASSIST-ME.

4.1 Reflections on the methodology and the process of searching

The literature search involved a quite general review of literature on e-assessment rather than looking specifically at the narrow focus of e + IBE + STM + competences, as this field yielded next to no information. However, these broader findings do relate to the specific ASSIST-ME focus.

The literature search revealed that the majority of reported use of e-assessment found was in the Higher Education (HE) and Further Education (FE) sector. While the focus of the ASSIST-ME project is for the primary and secondary phase, findings showed how it was possible to use HE as a proxy for the primary and secondary sectors and that general principles of e-assessment used in HE/FE could be used in this project.

Dermo (2009, in Voelkel, S. 2013) reported on how 'e-assessment is widely accepted by students as part of their university studies and they generally feel that it had a positive impact on their learning'. We are reminded that HE is used to e-assessment and e-learning, Dermo here saying this is in part due to technology having been embraced.

A small amount of the findings was specifically related to IBE. However, through reading, it became clear that in order to assess IBE (and related competencies), there's a need for learners to engage in IBE as part of their assessment and so ideas and e-

learning and assessment together could feed onto this process. There do exist conceptions that e-assessment is more compatible with objective, multiple-choice type questions, but actually our findings show that this conception is not correct and exemplars have shown how e-assessment and IBL are compatible.

Originally the literature search was focused on the last few years, but it became evident throughout the research that changes in e-assessment practices were slow and minor and so it was in fact appropriate to look back over the last 15 years.

4.2 Conclusions specifically related to objectives

4.2.1 Objective 1: Theories and Models relevant to development of digital assessments

A variety of theories and models for e-assessment have been identified. This section aims to draw out the advantages of e-assessment, as compared with paper/traditional tests, from these themes to establish good and interesting practice. These themes are followed up under objective 4, where implications for the development of e-assessment for ASSIST-ME are discussed.

Teaching/Learning/Assessment Link

Learners benefit when teaching, learning and assessment are linked and various sources in this review have demonstrated how e-assessment can facilitate this. Models identified in the search have blended teaching and short, repeated, formative tasks and has demonstrated how a sequence of these formative tasks can combine to give a more summative assessment. Gaming-style assessments can be a model where the teaching is all on-screen and through a series of smaller tasks (the formative element), learners, working at the edge of their ability can move towards mastery of skills (the summative element)

Stimulus types

One of the affordances of the technology is the types of stimulus that can be provided for learners to work with during e-learning and e-assessment. Digital learning objects can be used as a stimulus and learners can manipulate them to give them some grounding for their next actions. Worked exemplars can be included which support learners in knowing how to proceed with a problem/investigation. A variety of models for these has been found from examples where the entire process is revealed, to ones where hints and clues can be given to students or where only the next small step is revealed. A broader range of real-life scenarios and simulations are often more easily replicated on-screen. These cannot only offer more valid assessment stimuli, but can also increase students' motivation.

Feedback

Many successful models of e-assessment include elements of feedback, from the simple to the complex. The immediacy of the feedback offered by e-assessment has been shown to be a motivating factor for students both in terms of continuing with their learning/revision/understanding and increasing the likelihood that students undertake e-assessments out of choice. E-assessment feedback comes in many forms and can

be a form of learning in itself with feedback models, for example, linked to intelligent tutoring offering links to further learning.

Adaptivity

E-assessment enables a more interactive approach for learners. It seems that adaptive models not only personalise the learning and assessment, but act in a supportive way so that learners are more motivated and their time and effort is more focused on their ability and level of understanding and skills. These aspects of feedback and interactivity lend themselves to group work and peer assessment, both of which can be features of IBE.

These interactive and adaptive features that can come along with e-assessment can lead to more learner autonomy. It can also lead to improved self-assessment as learners are more likely to take notice due to the immediacy of the feedback offered through e-assessment. The on-demand nature of some formative and summative e-assessments has been shown to increase student motivation also.

Data for formative and summative purposes

Large amounts of data can be collected with e-assessment. This may be just scores, as with a paper tests, but it may also give group scores, compare scores to previous cohorts, compare sub-groups of the whole cohort and give information about progress. As well as this type of data, e-assessment can also give insights into how learners approached the task. This mass of data can enhance the teachers' understanding of the learning and act as a strong contributor to formative assessment and adapt the behaviour of the teacher and the learner. This is particularly relevant with this project, as it can capture individual student's input and can collect evidence of how learners approach problem solving including the sequences they used, the strategies they used, the number of attempts a learner took and the amount of time taken on different sections of the activity.

The use of learning analytics can feed into predictive modelling, user profiling and adaptive learning and so support a formative approach.

Re-submission

One aspect that e-assessment allows for, which can feed into meaningful, formative assessment, is the ability to re-submit answers. Original inputs can still be captured for use by the teachers so that they are aware of how easily a learner may have arrived at an answer, but for learners, the immediacy of this feedback and the opportunity to resubmit can make the process and learning more relevant than waiting a period of time to see the results. The immediacy of feedback with e-assessment allows learners to move directly on to the next stage which has advantages in formative assessment and with self, peer and diagnostic assessment. Learners can also submit a confidence rating with their answer which adds further meaning to the level of information gathered.

4.2.2 Objective 2: Strategies used in the evaluation of the models which could inform good practice

In the literature that was reviewed, there were few examples of how e-assessments had been evaluated. A model for development of e-assessments did not seem to follow the usual pattern of *design – trial – evaluate* and so few cases of evaluation techniques were identified.

In the evaluation of Operation ARIES! (Halperna et al., 2012) learning gains were measured in terms of short answer responses. This was compared across:

- different types of e-learning and
- immediacy of feedback.

In the evaluation of the ALTA system (Adaptive Learning Teaching Assessment for mathematics KS1 – 3) a less limited view of learner gains was adopted. It analysed:

- practical concerns: ease of use, teacher training, manageability for pupils and teachers and ability to use on computers in class and
- educational issues: engagement from pupils, whether formative assessment was promoted, how teaching and learning was supported, whether mathematical skills were developed, if there was any impact on pupils' ability and if self-assessment was enabled.

4.2.3 Objective 3: Existing relevant digital assessments

This report has identified both individual elements of digital assessments and complete assessment programmes which may comprise a number of elements including learning. The complete programmes considered in the review are ones that are established; some are used organisation-wide and some are national programmes. The exemplars are also in the subjects considered in this project: science, mathematics and technology.

Finding existing programmes has shown how e-assessment can be integrated into the learning process and be successful in terms of motivating learners, assessing competencies that are more difficult to assess using paper tests and can include elements difficult to replicate in a paper test, for example, simulations, videos and scenarios that lend themselves more readily to IBE and competency-based skills.

Mathematics

The mathematics exemplars included the use of questions, simulations, enrichment activities, activities designed to assess misconceptions and online games. They were designed to feed into summative assessment, formative assessment, self-assessment and diagnostic assessment. Some were adaptive in nature. They were aimed at a range of age groups: 5-16 years, 5-19 years, middle school, high school, and Higher Education.

Some reflect how technology is used in teaching and learning, others use digital learning objects or online/virtual manipulatives. Some use responses for diagnostic purposes, identifying where a learner is struggling and directing them to new questions

to address specific concepts. Others focus on learner progression to inform the next instructional steps. Some involve a pseudo-tutor with hints and feedback after errors; with these detailed feedback is also passed to teachers. Certain exemplars bank completed tasks in e-portfolios and others inform teaching through the collection of longitudinal records.

Science

The science exemplars often involve inquiry tasks and processes and critical thinking (closely linked to the ASSIST-ME focus). The exemplars found focused on summative and formative assessments. They were aimed at ages 11-19 years and middle school students.

They incorporate the use of feedback and some are interactive in nature. They utilise e-coaching as well as intelligent tutoring. Some include the facility to re-submit answers. Some employ videos and case studies and many use simulations.

Mathematics and Science

The examples found were mainly for use in revision and for independent learning. They were aimed at all secondary school ages. They were designed for formative assessment, diagnostic assessment and self-assessment.

One incorporated an item bank and another utilised digital learning objects and demonstrations of answers.

Technology

The technology examples found were used summatively as they were all high stakes qualifications and, as such, aimed at 14-19 year olds. Some aspects could also be used for formative assessment purposes and one claimed the on-going nature of the programme could be used for self-assessment.

Some demanded problem solving in a virtual world and others used interactive simulations. They utilised intelligent tutoring and online tutor support, as well as allowing students to progress at their own rate. They included the use of a timeline of student activities and an e-portfolio.

4.2.4 Objective 4: Implications for the development of the digital assessments relevant to the aims of ASSIST-ME

Information gathered through this process was used to form a set of recommendations to use when considering the development of e-assessments. The aspects of this that are relevant to developing formative and summative assessments in IBE in STM subjects are presented here. A key message with regard to making good use of e-assessment is to exploit the affordances of the technology and not to just translate a paper test to an online version.

Exploiting the affordances

The affordances of the technology should be exploited to enable the assessments to go beyond what paper can do. This could be via the inclusion of particular elements, for example, scaffolding questions, optional hints and clues, simulations and scenarios,

digital learning objects, all of which can support the learners to understand which way to proceed with IBE. E-assessment has the ability to be interactive which can be an advantage when using it in IBE. It could be argued that cognitive skills can be assessed on paper whereas to assess inquiry skills paper is not a suitable format.

Interactive tasks and simulations can give intrinsic, visual feedback to allow the learner to make decisions about the accuracy of their work and when they are ready to submit an answer (if that is necessary). However, there is a need to avoid 'random button pressing'. Interactive tasks should not be superficial or authoritative, but should give learners control and contribute to deeper, dialogical interaction among students.

E-learning and e-assessment relationship

E-assessment should be an integral part of the pedagogy and be closely linked to learning. The increase in learner control afforded by the technology allows for closer integration of teaching, learning and assessment. E-assessment loses its impact when it is just an add-on to usual classroom practice. It has also been found that the impact of e-assessment is insignificant where little time and attention is placed on it. With this in mind, it must be stressed how important it is that the format and functionality of the e-assessments are familiar to the learner. Many e-programmes firstly develop a good learning package and then add in the assessment element.

The relationship between e-learning and e-assessment is key (as is the relationship between learning and assessment) to the success of formative e-assessment. It is a matter of validity that the two need to relate to the same constructs. To ensure that the outcomes of e-assessment are used validly, there is a need to analyse which aspects of the curriculum could and do use technology in teaching/learning and to start with these as areas where e-assessment would be a valid approach.

The relationship between the teacher and student is also central to effective formative assessment, and the risks of using off the shelf e-assessments are that this relationship is distanced. Wylie and Dolan (2013) suggested that teachers should develop their own evaluative tools as part of their instructional practice. E-assessment can provide prompts for discussion, as long as teachers are skilled and/or supported in using e-assessment outputs (Wylie and Dolan, 2013).

Link to summative and formative assessment

Having adaptive e-assessment can better support the learner with self-assessments, as more work is levelled near to their level and so more of the feedback is relevant, more detail around the edges of their ability can be gathered and so learners can get a clearer picture of where they are, identify gaps or misconceptions in their current knowledge and where they need to go next.

E-assessment can support the bringing together of formative and summative assessment, as technology gives opportunities for the blurring of traditional lines between learning, formative assessment and summative assessment (Bennett, 2002). Is the blurring of the distinction between formative and summative assessment meaningful to students? Broadfoot et al. (2013) suggest that actually it is meaningful to have a more holistic view of assessment, when assessment is authentic.

Gaming models can fulfil the purposes of both formative assessment and summative assessment (Gee and Schaffer, 2010).

Feedback

The rich feedback that can be given to students when this is integrated with e-assessment give it a distinct advantage over other types of assessment. The immediacy of receiving feedback for the learner has been shown to be a motivator, including motivating learners to take more tests and for further learning. Another motivating factor comes about when the e-assessment is adaptive; motivation comes from when the activity is personalised to be pitched at the correct level for the individual learner.

If only one affordance of the technology were to be exploited, it should be the speed and detail of feedback that it is possible to give directly to students. Feedback should be instant, differentiated and individualised. Formative e-assessment is most effective when feedback relates to cognitive processes, not just a score or success/fail message. Feedback should be actionable, and students need to understand what to do with the feedback.

Interventions and intelligent tutoring

E-assessment provides opportunities for interventions which are speedy, based on evidence and good pedagogic principles. The great variety of intelligent tutoring options available allow for learner style and preferences to be accounted for, to ensure that e-learning is individualised.

Teacher Continual Professional Development (CPD)

The aims of the ASSIST-ME project assessments are wide and so CPD requirements could cover a range of areas: technology use in teaching and learning; e-assessment; formative assessment; links between summative and formative assessment; the pedagogic approaches of IBE and associated competencies; how each of these apply to the STM subjects; and very importantly, the interrelationships between each of these aspects. Most importantly teachers need to be released from their usual workload in order to take up CPD (Whitelock, 2006). There is a need, with the introduction of e-assessment, for there to be provision for the CPD of teachers who will be engaged with the technology. There is also a need for institutions to provide technical support so that teachers can feel confident when using the e-assessment. Technical support also ensures that teachers are aware of the features of the e-assessment and of all of the data outputs from the e-assessment so that it can be best used to enhance the teaching and learning.

Implementation

Due to the nature of this type of technology and the processes involved in the development of e-assessments, it is usual that the introduction of e-assessment would need to be a top down process, that is from the management, rather than an individual teacher bringing in a novel e-assessment technique; top-down change is needed, from policy makers, as practitioners will not be able to influence change of this nature from the bottom up. Some e-assessment projects begin with one or a handful of enthusiastic

teachers and succeed as small scale projects, but for e-assessment initiatives which are scaled up, a key factor in success will be the financial commitment and support of senior managers (JISC, 2007). Developers need to have a positive reason for using e-assessment and, once committed to e-assessment, they need to make it central to assessment.

During the implementation of the E-scape project, centres were concerned with support and adequate training for e-assessment alongside expectations of high cost. Dennick (2009) referred to economic issues and demands brought by e-assessment including staff to support the system, trainers, IT support and, software. Technical expertise and resources are required for the successful implementation of e-assessment (JISC, 2007), particularly to overcome the most common problems of interoperability with existing systems, network capacity and technical and pedagogical support for staff.

4.3 E-assessment and formative assessment

Many feel that high stakes assessment is incompatible with e-assessment innovation (Boyle et al., 2011). Beevers et al. (2011) argued that when there exists tension between using e-assessment for formative assessment and summative (high stakes) assessment, the latter will not get off the ground because of political reasons. The former is ideal for innovating with e-assessment. Beevers et al. predicted that formative assessment would be the vehicle for the e-assessment breakthrough because summative assessment is risk-averse and subject to political factors which inhibit the use of e-assessment.

4.4 Last words

The outcomes of the WP2 project will input to WPs 4 and 5 – the design of a range of combined assessment methods. The key messages contained in this document that the authors wish to highlight are summarised below.

1. When considering the validity of e-assessments, it is important to link the construct being taught/learned and the construct being assessed. This will require an understanding of the impact of technology on teaching / learning and the construct being assessed needs to reflect this. Technology impacts on cognitive processes and subject-related thinking and so to mitigate against negative impact, technology should be integrated with existing classroom practice.
2. IBE requires learners to interact with problems and situations; replicating this in assessment is more easily done through technology than on paper because the technology offers the affordance of interaction with authentic problem solving environments, instant feedback and adapting to the learners' responses.
3. The more elaborate the e-assessment package is the more effective it is; the benefits of the technology are not realised when a paper test is migrated to screen, but can be more fully exploited when the potential affordances of e-assessment technology are fully embraced.
4. The richness of the outputs from e-assessment allows teachers to monitor individual, group and sub-groups' attainment, to see progress over time and across interventions, all of which feed into a more complete formative education process.

5. If the developers were to take up one affordance of the technology then it should be to exploit the potential for immediate and detailed feedback to the student and the associated intelligent tutoring opportunities.
6. Using a gaming model (incremental assessments leading to mastery and progression) is an example of how to blur the distinction between formative and summative assessment, but importantly in terms of the aims of ASSIST-ME, gaming also shows how formative and summative assessment can be brought together in a IBE environment which develops and recognises competencies.

References

- ALTA (2009). *Adaptive Learning Teaching Assessment* (website). Retrieved from www.alta-systems.co.uk [accessed 11-05-13].
- Anderson, R. (2002). Reforming Science Teaching: What Research says about Inquiry. *Journal of Science Teacher Education*, 13(1), 1-12.
- Assessment Reform Group (2002). *Assessment for learning - 10 principles: Research-based principles to guide classroom practice*. Retrieved from <http://www.assessment-reform-group.org.uk> [accessed 12-05-13].
- Arieli-Attali, M., Wylie, E.C., & Bauer, M.I. (2012, April). *The use of three learning progressions in supporting formative assessment in middle school mathematics*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.
- Beatty, I. D., & Gerace, W. J. (2009). Technology-Enhanced Formative Assessment: A Research-Based Pedagogy for Teaching Science with Classroom Response Technology. *Journal of Science Education & Technology*, 18(2), 146-162.
- BECTA (2003): *What the research says about barriers to the use of ICT in teaching (online report)*. Retrieved from www.becta.org.uk [accessed 14-05-13].
- Beevers et al. (2011). What can e-assessment do for learning and teaching? Part 1 of a draft of current and emerging practice review by the e-Assessment Association expert panel. *International Journal of e-Assessment*, 1(2).
- Black, P. & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability* (formerly the *Journal of Personal Evaluation in Education*), 21(1), 5-31.
- Boyle, A (2006). *Evaluation of the 2006 pilot of the key stage 3 ICT test*. QCA Assessment Research team.
- Boyle, A. & Hutchison, D. (2009). Sophisticated tasks in e-assessment: what are they and what are their benefits? *Assessment & Evaluation in Higher Education*, 34(3), 305-319.
- Boyle, A., Sceeny, P and Sowerbutts, P. (2011). A history of three e-assessment programmes in England. *International Journal of e-Assessment*, 1(2).
- Broadfoot, P., Timmis, S., Payton, S., Oldfield, A. & Sutherland, R. (2013a). *Rethinking Assessment, series of discussion papers. Paper 1 Transforming education through technology enhanced assessment*. Graduate School of Education, University of Bristol (www.bristol.ac.uk/education).
- Broadfoot, P., Timmis, S., Payton, S., Oldfield, A. & Sutherland, R. (2013b). *Rethinking Assessment, series of discussion papers. Paper 2 Integrating FA and SA through technology enhanced assessment*. Graduate School of Education, University of Bristol(www.bristol.ac.uk/education).
- Broadfoot, P., Timmis, S., Payton, S., Oldfield, A. & Sutherland, R. (2013c). *Rethinking Assessment, series of discussion papers. Paper 3 Exploiting the collaborative potential of technology enhanced assessment in higher education*. Graduate School of Education, University of Bristol (www.bristol.ac.uk/education).
- Broadfoot, P., Timmis, S., Payton, S., Oldfield, A. & Sutherland, R. (2013d). *Rethinking Assessment, series of discussion papers. Paper 4: Learning analytics and*

- technology enhanced assessment*. Graduate School of Education, University of Bristol (www.bristol.ac.uk/education).
- Broadfoot, P., Timmis, S., Payton, S., Oldfield, A. & Sutherland, R. (2013e). *Rethinking Assessment, series of discussion papers. Paper 5 Ethical issues in technology enhanced assessment*. Graduate School of Education, University of Bristol (www.bristol.ac.uk/education).
- Broadfoot, P., Timmis, S., Payton, S., Oldfield, A. & Sutherland, R. (2013f). *Rethinking Assessment, series of discussion papers. Paper 6 National standards and technology enhanced assessment*. Graduate School of Education, University of Bristol (www.bristol.ac.uk/education).
- Burrow, M., Evdorides, H., Hallam, B. & Freer-Hewish, R. (2005). Developing formative assessments for postgraduate students in engineering. *European Journal of Engineering Education*, 30(2), 255-263.
- CCSSO (2008). *Formative Assessment: Examples of Practice*. A work product initiated and led by Caroline Wylie, ETS, for the Formative Assessment for Students and Teachers (FAST) Collaborative. Council of Chief State School Officers: Washington, DC.
- Center for Mathematics, Science and Computer Education. Online/ interactive mathematics activities. Retrieved from http://www2.eboard.com/eboard/servlet/BoardServlet;jsessionid=6FE830D99D518952E43C89FBDC402DBC?ACTION=NOTE_SHOW&ACTION_ON=NOTE&OBJEC T_ID=235808&SITE_NAME=cmsce&BOARD_NAME=mathatthelab&SESSION_ID=e m0i3gb2yq197279&TAB_ID=81941 [accessed 27-05-13].
- Cheung, A. C. K., (2011). *The Effectiveness of Educational Technology. Applications for Advancing Mathematics Achievement in K-12 Classrooms: A Meta-Analysis Best Evidence Encyclopedia (BEE)*. Retrieved from [www. Bestevidence.org](http://www.Bestevidence.org) [accessed 05-06-13].
- Carnegie Learning Cognitive Tutor (website). Retrieved from <http://www.carnegielearning.com/specs/cognitive-tutor-overview> [accessed 03-06-13].
- Clesham, R. (2009). *Making it real: Interactive assessment of national curriculum science process skills*. Paper presented at BERA (British Educational Research Association).
- Cook, J. & Crabb, N. (2002). *Genuine thinking or random button pressing? Self-assessment design in computer-based learning materials. Winds of change in the sea of learning*. Proceedings of the 19th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education (ASCILITE).
- Crisp, G. (2010). Interactive e-assessment – practical approaches to constructing more sophisticated online tasks. *Journal of Learning Design*, 3(3), 1-10.
- Crisp, V. & Ward, C. (2008). The development of a formative scenario-based computer assisted assessment tool in psychology for teachers: The PePCAA project. *Computers & Education*, 50(4), 1509-1526.
- CRLI (2009). Center for Research on Learning and Instruction, University of Szeged (website). Retrieved from www.edu.u-szeged.hu/~csapo/irodalom/DIA/Diagnostic_Assessment_Project.pdf [accessed 06-06-13].

- Daly, C., Pachler, N., Mor, Y. & Mellar, H. (2010). Exploring formative e-assessment: using case stories and design patterns. *Assessment & Evaluation in Higher Education*, 35(5), 619-636.
- Davis, M., McKimm, J. & Forrest, K. (2013). *E-Assessment – The use of Computer-Based Technologies in Assessment. How to Assess Doctors and Health Professionals* (pp. 41-52). Chichester, UK: John Wiley & Sons, Ltd.
- Dennick, R. E. G., Wilkinson, S. & Purcell, N. (2009). Online eAssessment: AMEE Guide No. 39. *Medical Teacher*, 31(3), 192-206.
- Doorey, N. A. (2012). Coming Soon: A New Generation of Assessments. *Educational Leadership*, 70(4), 28-34.
- Feldman, A. & Capobianco, B. M. (2008). Teacher Learning of Technology Enhanced Formative Assessment. *Journal of Science Education & Technology*, 17(1), 82-99.
- Furse E. (2009). Computer Based Revision. In Joubert, M. (Ed.), *Proceedings of the BSRLM, Vol.29, no.3*. Nov 2009, 25-30.
- Gee J.P. & Schaffer D.W (2010). Looking Where the Light is Bad: Video Games and the Future of Assessment. *Edge: the latest information for the education practitioner* 6(1), 3-19.
- Hughes, S. (2006). *Assessing mathematics onscreen: What are we assessing?* Proceedings of the British Educational Research Association Conference, Warwick University, September 2006.
- Hughes, S. & Rose, P. (2006). *The Computer Mediated Test Project. Report on Phase 1: The development and evaluation of key stage 2 and key stage 3 computer mediated mathematics test items*. Research Team, Test Development Team, Edexcel.
- Hughes, S., Custodio, I. Sweiry, E. & Clesham, R. (2011). *Beyond multiple choice: Do e-assessment and mathematics add up?* Proceedings of the Association for Educational Assessment Conference 2011, Queen's University, Belfast. November 2011.
- JISC (2007). *Effective Practice with e-Assessment: An overview of technologies, policies and practice in further and higher education* (online). Retrieved from <http://www.jisc.ac.uk/media/documents/themes/elearning/effpraceassess.pdf> [accessed 06-06-13].
- Johnson J. (2013). *Three Things Game Designers Need to Know about Assessment*. ETS Research Spotlight, Issue 8, April 2013. Retrieved from www.ets.org [accessed 28-05-13].
- Kennewell, S (2008). *Interactivity in the classroom and its impact on learning*. Paper presented at the British Educational Research Association Conference, Herriot Watt University, September 2008.
- Kennewell, S (2001). Using affordances and constraints to evaluate the use of information and communications technology in teaching and learning. *Journal of Information Technology in Teacher Education*, 10(1&2), 101-116.
- Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Martin, F., Davies, D., Pollitt, A. & Whitehouse, G. (2009). *E-scape portfolio assessment phase 3 report*. Retrieved from http://www.gold.ac.uk/media/e-scape_phase3_report.pdf [accessed 29-05-13].

- Llewellyn, D. (2007). *Inquire within: implementing inquiry-based science standards in grades 3-8*. Thousand Oaks: Corwin Pr.
- McKinsey and Company (no date). *Transforming Learning through mEducation*
Retrieved from
<http://www.mckinsey.com/Search.aspx?q=formativ%20assessment> [accessed 12-05-13].
- Meadows, M. & Billington, L. (2005). *A Review of the Literature on Marking Reliability*. Manchester: AQA [online]. Retrieved from
https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf [accessed 26-06-13].
- Miller, T. (2009). Formative computer-based assessment in higher education: the effectiveness of feedback in supporting student learning. *Assessment & Evaluation in Higher Education*, 34(2), 181-192.
- Neue, F. (no date). *Enhancements of the Perception™ Assessment Server for Metacognitive Analyses*, Neue Technologien und Lernen in Europa e.V. (fred.neumann@belab.de). PepCAA. Retrieved from www.pepcaa.odl.org [accessed 15-05-13].
- Neumann, D. L. (2010). Using Interactive Simulations in Assessment: The Use of Computer-Based Interactive Simulations. In: The Assessment Of Statistical Concepts. *International Journal for Technology in Mathematics Education*, 17(1), 43-51.
- NRich website (no date). Retrieved from www.nrich.maths.org [accessed 06-06-13].
- Pearson (2013). *Abacus Evolve mathematics programme (Pearson) for the primary sector* (website). Retrieved from
<http://www.pearsonschoolsandfecolleges.co.uk/Primary/Mathematics/AllMathematicsresources/AbacusEvolveFrameworkEdition/Structure/Structure.aspx> [accessed 24-06-13].
- Peat, M. & Franklin, S. (2002). Supporting student learning: the use of computer-based formative assessment modules. *British Journal of Educational Technology*, 33(5), 515-523.
- Pellegrino, J. W. & Quellmalz, E. S. (2010). Perspectives on the Integration of Technology and Assessment. *Journal of Research on Technology in Education*, 43(2), 119-134.
- Purvis, A. J., Aspden, L. J., Bannister, P. W. & Helm, P. A. (2011). Assessment strategies to support higher level learning in blended delivery. *Innovations in Education & Teaching International*, 48(1), 91-100.
- QCDA (2011). *Key stage 3 ICT onscreen assessment tasks*. Retrieved from <http://www.qcda.gov.uk/assessment/6555.aspx> [accessed May 20, 2011].
- Quellmalz E. S. & Pellegrino J. W. (January 2009). *Technology and Testing*, Vol. 323. Retrieved from www.sciencemag.org [accessed 02-06-13].
- Quellmalz, E. S. & Pellegrino, J. W. (2009). Technology and Testing. *Science*, 323(5910), 75-79.
- Ramadoss, R. & Wang, Q. (2012). Evaluation of a web-based assessment tool for learning grammar at the primary school level. *International Journal of Continuing Engineering Education & Lifelong Learning*, 22(3/4), 175-184.

- Redecker, C. & Johannessen, Ø. (2013). Changing Assessment -- Towards a New Assessment Paradigm Using ICT. *European Journal of Education*, 48(1), 79-96.
- Richardson, M., Baird, J-A., Ridgway, J, Ripley, M., Shorrocks-Taylor, D. & Swan, M. (2002). Challenging minds? Students' perceptions of computer-based World Class Tests of problem solving. *Computers in Human Behaviour*, 18, 633-649.
- Sainsbury, M. & Benton, T. (2011). Designing a formative e-assessment: Latent class analysis of early reading skills. *British Journal of Educational Technology*, 42(3), 500-514.
- SAM Learning (no date). Retrieved from www.samlearning.com [accessed 06-06-13].
- Scalise, K., Timms, M., Moorjani, A., Clark, L., Holtermann, K. & Irvin, P. S. (2011). Student learning in science simulations: Design features that promote learning gains. *Journal of Research in Science Teaching*, 48(9), 1050-1078.
- SCHOLAR in Scotland at the school/university interface (no date). Retrieved from www.scholar.hw.ac.uk [accessed 07-06-13].
- Shermis, M. D., Burstein, J. & Leacock, C. (2006). *Applications of computers in assessment and analysis of writing*. In: C. A. McArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing research* (403-416). New York, NY: Guilford Press.
- Sluijsmans, D. M., Prins, F. J. & Martens, R. L. (2006). The Design of Competency-Based Performance Assessment in E-Learning. *Learning Environments Research*, 9(1), 45-66.
- Stödberg, U. (2012). A research review of e-assessment. *Assessment & Evaluation in Higher Education*, 37(5), 591-604.
- Swithenby S. J. (2006). *Screen-Based Assessment*. New Directions, HEA Journal Issue 2, December 2006. Thousand Oaks, CA: Corwin Press. Online available at <http://journals.heacademy.ac.uk/doi/abs/10.11120/ndir.2006.00020023>.
- Threlfall, J; Pool, PC; Homer, MS & Swinnerton, BJ. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer, *Educational Studies in Mathematics*, 66(3), 335-348.
- Timmers, C. F., Braber-van den Broek, J. & Van den Berg, S. M. (2013). Motivational beliefs, student effort, and feedback behaviour in computer-based formative assessment. *Computers & Education*, 60(1), 25-31.
- Tippins, N. T. (2011). *Overview of Technology-Enhanced Assessments*. Technology-Enhanced Assessment of Talent (1-18). San Francisco: Jossey-Bass.
- University of Akron (2012). Retrieved from <http://www.uakron.edu/education/about-the-college/news-details.dot?newsId=40920394-9e62-415d-b038-15fe2e72a677&pageTitle=Recent%20Headlines&crumbTitle=Man%20and%20%20machine:%20Better%20writers,%20better%20grades> [accessed June 17 2013].
- Van der Kleij, F. M., Eggen, T. J. H. M., Timmers C.F. et al. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education*, 58 (1), 263-272.
- Voelkel, S. (2013). Combining the formative with the summative: the development of a two-stage online test to encourage engagement and provide personal feedback in large classes. *Research in Learning Technology*, 21. 19153.

- Wang, K. H., Wang, T. H., Wang, W. L. & Huang, S. C. (2006). Learning styles and formative assessment strategy: enhancing student achievement in Web-based learning. *Journal of Computer Assisted Learning*, 22(3), 207-217.
- Wang, T. H. (2007). What strategies are effective for formative assessment in an e-learning environment? *Journal of Computer Assisted Learning*, 23(3), 171-186.
- Wesiak, G., Al-Smadi, M., Höfler, M. & Gütl, C. (2013). Assessment for Complex Learning Resources. Development and Validation of an Integrated Model. *International Journal of Emerging Technologies in Learning*, 8, 52-61.
- Whitelock, D. (2006). *Roadmap for e-assessment*. Report for JISC, Open University. Retrieved from http://www.jisc.ac.uk/search.aspx?keywords=roadmap%20for%20e-assessment&filter=s&type=adv&sort=relevance&method=all&collection=default_collection [accessed 06-06-13].
- A research and development project being conducted by the Center for Research on Learning and Instruction, University of Szeged (2009). *Developing an online diagnostic assessment system for grades 1 to 6*. Retrieved from www.edu.u-szeged.hu/~csapo/irodalom/DIA/Diagnostic_Assessment_Project.pdf [accessed 22-05-13].
- Wylie, E. C. & Ciofalo, J. F. (2008). *Supporting teachers' use of individual diagnostic items*. Teachers College Record. Retrieved from <http://www.tcrecord.org/content.asp?contentid=15363> [accessed 13-06-13].
- Wylie, E. C. & Dolan, R. P. (2013). *The Role of Formalized Tools in Formative Assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Zakrzewski, S. & Steven, C. (2003). Computer-based assessment: quality assurance issues, the hub of the wheel. *Assessment & Evaluation in Higher Education*, 28(6), 609-623.

Appendices

Appendix 1. Search terms used in literature review

e-assessment	Computer Based Assessment OR E-assessment OR E-learning OR Integrated Learning Systems OR Technology enhanced assessment
e-learning	Learning analytics OR Intelligent tutoring OR Intelligent measurement OR Digital learning OR Digital objects
Formative assessment	Continuous measurement OR Embedded assessment OR Formative assessment OR Integrated assessment
Summative assessment	Summative Assessment OR Assessment
Inquiry based education	Inquiry based learning OR inquiry OR collaborative learning OR discovery learning OR cooperative learning OR constructivist teaching OR problem based learning OR Inquiry OR didactical engineering OR didactical learning OR didactical situations OR open approach OR problem based learning OR problem centred learning OR realistic mathematics education OR argumentation OR design OR project based learning
Competencies	21 st century skills OR Competence-based assessment OR Competency based learning OR Key competences
Subjects	STM OR STEM Mathematics OR Maths OR Math Science OR Physics OR Biology OR Chemistry Technology OR information communication technology OR information technology OR Computing

Appendix 2. Journals searched

Table showing the number of articles found in the targeted journals

Target Journal	Relevant articles found
British Journal of Educational Technology	2
Computer-Based Testing	0
Computers and Education	4
Education and Information Technologies	0
Educational Technology, Research and Assessment	1
European Journal of Education: special issue – ICT and Education	1
Frontiers in Artificial Intelligence and Information and Communication Technologies	0
International Encyclopaedia of Education (Technology and Learning - assessment)	0
International Journal of Computer-Supported Collaborative Learning	0
International Journal of E-assessment (Journal of the E-assessment Association)	3
International Journal of Educational Research	0
Journal of Applied Testing Technology	0
Journal of Computer Assisted Learning	2
Journal of Information Technology in Teacher Education	0
Journal of Research on Computing in Education	1
Journal of Science Education and Technology	3
Journal of Technology, Learning, and Assessment	0
Learning, Media and Technology	0
Research in Learning Technology (Journal of the Association of Learning Technology)	1

Appendix 3. Sources viewed with no relevant content

Citation/link
Cambridge Assessment website www.cambridgeassessment.org.uk
BECTA (2003) <i>What the research says about barriers to the use of ICT in teaching</i> BECTA ICT Research, Coventry www.becta.org.uk
Inspired by Technology, Driven by Pedagogy. A Systemic Approach to Technology-Based School Innovations. Centre for Educational Research and Innovation, OECD
http://www.mckinsey.com/Search.aspx?q=formative%20assessment Santiago, P., McGregor, I., Nusche, D., Ravela, P. and Toledo D. (2012) <i>OECD Reviews of Evaluation and Assessment in Education: MEXICO</i>
http://isolveit.cast.org/home iSolveit mathematics puzzles to develop logic and reasoning skills. iPad based apps with very limited focus not related to curriculum, more like puzzles e.g. sudoku.
TAO (Testing Assiste par Ordinateur) http://www.taotesting.com/ Open source e-testing platform from MCQs to simulations. Not relevant for IBE.
Maths.org. No relevant content (except link to NRich which has been explored).
Interactive Teaching and ICT, Swansea Metropolitan University. More concerned with ICT and interactive whiteboards than our remit.
International E-learning Association
Computer Aided Learning Conference
OECD Innovative Learning Environments Project 2010
Quest Atlantis – great resource (educational tasks in gaming environment) but relates to teaching and learning and not to assessment.
MacArthur Foundation http://www.macfound.org/programs/learning/ focus on learning not assessment
e-asTTle project, New Zealand. Mathematics online assessments: on-screen versions of paper tests. Teacher chooses level, length curriculum strands, 1st stage's level chosen by teacher, 2nd stage adaptive, MCQs and short answer Qs.
Journal of Research in Science Teaching: vol 48, pp 1050-1078 <i>Student Learning in Science Simulations: Design features that promote learning gains</i> . Focus on virtual laboratories and science-simulation software with no reference to assessment or IBE.
Doorey, A. How 2 common core assessment consortia were created and how they compare. (December 2012/January 2013) Educational Leadership.

Sets out plans for project, no new information for this project.

BECTA: Closed in 2011.

Udacity www.udacity.com: free interactive courses (lots re computer science and mathematics) Comprise short video lectures plus integrated quizzes (non-adaptive). Virtual field-trips, forums with peers. More focused on the learning than the e-assessment model.