LEARNING PROGRESSIONS AND ONLINE FORMATIVE ASSESSMENT NATIONAL INITIATIVE

LITERATURE REVIEW: FORMATIVE ASSESSMENT EVIDENCE AND PRACTICE

acara AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

Education Services Australia

aitsl Australian Institute for Teaching and School Leadership Limited

Contents

TABLES

FIGURES

EXPERT PANEL

Dr Rod Lane (Chief Investigator) – Deputy Head of Department – Learning & Teaching

Professor Rauno Parrila (Chief Investigator) – Professor of Literacy (Reading)

A/Prof Matt Bower – Associate Professor of Technology-Enhanced Learning

Professor Rebecca Bull – Professor of Numeracy

A/Prof Michael Cavanagh – Associate Professor of Mathematics Education

Dr Anne Forbes – Senior Lecturer in STEM Education

A/Prof Tiffany Jones – Associate Professor in Sociology of Education

Dr Maryam Khosronejad – Education (Professional Development)

Dr David Leaper – Assessment Analysis and Development

Professor Liz Pellicano – Professor of Austism Education

Dr Sarah Powell – Creative Arts (Music & Dance)

Professor Mary Ryan – Professor of Literacy (Writing)

Dr Iliana Skrebneva – Senior Research Assistant

RECOMMENDED FULL CITATION

Lane, R, Parrila, R, Bower, M, Bull, R, Cavanagh, M, Forbes, A, Jones, T, Leaper, D, Khosronejad, M, Pellicano, L, Powell, S, Ryan, M, and Skrebneva, I (2019) Formative Assessment Evidence and Practice Literature Review. AITSL: Melbourne

ABBREVIATIONS AND/OR GLOSSARY

Term	Definition
Adaptive assessment	Assessment items and the nature of the feedback generated are based on the learner's current ability.
AITSL	Australian Institute for Teaching and School Leadership
ANCOVA	Analysis of covariance. The output of the test shows the effect of an independent variable where the influence of covariants are removed.
ANOVA	Analysis of variance. Also known as the Fisher analysis of variance. Tests two or more means for differences.
Assessment as learning	A type of formative assessment which focuses on teaching students the metacognitive processes to evaluate their own learning and make adjustments.
Assessment for learning	A type of formative assessment that is used by teachers to gain an understanding of their students' knowledge and skills in order to guide instruction.
ASSISTments	A web-based Mathematics cognitive tutor. Like the adaptive systems described above, ASSISTments scaffolds problems into requisite skills and knowledge components.
Benchmark interim assessment	A comparison of student understanding or performance against a set of uniform standards within the same school year. It may contain hybrid elements of formative and summative assessments, or a summative test of a smaller section of the curriculum.
BYOD	Bring your own device
CAI	Computer-assisted interventions
CAT	Consequential Assessment Technique
CBM	Curriculum-based measurement
CCT	Classroom connectivity technology
CDDRE	Center for Data-Driven Reform in Education
CEM	Curriculum-embedded measures
COCA	Concepts of Comprehension Assessment
Cognitive conflict	A psychological state involving a discrepancy between cognitive structures and experience, or between various cognitive structures (that is, mental representations that organise knowledge, beliefs, values, motives, and needs). This discrepancy occurs when simultaneously active, mutually incompatible representations compete for a single response.
Cognitive model	Model outlining the prerequisite cognitive and learning skills underlying successful progression. For example, does the process require a significant amount of working memory, attention, motivation, persistency, cognitive ability, language skills, etc?
Conceptual change	Learning that involves the fundamental restructuring of students' pre-instructional ideas.
Cronbach's alpha	Cronbach's alpha is a measure of internal consistency; that is, how closely related a set of items are as a group.

Curriculum embedded measures	Formative assessments of recently taught content or skills designed to provide information that can guide instructional modifications for individual students.
DAT	Diagnostic Assessment Tools
Data literacy	Data literacy is the ability to read, work with, analyse, and argue with data. Much like literacy as a general concept, data literacy focuses on the competencies involved in working with data. Some of the competencies necessary to be able to demonstrate data literacy include: developing a habit of mind and practice regarding data use; using inquiry processes; asking significant questions, collecting and organising data; knowing and understanding data properties; putting data in context (using pedagogical content knowledge); synthesising, probing and prioritising data; and transforming data into application.
Demonstrate	To show or make evident knowledge and/or understanding.
DER	Digital Education Revolution
DFI	Dynamic Forecasting Intervention
Direct instruction	Teacher-centred pedagogy in which the teacher explicitly communicates a description of the concept to be learned or steps undertaken in a given practice to be learned
ERI	Early Reading Intervention
Experimental design	Experimental designs are used to examine the effect of a treatment or intervention on some outcome. In the simplest two-group case, a treatment is implemented with one group of participants (the treatment group) and not with another (the control group).
FAPD	Formative assessment professional development
FAST-R	Formative Assessments of Students Thinking in Reading
Feedback	Informational response or information on reactions to an individual's, group's or organisation's performance (including the performance of a task or explanation of an idea for example); intended as a basis for improvement.
Fidelity	The degree to which an intervention or program is delivered as intended.
Formative assessment	Formative assessment is any interaction that generates information on student learning which is then used by teachers and students to inform teaching and learning content and strategies.
Formative diagnostic assessment	A process of questioning, testing, or demonstration used to identify how a student is learning, where the student's strengths and weaknesses lie, and potential strategies to improve that learning. It focuses on individual growth.
I-Gmath	A synchronous peer-tutoring system on mobile tablet devices.
Interim assessments	Interim assessments are periodic diagnostic assessments typically administered three or four times during the school year to help teachers use evidence to differentiate instruction and make better instructional decisions, often in preparation for year-end summative assessments used as accountability measures.
ISI	Individualising Student Instruction intervention
ISI/A2i	A large-scale US software and instruction development project focused on improving K–3 students' reading skills.

Learning progression	Learning progressions, also known as progress maps, developmental continuums and learning trajectories, describe research-based, descriptive continuums of how students develop and demonstrate deeper, broader, or more sophisticated understanding over time.
Likert-type scales	A tool for measuring attitudes by asking people to respond to a series of statements about a topic, in terms of the extent to which they agree with them, and so tapping into the cognitive and affective components of attitudes.
MAP	Measures of Academic Progress
Mastery learning	The mastery learning model involves four components: defining mastery, planning for mastery, teaching for mastery, and grading for mastery. Formative assessments are used to provide both students and teachers with feedback about whether a particular instructional goal has been mastered. Students who do not meet the criteria for mastery are given correctives, such as alternative textbook readings, workbooks, or other varied learning tools. On completion of the correctives, the students take a second formative assessment. If they fail this test, they are given additional opportunities to study. Virtually all students achieve mastery before moving to the next unit.
mCLASS	A universal screener that measures the development of reading skills of all students in grades K–5 through two main assessments: Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and the Text Reading Comprehension (TRC) assessments.
Meta-analysis	A quantitative statistical analysis that is applied to separate but similar experiments of different and usually independent researchers and that involves pooling the data and using the pooled data to test the effectiveness of the results.
Metacognition	Metacognition is alertness to and reflection upon one's own or others' thought processes in learning and other contexts – or to put it more simply, thinking about thinking processes.
Metatalk	The use of language knowledge to have metalinguistic conversations.
NAEP	National Assessment of Educational Progress
NAPLAN	National Assessment Program in Literacy and Numeracy
NCLBA	*No Child Left Behind Act*
NPT	Nonreciprocal peer tutoring
OGT	Ohio Graduation Tests
PAL	Peer-assisted learning
PARS	Personalised Assessment Reporting System
Pedagogical content knowledge (PCK)	Pedagogical content knowledge generally refers to teachers' expertise both in the specific subject being taught (its information and skills) and in the best methods for teaching various components of the subject being taught (including different methods for the different information and skills in question, as needed).
PISA	Programme for International Student Assessment
PMA	Progress monitoring assessments so known as learning progress assessments (LPA)
POE	Predict–Observe– Explain
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
Progress monitoring	Progress monitoring is monitoring (over time) a student's or whole class's learning progress. It involves the gathering and evaluation of data to ascertain both the responsiveness to and effectiveness of classroom practice on the student's or class's learning goals and outcomes.

PSADRI	Problem-solving Assessment, Diagnosis, and Remedial Instruction
Reliability	Consistency or dependability of test performance across occasions, scorers and specific content.
RPTMC	Reciprocal peer-tutoring-enhanced mathematical communication
Rubric	Rubrics are assessment tools with three characteristics: a list of criteria for assessing the important goals of the task; a scale for grading the different levels of achievement; and a description for each qualitative level.
Scripts	Scripts offer specific questions structured in steps to follow an expert model of approaching a task from beginning to end. They are designed to analyse the process being followed during a task, although they can also be used to analyse the final outcome.
SDA	Student directed assessment process
Self-efficacy	A person's sense of being able to deal effectively with a particular task. Also, beliefs about personal competence in a particular situation.
SID	Scaffolding instructional discourse
Stratified randomised design	A method of sampling where the researcher divides the group into subgroups that are proportionate to the different strata.
Stratified randomised design	A research design that considers a number of different groups (strata) at the same time as randomly assigning groups within each stratum to control or treatment groups.
Summative assessment	Assessment of a test-taker's knowledge and skills typically carried out at the completion of a program of learning, such as the end of an instructional unit.
Systematic review	A systematic review is a synthesis and evaluation of primary research papers. It's conducted using a rigorous and clearly detailed methodology for the approach to searching and the process of selecting appropriate studies.
Task model	Model showing the sequence of activities to be successfully completed to meet learning outcomes and how learners typically progress through them (learning progression).
TTCT	Torrence Tests of Creative Thinking

Executive summary

Formative assessment has been defined as any interaction that generates data on student learning and is used by teachers and students to inform teaching and learning, to address specific student learning difficulties and to support learning growth over time.

A wide variety of assessment strategies, tools and resources currently exist to support and improve teachers' capacities to collect and analyse reliable data on student achievement and to adjust their teaching to meet each student's needs for enhanced learning outcomes. There is a widespread assumption in the academic literature that formative assessment leads to better learning outcomes for students.

Presently, we know little about the most effective ways of implementing formative assessment, including optimal school and educational system structures and supports. A systematic review of evidence is made more difficult by a lack of clarity and consensus regarding the nature and definition of formative assessment.

This review, commissioned by the Australian Institute for Teaching and School Leadership (AITSL), synthesises national and international research on the effective formative assessment practices of teachers and school leaders, including their current capacities, challenges and needs. It presents the findings of a review of peer-reviewed studies meeting robust experimental design criteria that examine formative assessment practices in Australian and international K–12 contexts. It further delivers an analysis and critical review of research relevant to formative assessment practices, including (but not limited to) the use of online assessment tools.

Key findings

Extent of Australian and international contemporary and seminal research on effective formative assessment practice (Chapters 1 and 2)

A comprehensive database search of studies relevant to formative assessment practices in K–12 contexts identified 5,867 studies. References were screened for quality and relevance (see Chapter 2). Further screening of papers against the tight inclusion criteria by five discipline experts reduced the pool of included papers to 71. Findings from these studies were then analysed to answer the review's focus questions. The authors note there are few rigorously designed experimental studies on formative assessment's impact on student learning, especially in Writing and The Arts.

Language, models and definitions of formative assessment (Chapter 1)

The conception of formative assessment has broadened over the past 50 years, from the notion of formative evaluation (Scriven 1963) to a broad range of practices ranging from process-oriented 'formative learning assessment' to 'instrument-based formative assessment.' Unlike summative assessment, there is currently no agreed-upon definition of formative assessment, nor does formative assessment represent a well-defined set of practices. The varying definitions have made it difficult to compare studies of formative assessment's effectiveness.

This review describes two periods of development in the definition of formative assessment: (1) the 1960s to the mid 2000s, when a loose consensus was developed and, (2) the period of conceptual confusion from 2000 onwards, since when the boundaries of formative assessment have continued to expand.

Sparks' (2015) model of assessment types is presented as a potential solution to these definitional issues. Sparks differentiates formative learning assessment from formative diagnostic assessment, benchmark interim assessment and summative assessment. Chapter 1 concludes with an explanation of the formative assessment's domain general principles and an overview of the benefits of using technology to support formative assessment.

Findings for the impact of formative assessment on teaching practice and student learning progress/outcomes are mixed (Chapters 3–7)

Research shows that formative assessment in different fields is often non-transferable and, when poorly designed, can lead to inaccurate information and ill-conceived pedagogical responses (Bennett 2011). The existing meta-analyses also have limitations that need to be considered. A number of these meta-analyses contain questionable claims about effect sizes that either summarise research too disparate to be synthesised meaningfully or are based on methods whose details were not published (Bennett 2011). To address this in the current review, we developed an evaluation framework for assessing the quality and rigour of formative assessment studies and meta-analyses and involved measurement and discipline experts in the evaluation and synthesis of this research.

Where studies report significant benefits, some isolate those benefits to specific groups of children (low achieving, high achieving). Others lack a sufficiently large sample size to look at effects for specific groups of students. Further, frequent omission of details regarding control group activities make determining formative assessment's impact on student learning more difficult. When studies involve a wholesale change to instruction (including changes to resources, professional development, time on task etc.), it is impossible to isolate formative assessment's influence. Necessary information regarding fidelity of implementation, instructional/cognitive models and learning progressions is also missing from many studies.

Greater benefits do appear where targeted, individualised feedback is provided instantaneously and more frequently. Studies show a small number of yearly assessment points is not sufficient in providing timely feedback to students on specific learning tasks. Results suggest frequent and embedded formative feedback may be critical to effective implementation of formative assessment practices.

The effectiveness of particular tools and resources (including online tools) that support teachers' professional judgements of learners' needs and the implementation of formative assessment practices in school settings (Chapters 3–7)

We found formative assessment tools and resources (online or otherwise) are most likely to be effective when based on:

1. a valid task model showing the sequence of activities to be successfully completed to meet learning outcomes and how learners typically progress through them (learning progression)
2. a valid cognitive model outlining the prerequisite cognitive and learning skills underlying successful progression (e.g. does the process require a significant amount of working memory, attention, motivation, persistency, cognitive ability, language skills etc?).

Studies in which formative assessments were followed by evidence-based interventions tended to produce better results.

Many of the reviewed studies included some technology/software component (e.g. mCLASS, ISI/A2i for Reading) that is likely to be helpful in improving student learning outcomes. Studies with rigorous experimental designs and controls are, however, required to validate these causal links. Existing studies suggest there is a good reason to believe that experimental studies would show the positive effect of using technology for formative assessment.

Research showed a mixed picture regarding the use and effectiveness of online formative assessment interventions across different disciplines. The impact of technology-driven formative assessment on student learning outcomes depends on various factors, such as the characteristics of learners (low achieving, high achieving), study sample size, experimental design and tool selection. The Using Sources Tool, for example, was found useful in facilitating teachers' evaluation of student Writing tasks. Progress monitoring, in particular curriculum-based measurement (CBM) mazes, are shown to be effective in improving student reading. Adaptive computerised programs can be effective in improving learning in subjects where there is a clearly identifiable skill hierarchy and relationships between skills.

Some outcomes, however, are difficult to assess using online tools, including 'creativity' in The Arts. A variety of assessment tools including observation, student–teacher collaboration and self-, peer and teacher feedback are required to effectively improve these skills.

The general principles for the effective application of formative assessment employment also apply to online assessment. Basing online assessments on a valid model of task components and the prerequisite cognitive and learning skills underlying successful progression is crucial. Interventions need to be evidence-based and aligned with validated learning progressions for the targeted concept or skill. We also found elaborate feedback with prompts is generally more effective than feedback that only recognises errors or provides correct answers.

Our research found optimal use of computer-based formative assessment is dependent on teachers' pedagogic preferences and orientations. When using technology to support formative assessment, it is vital that teachers have the requisite hardware and software knowledge and skills, and that formative assessment using digital technologies is supported and integrated within regular classroom activities. Teachers also need ongoing professional development to administer the assessments, interpret the results, make valid inferences and translate the information obtained to effective instruction.

Features of effective formative assessment professional development (Chapter 8)

Attitudes to formative assessment can create barriers for implementation, especially when ideas and practices are incompatible with teachers' current views. Ongoing support is required if new ideas and strategies are to be effectively implemented. The research shows brief interventions, such as short-term, product-oriented workshops, are less likely to effectively change practice. By contrast, long-term, process-oriented professional development with ample opportunities for collaboration, feedback and discussion appears more effective in successfully changing teachers' classroom assessment practices. Professional learning that is work-embedded and situated within school needs is preferred over one-day workshops or formally presented interventions.

Several studies reported that practice-centred collaboration, often in the form of school-based professional learning communities, is a critical ingredient for effective formative assessment practices. Professional development is most effective when teachers actively engage in instructional inquiry as part of a collaborative professional community that is focused on instructional improvement and student achievement. Collectively sharing developmental work within the school site or across networks is also an important factor in formative assessment professional development success.

The optimal school and education system structures, supports and conditions for effective implementation of formative assessment practices, including implementation of assessment tools, and any barriers to their effective implementation (Chapter 8).

We found that environmental conditions and teacher-level factors both play a role in effective implementation of formative assessments. Research highlights the need for school leaders who understand formative assessment, can provide a rationale for its use and can create a supportive and non-threatening environment where the effective use of assessment data is modelled for staff. Leaders who can establish a schoolwide formative assessment culture with vision and expectations for assessments and a school climate promoting trust, mutual respect and cooperation will create the best environment for formative assessment success. This should further be reinforced with high-quality professional development and effective support for formative assessment implementation and a commitment to giving teachers regular, protected meeting times for meaningful examination of assessment practices.

It is important that accountability pressures on teachers do not lead to unintended impacts on instruction and assessment practices. Rather, decentralised organisational structures and distributed school leadership should focus on building a broader base of engagement and expertise, and a greater sense of shared vision and ownership. Strategically aligning expertise and resources to support teachers' learning about effective practice is necessary for achieving optimal implementation.

Teachers require both assessment knowledge and data literacy to effectively implement formative assessment. Increased focus on assessment literacy in initial teacher education and in-service teacher PD is therefore needed. Promoting a classroom philosophy that regards mistakes as opportunities to learn and encourages honest reflection is key to achieving better results for students. Implementing effective formative assessment also requires sound pedagogical content knowledge, so that teachers can break down critical concepts, find appropriate entry points for all students, and redesign instruction to match students' assessed understandings and misconceptions.

1 Chapter 1: Introduction

1.1 Background to the literature review

The *Through Growth to Achievement: Report of the Review to Achieve Educational Excellence in Australian Schools* (Excellence Review) (Gonski et al. 2018) recommends the development of a new online, on-demand assessment tool aligned with Australian Curriculum Learning progressions and supported by targeted professional development. A comprehensive review of national and international literature on formative assessment (or assessment for learning) is required to inform the discovery phase of this project.

An extensive body of literature outlines the purported benefits of formative assessment for student learning. Recent reviews, however, highlight a number of issues with these studies. They include a failure to accurately define the distinctive features of assessment and formative assessment, under-representation of measurement principles in the conceptualisation of formative assessment and a lack of understanding of the domain-specific nature of effective practice. Key research shows formative assessment in different fields is often non-transferable and, when poorly designed, can lead to inaccurate information and ill-conceived pedagogical responses (Bennett 2011).

There are also limitations to existing meta-analyses to be considered in a thorough review of the extant literature on formative assessment. A number of these meta-analyses' claims about effect sizes are questionable because they either summarise research too disparate to be synthesised meaningfully or fail to publish their detailed methods (Bennett 2011). A key implication of this is the need to: (1) develop an evaluation framework for assessing the quality and rigour of formative assessment studies and meta-analyses, and (2) involve measurement and discipline experts in the evaluation and synthesis of the research. Detailed aims and methodology for this evaluation and synthesis process are outlined below.

1.2 Aims of the literature review and key questions addressed

This literature review will evaluate and synthesise national and international research on the effective formative assessment practices of teachers and school leaders, including the use of online assessments. The review will address:

a. Australian and international contemporary and seminal research on effective formative assessment practices
b. language, models and definitions of formative assessment
c. the evidence base for the impact of formative assessment on teaching practice and student learning progress/outcomes
d. the role of teachers and school leaders in effectively implementing formative assessment practices
e. the use and impact of tools and resources that support teachers' professional judgements of learners' needs and the implementation of formative assessment practices in school settings
f. the optimal school and education system structures, supports and conditions for effective implementation of formative assessment practices, including implementation of assessment tools and any barriers to their effective implementation.

The review aims to provide an evidence base to inform AITSL's capacity building for teachers and school leaders. This includes the proposed development of an online, on-demand assessment tool and tailored teaching resources to maximise student learning growth, professional learning for building the assessment capacity of teachers and school leaders, and research evidence summaries and translations. These resources have the potential to help Australian schools develop a shared understanding of formative assessment, and more effectively identify and respond to students' learning progression needs (Gonski et al. 2018; Cawsey et al. 2019).

1.3 Frame for the literature review

Formative assessment is a critical component of effective and responsive classroom instruction. It can provide teachers, school leaders and learners with accurate information about: (1) the learner's current performance level, (2) gaps in their learning outcomes, (3) growth in their learning across time, (4) specific difficulties they experience with current learning tasks, and (5) possible reasons for those difficulties. The goal of formative assessment is to provide information that leads to more effective teaching and learning practices than would be possible without it.

An understanding of a curriculum's instructional goals is needed to provide valid information on whether a learner has reached the required level of performance and if there are any gaps in their learning. In this sense, any summative assessment that sufficiently covers the desired learning outcomes can be used formatively when administered during the course of instruction and followed by different instructional choices for those who have achieved the learning outcomes and those who have not.

For example, a low score on a spelling test can show lack of learning and be interpreted as a sign more instruction is needed. However, a thorough understanding of the different component skills needed for successful performance is required to identify a specific locus of difficulties. For this example, we could design items requiring knowledge of only simple phoneme-grapheme correspondences, more complex phoneme-grapheme correspondences, and word-specific spellings (and, therefore, prior exposure to them). Different performance patterns across the items would then suggest multiple foci for further instruction.

To design a formative assessment that generates valid information about task components requires a model – usually theoretical – of what those components are and how learners typically progress through them. It also requires knowledge of effective instruction in different components. Finally, a carefully designed sequence of formative assessments and instruction can provide diagnostic information about the learner beyond task-component knowledge, such as cognitive-component knowledge. For example, if an initial formative assessment of spelling shows some students likely lack knowledge of specific phoneme-grapheme correspondences, instruction can focus on teaching those correspondences based on the instructional knowledge that many students will benefit from direct instruction. If the subsequent follow-up formative assessment shows only some students benefited from direct instruction despite full attendance and attention, this assessment–instruction–assessment sequence can indicate that a different approach is required or that the instruction needs to focus first on underlying prerequisite learning skills (such as phonemic awareness) before content learning is successful.

Valid inferences at this level require both a model of task components and a model of the prerequisite cognitive and learning skills underlying successful progression in spelling. To design formative assessment–instruction sequences of this kind, and to follow them up with appropriate summative assessments verifying or falsifying the hypothesised instructional or cognitive difficulties, requires an understanding of the curriculum, each learning task, the cognitive models of learning in the domain, the effective instruction to match various learning needs (differentiated instruction), and valid assessments.

What is clear from the above is that a review of formative assessments requires substantive expertise in both assessment and in the disciplines reviewed. To evaluate and compare the effectiveness of different approaches to formative assessment, one must first situate exemplar studies in one or more of the above levels. To the extent they fall beyond assessment of performance levels, one must also examine the task and the cognitive and instruction models that either implicitly or explicitly guided the assessment design and the inferences made on the basis of the results. In particular, where results are less than impressive, this level of precision is necessary to suggest the locus of failure. For example, in a spelling study, failure possibly emanates from an outdated task-component model or an ineffective instructional model. Only by understanding both models can we make valid comparisons between studies. Therefore, our proposed approach to reviewing formative assessment literature relies as much on assessment expertise as on the expertise of subject area specialists who can integrate assessment knowledge with task, cognitive and instruction model knowledge and theories in their fields.

1.4 Language, models and definitions of formative assessment

1.4.1 Introduction

The relative advantages and disadvantages of formative and summative assessment have been discussed since Scriven first coined the term ‘formative evaluation’ in 1967. Misconceptions about the dichotomy between these two forms of assessment mean formative assessment is seen as a tool for ‘improving learning’, and summative assessment as a static measure of learning (Lau 2016). Summative assessment has also become synonymous with large-scale testing programs such as the NAEP (National Assessment of Educational Progress) in the United States, the Key Stage SATs (Standard Attainment Tests) in the UK, and NAPLAN (National Assessment Program – Literacy and Numeracy) in Australia. These assessments are used by policy makers to monitor the educational system’s progress and ensure national standards are maintained. Such assessments often attract media attention that focuses on the level of student performance as a whole and may be critical of the distorting influence the tests have on the delivered curriculum. In Australia, coverage of these issues coincides with the dates of national assessments and the release of results.

During the 1990s, similar concerns about the dominance of summative assessment led to the formation of the Assessment Reform Group (ARG) in the UK, which aimed to promote evidence-based educational decision-making. The ARG funded a literature review of more than 250 formative assessment studies, which was published by Black and Wiliam (1998a, 1998b). Both articles were highly influential in the UK and internationally. Black and Wiliam defined formative assessment as occurring ‘when the evidence [of learning] is actually used to adapt the teaching to meet the student needs’ (1998a, p. 140). Their conclusions strongly supported implementation of formative assessment programs to improve education quality:

> There is a body of firm evidence that formative assessment is an essential component of class room work and that its development can raise standards of achievement. We know of no other way of raising standards for which such a strong *prima facie* case can be made. Our plea is that national and state policy makers will grasp this opportunity and take the lead in this direction (Black & Wiliam 1998a, p. 148).

This endorsement of formative assessment found a ready audience in the US as the *No Child Left Behind Act* (NCLBA) of 2002 was being implemented. In the context of this burgeoning interest, the label ‘formative assessment’ was applied to an ever-broadening range of practices, leading to confusion over its meaning.

The purpose of this section is to examine definitions of formative assessment to gain a clear sense of what formative assessment is and its scope. First, we show how the term developed up to the early 2000s, before depicting how 'formative assessment' became attached to a range of assessments. We then examine the benefits of using technology to deliver formative assessment.

1.4.2 The development of formative assessment (1960s – mid-2000s)

The distinction between 'summative' and 'formative' evaluation was first made by Scriven (1963); however, it was Bloom, Hastings and Madus' (1971) interpretation that led to their popularity in assessment discourse. For Bloom, Hastings and Madaus, formative evaluations 'provide feedback to students on their learning of particular portions of the learning sequence', while summative evaluations are 'used at the end of the course, term or educational program'. Over time, the term 'evaluation' was replaced with 'assessment' to emphasise that the focus is on students rather than programs (Bennett 2011).

Since Bloom et al.'s work (1971), a general consensus has developed regarding the definition of summative assessment. The Standards for Educational and Psychological Testing define summative assessment as an 'assessment of a test taker's knowledge and skills typically carried out at the completion of a program of learning, such as the end of an instructional unit' (2014, p. 224). Until the mid-2000s, there were also signs a broad consensus could be reached on the definition of the term 'formative assessment'.

1.4.3 The period of conceptual confusion (2000s –)

Since the mid-2000s, the term 'formative assessment' has been associated with an ever-widening array of assessment practices, causing confusion about its definition. According to Gewertz (2015, para. 1), asking five teachers would probably result in 'five different answers'. The reason for this conceptual confusion has been ascribed to poor timing (Shepard 2005, p. 2), as formative assessment came to the attention of US educators at the same time as the NCLBA (2002). The Act linked federal funding to student performance on state-mandated assessments, putting extreme pressure on educators to improve student performance, and making them receptive to the promise heralded by influential research on formative assessment (Black & Wiliam 1998a, 1998b).

Black and Wiliam's (1998a) conclusion that formative assessment helped students to master course content and significantly boost their scores on external achievement tests was potentially attractive to US educators. Perhaps most alluring was the finding that formative assessment could boost learning gains by 'typical' effect sizes of between 0.4 and 0.7. As Black and Wiliam put it, 'an effect size gain of 0.7 in the recent international comparative studies in Mathematics would have raised the score of a nation in the middle of the pack of 41 countries (for example, the US) to one of the top five' (1998a, p. 140). Rather than seeing a thoughtful integration of Black and Wiliam's (1998a) findings, US educators, desperate to see the improvements demanded by the NCLBA, grasped at formative assessment as a panacea to lift the performance of their schools.

The popularity of formative assessment did not escape the notice of test-publishing companies. A number of publishing firms repackaged, or even merely relabelled, previously sold 'benchmark' or 'interim' tests as 'formative' (Popham 2006, p. 86). At the same time, the US Educational Testing Service launched the *ETS Formative Assessment Item Bank*, consisting of 11,000 questions from which teachers could construct formative assessment tests (Bennett 2011). The market boomed and commercial test publishers prospered. By the 2006–07 academic year, it was estimated expenditure on 'formative assessment' resources accounted for 30% of the $2.1 billion spent on assessment in the United States (Cech 2008).

The impact of these diverging forms of assessment being labelled ‘formative’ can be seen in the second edition of the *Handbook of Formative Assessment,* where Cizek, Andrade and Bennett lament that, since the first edition, definitional clarity ‘may even have degraded’ (2019, p. 8). The definitional issue can be summarised as a conflict between two groups: on the one side are those for whom formative assessment is a process, and on the other are those comfortable with it as an instrument (Bennett 2011). The former is composed mostly of scholars and educators, who view formative assessment as embedded in the instruction, and something to be developed by programs to improve teachers’ skills for effective implementation (Table 2 below is a product of this perspective). The latter group is comprised mostly of publishers, who look to complement educational programs by producing assessments aligned to the teaching cycle.

Scholars in the field reacted to the practice of formative assessment-as-an-instrument. Shepard (2005) maintains the term ‘formative evaluation’ should be used to refer to program-level assessments, and that the term ‘formative assessment’ should only apply to assessments that are closer to the instruction. Both Popham (2006) and Shepard (2005) assert benchmark/interim tests are not supported by formative assessment research. Popham (2006) objected to benchmarking or interim tests, arguing results could not be incorporated into the teaching cycle. That students and teachers could use feedback to inform instruction was seen as central to the definition of formative assessment, and ‘benchmark’ or ‘interim’ tests generally failed to return results in sufficient time for teachers to use them. According to Popham, these tests could not be considered formative since neither learning nor teaching was affected by feedback (Popham 2006, p. 86). In addition, for assessments to inform teaching they needed to be integrated into the curriculum – typically this was not the case (Shepard 2005). A final objection was that benchmarking and interim tests were too general to be useful at an individual level and could only inform ‘relatively gross instructional-program-level decisions’ (Shepard 2005, p. 6).

It should be noted that some of Popham's (2006) and Shepard’s (2005) objections appear to be matters of degree: if these benchmarking/interim tests could return results in time to be incorporated into the instructional cycle and the test items aligned with the curriculum, then they could be considered formative assessments. Also, the seemingly disparate perspectives of formative assessment as a process and as an instrument may not be as distant as they seem. To be effective, formative assessment must combine a process with an instrument (Bennett 2011). The challenge is how to make sense of these polarised positions on the definition of formative assessment.

1.4.4 Developing clarity around the types of assessment

Sparks’ (2015) model offers a useful summary of the spectrum of assessment types. Sparks identifies four main categories of assessment: formative learning assessment at one extreme, followed by formative diagnostic testing, interim/benchmark testing, and then summative assessment at the other extreme. Assessment types are differentiated according to questions such as, ‘Who is being measured?’, ‘How often?’, ‘For what purpose?’ and ‘What strategies are used?’ (Table 1). This effectively demonstrates that the previously considered strict dichotomy between formative and summative assessment is better approached as a matter of degree, depending on the assessments’ use.

While recognising the diversity of formative assessment and including both process-led (‘formative learning assessment’) and instrument-dominated approaches (‘formative diagnostic assessment’ and ‘benchmark/interim assessment’), Table 1 allows us to choose which aspect of formative assessment to focus on when seeking clarity. In the case of this literature review, we can now focus on the first two columns of the table. The definition categories of ‘formative learning assessment’ and ‘formative diagnostic assessment’ above align with earlier formative assessment definitions. Black and Wiliam maintain that assessments function formatively:

> ... to the extent that evidence about student achievement is elicited, interpreted and used by teachers, learners or their peers to make decisions about the next steps in instruction that are likely to be better or better founded than the decisions they would have taken in the absence of the evidence elicited (2009, p. 9).

Table 1: Types of assessment

Formative Learning Assessment	Formative Diagnostic Assessment	Benchmark/Interim Assessment	Summative Assessment
What is it?			
Formative learning is the process of teaching students how to set goals for their learning, to identify their growth towards those goals, to evaluate the quality of their work, and to identify strategies to improve.	Formative diagnostic assessment is a process of questioning, testing or demonstration used to identify how a student is learning, where her or his strengths and weaknesses lie, and potential strategies to improve that learning. It focuses on individual growth.	Benchmark or interim assessment is a comparison of student understanding or performance against a set of uniform standards within the same school year. It may contain hybrid elements of formative and summative assessments, or a summative test of a smaller section of content, such as a unit or semester.	Summative assessment is a comparison of the performance of a student or group of students against a set of uniform standards.
Who is being measured?			
Individual students measure themselves against their learning goals, prior work, other students' work and/or an objective standard or rubric.	Individual students. The way they answer gives insight into their learning process and how to support it.	Individual students or classes.	The educational environment: Teachers, curricula, education systems, programs etc.
How often?			
Ongoing: It may be used to manage a particular long-term project or be included in everyday lessons. Feedback is immediate or very rapid.	Ongoing: Often as part of a cycle of instruction and feedback over time. Results are immediate or very rapid.	Intermittent: Often at the end of a quarter or semester, or a midpoint of a curricular unit. Results are generally received in enough time to affect instruction in the same school year.	Point in time: Often at the end of a curricular unit or course, or annually at the same time each school year.
For what purpose?			
To help students identify and internalise their learning goals, reflect on their own understanding and evaluate the quality of their work in relation to their own or objective goals, and to identify strategies to improve	To diagnose problems in students' understanding or gaps in skills, and to help teachers decide next steps in instruction.	To help educators or administrators track students' academic trajectory toward long-term goals. Depending on the timing of assessment feedback, this may be used more to inform instruction or to evaluate the quality of the learning environment.	To give an overall description of students' status and evaluate the effectiveness of the educational environment. Large-scale summative assessment is designed to be brief and uniform, so there is often limited information to

their work and understanding.			diagnose specific problems for students.
What strategies are used?			
Self-evaluation and metacognition, analysing work of varying qualities, developing one's own rubric or learning progressions, writing laboratory or other reflective journals, peer review etc.	Rubrics and written or oral test questions, and observation protocols designed to identify specific problem areas or misconceptions in learning the concept or performing the skill.	Often a condensed form of an annual summative assessment, e.g. a shorter-term paper or test. It may be developed by the teacher or school, bought commercially, or be part of a larger state assessment system.	Summative assessments are standardised to make comparisons among students, classes or schools. This could a single pool of test questions or a common rubric for judging a project.

1.4.5 Domain general principles of formative assessment

In some instances, principles for effective formative assessment have been proposed. The first edition of the *Handbook of Formative Assessment* (Cizek 2010, p. 8) included a list of ten key elements extracted from definitions and models by leading researchers for their 'potential to maximise the achievement, development, and instructional benefits' of formative assessment (Cizek 2010, p. 7). Cizek makes it clear this is not a defining list and that a given formative assessment program need not include all items. An updated list provided by Cizek, Andrade and Bennett (2019) is shown in Table 2.

Table 2: Ten key elements of formative assessment

1	Focuses on goals that represent valuable educational outcomes with applicability beyond the learning context
2	Communicates clear, specific learning goals
3	Provides examples of learning goals including, when relevant, the specific grading criteria or rubrics that will be used to evaluate the student's work
4	Identifies student's current knowledge/skills and necessary prerequisites for the desired goals
5	Requires development of plans for attaining the desired goals
6	Includes frequent assessment, including student self-assessment, peer assessment, and assessment embedded within learning activities
7	Includes feedback that is non-evaluative, specific, timely, related to the learning goals, and provides recommendations for how to improve
8	Encourages students to self-monitor progress toward the learning goals
9	Promotes metacognition and reflection by students on their work
10	Encourages students to take responsibility for their own learning.

Wiliam and Thompson (2008) provide another useful framework for conceptualising formative assessment's domain general principles. Their book *Embedded Formative Assessment* provides five strategies they see as vital to successful formative assessment practice in the classroom. These strategies are based on a matrix examining three questions: (1) where is the learner going? (2) where is the learner now? and (3) how to get the learner there? Each of these questions can be answered from the perspective of the teacher, peers and learner (see Table 3).

Table 3: Five strategies vital to successful formative assessment practice

	Where learner is going?	**Where learner is now?**	**How to get learner there?**
Teacher	Clarifying, sharing and understanding learning intentions and criteria for success.	Engineering effective classroom discussions, activities, and learning tasks that elicit evidence of learning – developing effective classroom instructional strategies that allow for the measurement of success.	Providing feedback that moves learning forward – working with students to provide them the information they need to better understand problems and solutions.
Peer			Activating learners as instructional resources for one another – getting students involved with each other in discussions and working groups can help improve student learning.
Learner			Activating learners as owners of their own learning – teaching students to monitor and regulate their learning increases their rate of learning.

1.5 Effective use of technology to support and enable formative assessment

1.5.1 Benefits of using technology for formative assessment

Researchers generally acknowledge the important role that technology can play in supporting formative assessment processes (e.g. Bhagat & Spector 2017). Technologies like mobile devices, computers, tablets and online resources are increasingly common in schools (Kaware & Sain 2015), so researchers are accordingly interested in understanding and reviewing the efficacy of using digital technologies for formative assessment (Gikandi, Morrow & Davis 2011).

Moving beyond the use of simple closed-item responses, technological advancement has generated new opportunities for formative assessment. Technology-enhanced learning environments can provide tools and systems for creating learning situations requiring complex thinking, problem-solving and collaboration strategies and also allow for the assessment of these competencies (European Commission 2011). Innovative computer-based assessments, for example, can score student performances on complex cognitive tasks involving various cognitive processes, including open-ended performances such as written essays and student collaboration on constructed response formats (Roschelle et al. 2000).

Cited benefits of using technology include:

1. Greater diagnostic information – technology can be used to provide feedback not only about the quality of student responses but also about the pedagogical or learning strategies to address, prevent and correct misconceptions and skill deficits (van der Kleij et al. 2015). Mathematics teachers, for example, use problem-solving diagnosis and remedial instruction (PSADRI) systems to collect diagnostic details on student learning (Hsiao et al. 2017).
2. Ability to capture and assess procedural knowledge – technology can help track the full record of the problem-solving process adapted by the learner (Bennett & Gitomer 2009).
3. Efficiency and cost effectiveness – automated assessment can save time and reduce the number of errors (Bennett & Gitomer 2009). Automated scoring also constitutes a cost saving compared to human scoring processes.
4. More timely feedback to students – automating assessment processes enables students to receive feedback in faster and hence more frequent feedback cycles (van der Kleij et al. 2015). For instance, Mathematics teachers can use computerised dynamic adaptive tests for formative assessment and immediate feedback (Wu et al. 2017).
5. Adaptive feedback – using computerised adaptive tools enables assessment items and the nature of the feedback generated to be based on the learner's current ability (Veldkamp, Matteucci & Eggen 2011). Examples include formative assessment-based personalised web learning systems that provide hints and supplementary material/tasks based on students' mathematical knowledge levels (Wongwatkit et al. 2017).
6. Peer feedback and collaboration – technology can be used to facilitate peer feedback and collaboration. Online peer tutoring systems, for example, have a growing role in formative assessment in Mathematics (Chappell et al. 2015; Tsuei 2017; Yang et al. 2015).
7. Interactive learning – the use of visualisation techniques in digital technologies allows students to receive feedback and learn interactively (Bhagat & Spector 2017). Science students can use web-based interactive tools with intrinsic feedback (e.g. SmartGraphs) for exploring and understanding how to work with data in various visual formats (Zucker, Kay & Staudt 2014).
8. Improving learner self-regulation – online formative assessment tools can offer students the ability to effectively manage their behaviour and cognitive processes, such as monitoring academic progress, self-evaluation, time management, goal-setting, motivation, positive reactions to feedback etc. (McLaughlin & Yan 2017).
9. Easy access to student work – teachers can use technology to easily access student work during the assessment process, monitor student contributions and provide formative assessment feedback (McLaughlin & Yan 2017). For instance, software like mCLASS allows teachers to monitor individual student progress more frequently by including in-between assessment tasks in reading discipline (Konstantopoulos et al. 2013; 2016).

1.5.2 Conclusion

This review of definitional issues shows the evolution of our understanding about formative assessment over time and the potential benefits of technology to support and enable it. The impact of using technology for formative assessment and the nature of best practice is an ongoing area of investigation. The conception of formative assessment has broadened from the notion of formative evaluation (Bloom et al. 1971) to include a range of practices on a spectrum from process-oriented 'formative learning assessment' to 'instrument-based formative assessment'. Given the purpose of this report is to review *the evidence base related to the impact of formative assessment on student learning progress and outcomes,* we will limit our attention to rigorously designed studies that isolate the impact of formative assessment on student learning. The following chapter will present this review's methodology, including search terms used, inclusion/exclusion criteria, screening method and approach for synthesising research.

2 Chapter 2: Methodology

2.1 Overview

This chapter describes the review's literature search process, screening process, inclusion and exclusion criteria, and analysis approach. The review involved a structured process of formulating research questions, defining search terms, selecting databases, conducting the literature search, formulating inclusion criteria, applying these to selected relevant literature, and the extraction of data. We consulted a library professional to validate the search terms and strategy adopted for the review. Initial literature searches were conducted in May–June 2019.

2.2 Locating existing systematic reviews/meta-analyses on formative assessment

The starting point for this review was a meta-analysis of studies conducted between 1988 and 2014 by the US Department of Education (Klute et al. 2017) and focusing on formative assessment's effect on elementary school student achievement. Klute et al. (2017) identified the following 16 references as relevant to the current review: Andrade and Cizek (2010); Bennett (2011); Black and Wiliam (2009); Briggs et al. (2012); Brookhart (2010); Clark (2012); Filsecker and Kerres (2012); Heritage (2010, 2013); Kingston and Nash (2011, 2012); Marzano (2010); McMillan, Venable & Varier (2013); Moss and Brookhart (2009); Noyce and Hickey (2011); Wiliam (2011).

A library database search identified a further 12 systematic reviews and meta-analyses relevant to formative assessment in K–12 contexts. These included: Burns et al. (2010); Gersten et al. (2009); Graham, Hebert and Harris (2015); Haelermans, Ghysels and Prince (2015); Hartmeyer, Stevenson and Bentsen (2018); Heitink et al. (2016); Hellrung and Hartig (2013); Kingston and Broaddus (2017); Miller, Scott and McTigue, (2018); Rubie-Davies and Rosenthal (2016); Sanchez et al. (2017); Wang et al. (2016).

The bibliographies of each of the above reports were obtained and original cited papers located and reviewed. A manual search of key journals on assessment in each of the subject domains was undertaken to identify relevant studies exploring the impact of formative assessment on teaching practice and student learning. While these studies provided an important starting point, a broader search of the literature was required to identify more recent work conducted across K–12 contexts.

2.3 Search strategy

We followed guidelines suggested by the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) to provide a structure and process for the literature search and screening process (Moher et al. 2009). The review involved four phases: (1) a broad search to identify possible relevant studies, (2) the screening of identified studies at title, abstract and full-text levels filtered through an *a priori* set of inclusion criteria, (3) a quality analysis of all remaining articles using an *a priori* rubric of research quality indicators, and (4) a synthesis of the final corpus of studies (Torgerson 2007).

2.4 Databases used

Advice was sought from a senior librarian regarding the databases used for this search. We undertook a thorough and systematic search of six scientific databases to retrieve relevant literature: EBSCO (Education Resource Complete, Academic Search Premier, Business Source Premier, EconLit), ERIC (Education Research Information Centre); PsycINFO; ScienceDirect; Scopus, and A+education. These databases were selected to ensure all relevant journals addressing the inclusion criteria were searched.

2.5 Inclusion and exclusion criteria

The inclusion criteria used to search for and evaluate studies were:

a. **Study design and quality** – the studies needed to be empirical and use a group comparison research design (e.g. randomised experiments, non-randomised quasi-experiments, regression discontinuity designs) and single-subject research. Studies needed to include well-designed control conditions so that treatment effects could be confidently attributed to the formative assessment interventions. Correlational, qualitative and other descriptive research designs as well as theoretical papers, reviews and opinion pieces were excluded from the selection. Studies needed to be published in peer-reviewed journals, book chapters or conference proceedings.

b. **Language** – studies needed to be published in English.

c. **Timeframe** – studies published between January 2009 and June 2019 were included. The search was restricted to studies published in the past decade to ensure that the technologies discussed were relevant to current classrooms and school contexts. The pace of change in information and communication technologies has been rapid over the past ten years, driven by the implementation of policies and programs aimed at increasing students' access to and use of ICT, such as the government-funded Digital Education Revolution (DER) reform package (2008–13) and school ‘bring your own device’ (BYOD) programs (Dandolo Partners 2013).

d. **Student sample** – this review included formative assessment studies with students enrolled in Kindergarten to Year 12.

e. **Subject area context** – studies needed to focus on formative assessment (or assessment for learning) and its impact on teaching practice or student learning outcomes. Separate searches were undertaken to identify studies exploring optimal school and education system structures, supports and conditions for effective implementation of formative assessment practices.

2.6 Keywords used in literature search

Keyword searches of abstracts in the academic databases involved four steps:

Step 1: Keyword searches for terms related to formative assessment and research design. These search phrases were informed by Klute et al.’s (2017) previous work:

- ‘formative assessment*’ OR ‘formative learning*’ OR ‘formative teach*’ OR ‘assessment for learning’ OR ‘diagnostic assessment*’ OR ‘formative evaluation*’ OR ‘mastery learning’ OR ‘classroom questioning’ OR ‘curriculum-based assessment*’ OR ‘curriculum-based measurement’ OR ‘curriculum-embedded measure’ OR ‘curriculum-embedded measurement’ OR ‘progress monitoring’ OR ‘learning progress assessment.’
- The words (or variants of) ‘random*’ OR ‘RCT Quasi-experiment*’ OR ‘QED Experiment*’ OR ‘impact’ OR ‘effectiveness’ OR ‘causal post-test’ OR ‘post-test’ OR ‘pre-test’ OR ‘pre-test’ OR ‘Efficacy trial’, OR ‘Multisite’ OR ‘multi*site’ OR ‘Comparison Treatment’ OR ‘Control group*’ OR ‘comparison group*’ OR ‘matched group*’ OR ‘Equivalence’ OR ‘Baseline’ OR ‘Propensity score’. These terms had to appear in the abstract.
- Keyword searches in this first phase were restricted to studies published between January 2009 and December 2014 in primary and secondary school contexts. This was necessary because Klute et al.’s study (2007) was restricted to research undertaken with elementary school students. By combining the research from Klute et al. with the current search we were able to locate all relevant papers addressing formative assessment in K–12 contexts (2009–14).

Step 2: The second step was to locate relevant literature from 2014 to 2019. It involved the same set of keywords used in Step 1 but focused on papers published between 2014 and 2019. It included studies in elementary, primary and secondary school contexts (K–12).

Step 3: The third step was designed to locate studies relevant to formative assessment in online contexts. Search terms included:

- ‘formative assessment*’ OR ‘formative learning*’ OR "formative teach*’ OR ‘assessment for learning’ OR ‘diagnostic assessment*’ OR ‘formative evaluation*’ OR ‘mastery learning’ OR ‘classroom questioning’ OR ‘curriculum-based assessment*’
- The words (or variants of) ‘online assessment tool’ OR ‘electronic assessment tool’ OR ‘mobile assessment tool’ OR ‘virtual assessment tool’ OR ‘digital assessment tool’ OR ‘on-line assessment tool’ OR ‘on demand assessment tool’ OR ‘adaptive assessment tool’ OR ‘web-based assessment tool’. These terms had to appear in the abstract.

The same inclusion criteria (in terms of time and student sample) were applied to this search.

Step 4: This step involved a search for studies exploring optimal school and education system structures, supports and conditions for effective implementation of formative assessment practices.

The search included the following terminology:

- ‘formative assessment* OR ‘formative learning* OR ‘formative teach* OR ‘assessment for learning” OR ‘diagnostic assessment*’ OR ‘formative evaluation*’ OR ‘mastery learning’ OR ‘classroom questioning’ OR ‘curriculum-based assessment*’
- The words (or variants of) ‘role of teacher*’ OR leader* OR principal* OR admin*OR ‘role of admin*’ OR ‘role of headm* OR ‘professional development’ OR ‘teacher development’ OR implement*. These terms had to appear in the abstract.

Searches were conducted sequentially, with overlapping (duplicate) documents excluded from each subsequent search.

2.7 Results of the database search

The number of studies identified in each search is summarised below and in the PRISMA diagram (Figure 1).

- Step 1 – 1,539 references identified.
- Step 2 – 3,127 references identified.
- Step 3 – 1,639 references identified.
- Step 4 – 1,057 references identified.

An additional 30 relevant studies were located by scanning the reference lists of papers identified by the database searches. This resulted in a total of 7,392 references. When duplicates were excluded, only 5867 references remained. These references were exported to Endnote Version 9 software for further analysis.

2.8 The screening process

The next stage involved the project team's five discipline experts screening papers against the five inclusion criteria in Section 2.5. Screening was conducted in three phases. In Phase I, reviewers focused on report titles and abstracts, screening for topic, sample, timeframe and outcome relevance. Reviewers were instructed to keep any article that could potentially meet the inclusion criteria. In Phase II, reviewers read the papers in full, identifying studies that met the inclusion criteria for the review. After the Phase II screening process, 129 articles were identified as meeting the inclusion criteria. In Phase III, reviewers categorised these 129 studies by domain area: Mathematics, Reading, Writing, Science and The Arts. Studies relating to professional development and optimal characteristics were classified separately. Four of these six disciplines (Mathematics, Reading, Writing, and Science) included research involving online/digital formative assessment tools. Of the 129 studies identified via screening as eligible for evidence review, discipline experts in each subject area determined that 71 studies met the standards.

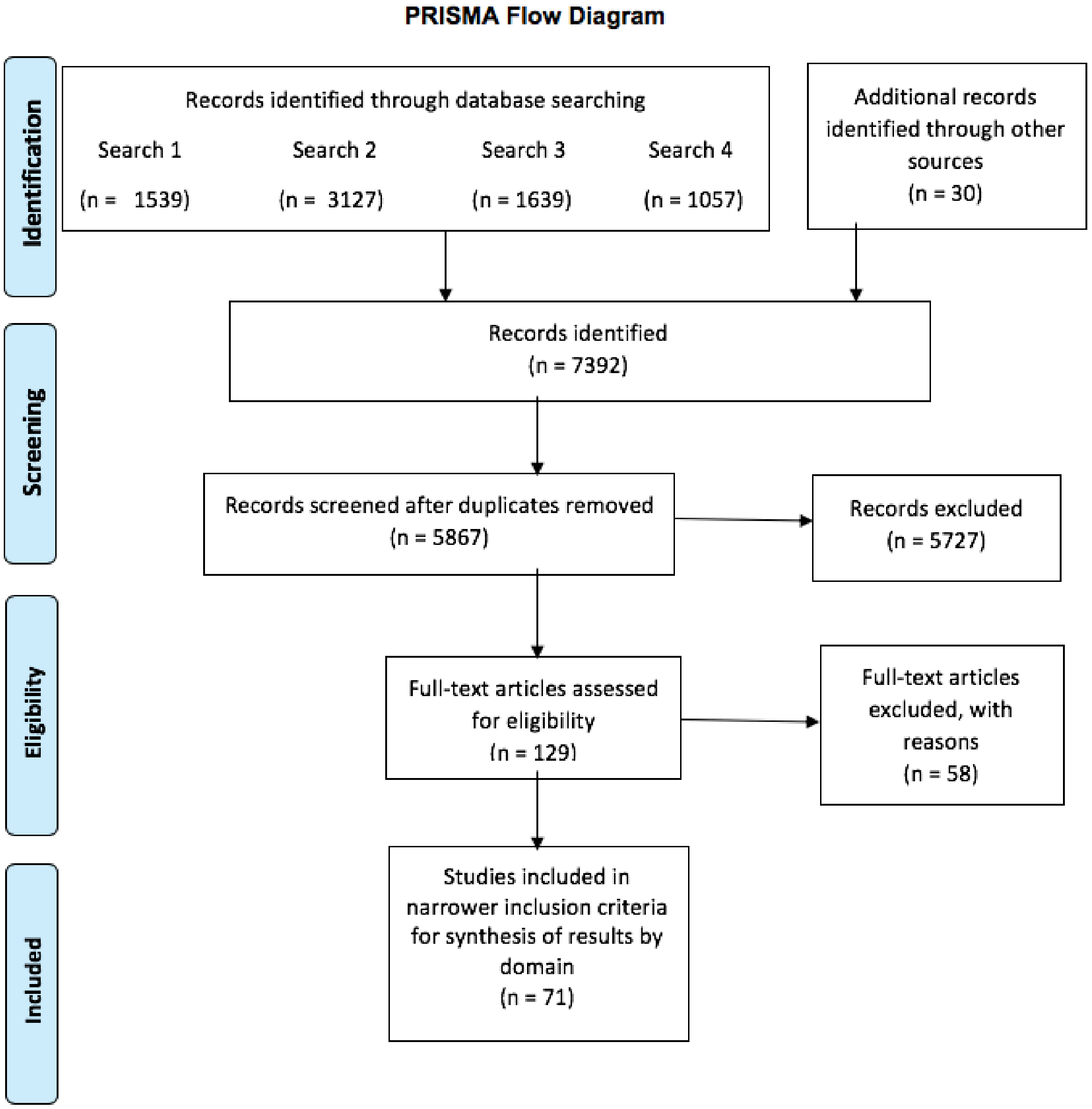
PRISMA Flow Diagram
Identification
Screening
Eligibility
Included
Records identified through database searching
Search 1
(n = 1539)
Search 2
(n = 3127)
Search 3
(n = 1639)
Search 4
(n = 1057)
Additional records identified through other sources
(n = 30)
Records identified
(n = 7392)
Records screened after duplicates removed
(n = 5867)
Records excluded
(n = 5727)
Full-text articles assessed for eligibility
(n = 129)
Full-text articles excluded, with reasons
(n = 58)
Studies included in narrower inclusion criteria for synthesis of results by domain
(n = 71)

Figure 1: PRISMA flow diagram

2.9 Analysis approach

Once relevant articles had been identified, each paper was evaluated by a discipline expert against an evaluation framework to assess the study's quality and rigour and to summarise the research findings. This involved extracting the following information for each of the studies:

1. Domain area addressed – Mathematics, Science, the Arts, Writing or Reading.
2. Location of study – city/country.
3. Description of sample characteristics – typical sample vs. atypical sample.
4. Form of formative assessment – What is the source of the assessment tool used? Who is the author? (for example, classroom teacher, school-based learning community, assessment expert working with teachers, ready-made package, standardised/non-standardised?)
5. Validity of the assessment? Did the assessment measure what it claimed to measure?
6. Impact of the formative assessment on student learning outcomes.
7. What is being measured? L = Learning outcomes; PR = Progress; G = Gaps in understanding; S = Specific difficulties; R = Reasons for difficulties (cognitively diagnostic/task diagnostic).
8. Who is the feedback to? L = learner; T = teacher; S = software. Whose behaviour is expected to change as a result of this feedback?
9. Type of feedback to learner: NA = Not applicable; S = Score/grade provided only; SF = Score/grade and feedback re correct answer; SE = Explanation of the difference: correct results and explanation of differences between their result and the correct result; SEI = Explanation and improvement suggestions: as previous but now students also receive some specific suggestions for improvement; SEA = Explanation and specific activities: students are given information about the correct results, some explanation and specific activities to undertake.
10. Type of feedback to teacher: NA = Not applicable; S = Overall score only; SS = Separate scores provided for specific aspects of performance; I = Possible explanation of the problem areas and suggestions for additional instructional focus; A = Possible explanation of the problem areas and specific instructional activities to undertake.
11. Is evaluation based on a theoretically valid task model, i.e. the sequence of activities to be successfully completed to meet learning outcomes and how learners typically progress through them (learning progression)?
12. Is the intervention based on a theoretically valid cognitive model, i.e. a model of prerequisite cognitive and learning skills underlying successful progression? For example, does the process require a significant amount of working memory, attention, motivation, persistency, cognitive ability, language skills etc.?
13. Are the actions/interventions following the assessment task evidence-based, i.e. is the instructional model valid?
14. What tools/resources are used in the assessment process and intervention? (Could be teacher designed or commercial).
15. Other comments, observations or questions.

See Appendix 1 for a summary of this information for each included study.

3 Chapter 3: Review results – Research on effective formative assessment practices in Mathematics

3.1 Summary

3.1.1 Evidence for the impact of formative assessment on student learning in Mathematics

- The results across the Mathematics studies are mixed.
- Some studies report statistically significant gains for students in some aspects of Mathematics, but not in others. In other studies, benefits are isolated to specific groups of children (low achieving, high achieving) or require extensive support or involvement of experts.
- Some studies found no statistically significant change in students' maths scores.
- Many studies lacked detail regarding the activities of the control group and/or could not differentiate effects due to formative assessment versus effects due to time spent on mathematical activities. Some studies provided the intervention group with teacher professional learning opportunities and/or curriculum materials not afforded to the control group so it is difficult to determine the specific elements of the formative assessment program or any other variables that may have influenced the results.
- Some studies report that teachers improved their formative assessment knowledge and practices, but students showed no statistically significant improvement.
- Studies focused on professional development to support teachers in making better use of formative assessment tools report no advantage for students of teachers who had completed the professional development, with the main conclusion being that students who completed more assessments showed greater gains in maths achievement.

3.1.2 Evidence for the effectiveness of particular tools and resources in Mathematics

- Two maths studies included a robust experimental design and provided sufficient detail to judge the impact of the formative assessment intervention. Both studies reported no statistically significant effects on students' achievement in maths.

3.1.3 Implications for effective implementation of formative assessment practices

- Teacher professional learning appears to be crucial to foster teachers' pedagogical content knowledge and hence their use of impactful formative assessment practices.
- Even with sustained teacher professional learning, there is not necessarily any increase in general teaching quality.
- Providing very detailed formative assessment guidelines and teaching materials can help teachers whose lessons have low instructional quality, but they may also constrain teachers whose lessons already possess high instructional quality.
- Frequent use of high-quality formative assessment practices is crucial; it must be regular and sustained to have any chance of impacting student learning.
- There are challenges for teachers in precisely determining a student's learning difficulties and instantly deciding on a course of action to remedy them during a lesson.

- Future studies that focus on professional development need to consider student learning in conjunction with intensive examinations of teacher's instructional practices.

3.1.4 Chapter overview

This chapter examines studies exploring formative assessments in Mathematics. It is divided into two sections: Part A explores formative assessment in Mathematics in general and Part B looks specifically at online formative assessment in this domain.

3.2 Part A: Formative assessment in Mathematics

Twenty-four Mathematics studies were originally identified for this review. Of these, four studies were considered unsuitable for inclusion due to the following reasons:

- two studies involved students who were not of school age: Fantuzzo, Gadsden and McDermott (2011) focused on the early childhood years; Hudesman et al. (2014) focused on college students
- two studies focused on topics unrelated to formative assessment: Boakes (2009) investigated the use of origami to improve students' spatial reasoning; Samo, Darhim and Bana Kartasasmita (2017) examined ways to improve higher-order thinking in Mathematics.

Twenty Mathematics studies were retained and included in the review. Of these, 12 were conducted in North America (Axtell, McCallum & Bell 2009; Bond & Ellis 2013; Bryant et al. 2011; Carlson, Borman & Robinson 2011; Clarke et al. 2011; Clarke et al. 2014, 2015; Konstantopoulos, Miller & van der Ploeg 2013; Menesses & Gresham 2009; Phelan et al. 2011, 2012; Randel et al. 2016). There was one study each in Belgium (Baten, Praet & Desoete 2017), Germany (Pinger et al. 2018), Indonesia (Sumantri & Satriani 2016), Jordan (Abu-Hamour & Mattar 2013), Nigeria (James & Folorunso 2012), Singapore (Wong 2017), Sweden (Andersson & Palm 2017) and The Netherlands (van den Berg, Bosker & Suhre 2018).

3.2.1 Sample characteristics

The majority of studies focused on specific Mathematics content, such as early number learning (Baten, Praet & Desoete 2017; Bryant et al. 2011; Clarke et al. 2011, 2014, 2015; Konstantopoulos, Miller & van der Ploeg 2013; Menesses & Gresham 2009), introductory algebra (Phelan et al. 2011, 2012), probability and statistics (Bond & Ellis 2013) and Pythagoras' Theorem (Pinger et al. 2018). Some studies also reported on mathematical or instructional processes, such as metacognition (Baten, Praet & Desoete 2017; Bond & Ellis 2013; Wong 2017), fluency and automaticity (Axtell, McCallum & Bell 2009) and peer tutoring (Menesses & Gresham 2009).

Many studies have a particular interest in examining the effects of interventions for low-performing Mathematics students (Abu-Hamour & Mattar 2013; Axtell, McCallum & Bell 2009; Baten, Praet & Desoete 2017; Menesses & Gresham 2009) or districts (Carlson, Borman & Robinson 2011), and Tier 2 lessons, in which small groups of students are withdrawn from their usual lessons for intensive remedial instruction (Bryant et al. 2011; Clarke et al. 2014).

Some studies compared different assessment types, such as multiple choice and extended response questions (Sumantri & Satriani 2016), investigated the provision of feedback on formative tests (James & Folorunso 2012) or examined the frequency with which formative assessment is provided by the teacher (van den Berg, Bosker & Suhre 2018). The studies also investigate specific formative assessment programs or tools (Konstantopoulos, Miller & van der

Ploeg 2013; Pinger et al. 2018), teaching resources (Axtell, McCallum & Bell 2009; Baten, Praet & Desoete 2017) or teacher professional learning (Andersson & Palm 2017; Pinger et al. 2018; Phelan et al. 2011, 2012; Randel et al. 2016).

3.2.2 Typical formative assessment practices in Mathematics

The formative assessment activities described by the studies are often about providing information to teachers and/or students about the correctness of students' responses to mathematical tasks. These tasks include closed questions about basic number facts (Clarke et al. 2011, 2014, 2015; Menesses & Gresham, 2009), speed and accuracy tests (Abu-Hamour & Mattar, 2013; Axtell et al. 2009), multiple-choice questions (Sumantri & Satriani, 2016) and open-response tasks that require students to show working and explain reasoning (Phelan et al. 2011, 2012; Pinger et al. 2018). Formative assessment practices in Mathematics are also designed to help teachers become aware of students' mathematical misconceptions or knowledge gaps, along with information on how students could improve performance in relation to the learning goals (Pinger et al. 2018; Randel et al. 2014).

Feedback is provided by the teacher (Clarke et al. 2011, 2014, 2015; van den Berg, Bosker & Suhre 2018), peers (Bond & Ellis 2013; Menesses & Gresham 2009), visual and auditory computer displays (Baten et al. 2017) or as self-assessment (Wong 2017). Formative assessment includes progress monitoring using charts or graphs to display students' performance over time, perhaps using a points system (Menesses & Gresham 2009) or rewards (Bryant et al. 2011). Feedback is most commonly provided to the student directly (Baten, Praet & Desoete 2017; Bond & Ellis 2013; James & Folorunso 2012; Menesses & Gresham 2009; Pinger et al. 2018; Sumantri & Satriani 2016; van den Berg, Bosker & Suhre 2018; Wong 2017) or through the teacher (Abu-Hamour & Mattar 2013; Axtell, McCallum & Bell, 2009; Clarke et al. 2011, 2014, 2015). In some studies, feedback is only given to the teacher, who then makes instructional decisions about students' learning needs (Andersson & Palm 2017; Carlson, Borman & Robinson 2011; Konstantopoulos, Miller & van der Ploeg 2013; Phelan et al. 2011, 2012; Pinger et al. 2018; Randel et al. 2014).

3.2.3 Evidence of impact on student learning

3.2.3.1 Highest quality studies

Two Mathematics studies included a robust experimental design and provided sufficient detail to judge the impact of the formative assessment intervention. A study in Germany by Pinger et al. (2018) focused on secondary students' learning of Pythagoras' Theorem to investigate the interplay between formative assessment and teaching quality. A quasi-experimental study design was implemented for 15 teachers in the control group (n = 361 students) and 20 teachers in the intervention group (n = 498 students). Both groups implemented the same teaching program of 13 lessons, each of 45 minutes duration. All teachers received initial training about the teaching unit and were provided with obligatory teaching materials to ensure all students worked on the same tasks. Intervention class teachers received an additional training session on formative assessment practices and were shown how to administer a diagnostic tool. The tool included an assessment component containing one or two mathematical problems with space for the student to write down the solution, and process-oriented student feedback to indicate strengths, weaknesses and strategies to improve. Following lessons 5, 8 and 11 in the teaching program, intervention class teachers marked the diagnostic tool and returned it to students along with process-oriented feedback.

No statistically significant effects were found on students' achievement in the post-test. Interestingly, although there was a positive association between teachers' process orientation and use of instructional time with students' achievement, this was suppressed by the formative assessment intervention. The authors concluded that implementing challenging tasks and supportive feedback via predesigned tools does not automatically change teachers' general instructional quality. Pinger et al. (2018) noted that while professional development can foster teachers' pedagogical content knowledge, it may do so without any increase in general teaching quality. Furthermore, while the detailed guidelines and teaching materials helped teachers whose lessons had a low degree of instructional quality, these resources might have constrained teachers who already possessed high instructional quality.

A study of Classroom Formative Assessment (CFA) was conducted in The Netherlands by van den Berg, Bosker and Suhre (2018). CFA is designed for teachers to assess students' mastery of a learning goal during the lesson and provide immediate instructional feedback to correct students' misconceptions. The study used a quasi-experimental pre-test – post-test design to compare two interventions: (1) a CFA model of daily and weekly goal-directed instruction, assessment and immediate instructional feedback for students who required additional support, and (2) the usual practice in Dutch schools of half-yearly mathematics tests and weekly pre-teaching sessions for low-achieving student groups. To ensure fidelity to the CFA model, teachers in the CFA intervention participated in a professional development program that included an initial workshop followed by a lesson observation and reflective conversation with a coach who provided advice to teachers on their CFA class implementation. Teachers in the control condition also received professional development, though not as intensively as the CFA teachers.

Three CFA teacher lessons were observed in order to determine the extent to which they implemented the model. Results indicate that CFA teachers did assess their students' mastery of the learning goals and provide immediate instructional feedback more often than teachers in the control condition. However, the teachers' participation in the CFA condition did not significantly enhance student performance. The authors suggest a possible explanation for the absence of an effect is that, though CFA teachers made more frequent use of assessments and immediate instructional feedback than the control teachers, they did not do so as often as intended. The study also highlights how, even with detailed formative assessment information, teachers are challenged by precisely determining a student's learning difficulties and instantly deciding on a course of action to remedy them during a lesson.

3.2.3.2 Summary

Both studies included a major focus on professional development designed to support teachers in gathering, interpreting and using formative assessment data. They included well-designed control conditions so that treatment effects can be confidently attributed to the formative assessment interventions but provide mixed results for their impact. Pinger et al. (2018) and van den Berg, Bosker and Suhre (2018) reported no statistically significant effects on students' Mathematics achievement. The two studies highlight some of the challenges teachers face when putting their pedagogical knowledge into practice and making in-the-moment decisions to benefit students' learning during Mathematics lessons.

3.2.3.3 Lower quality studies

Many studies included in this review investigated various aspects of formative assessment in Mathematics instruction, but also provided the intervention group with teacher professional learning opportunities and/or curriculum materials not afforded to the control group. For example, Clarke et al. (2011, 2014, 2015) developed an Early Learning in Mathematics (ELM) program in number, geometry, measurement and mathematical vocabulary for Grade 1 students at risk for mathematical difficulties. The ELM intervention provided experimental class teachers with professional development focused on implementing the ELM curriculum and ELM materials, which included scripted lessons. Clarke et al.'s 2011 study focused on at-risk students in the general education or Tier 1 classroom setting. They used a randomised block design with 64 classrooms randomly assigned within schools to treatment (ELM) or control (standard district practices) conditions. The study reported post-test scores of at-risk treatment students were significantly greater than their control peers, and the gains of at-risk treatment students were greater than the gains of peers not at risk, effectively reducing the achievement gap.

Clarke et al. (2014) implemented the ELM in a Tier 2 intervention, in which low-performing students were withdrawn from class for small-group remedial instruction, so teachers could provide timely academic feedback to confirm correct student responses and address potential misconceptions. The intervention was intensive, with students receiving 60 lessons of approximately 30 minutes each over a 20-week period. Even so, the authors reported mixed results with some statistically significant gains for students' conceptual understanding but not for procedural fluency; nor did the researchers find any association between higher levels of implementation fidelity of ELM and student outcomes. Clarke et al. (2015) also found that students in ELM classrooms did not achieve Mathematics outcomes that were significantly different than those achieved by students in control classrooms. In all three studies, the authors conceded they were unable to determine the specific elements of the ELM program or any other variables that may have influenced their results.

Andersson and Palm (2017) implemented a professional development program (PDP) focused on a range of formative assessment practices. These included guidance for teachers on providing feedback and how to adapt their instruction by more frequently gathering information about student learning. Participants were 22 Year 4 teachers who met with a PDP manager for six hours each week during one school term for a total of 144 hours. In addition, the teachers had another 72 hours available for reading, planning and reflecting on the PDP formative assessment activities. Between meetings, teachers were asked to implement the activities in their classes. The researchers made two unannounced classroom observations of each teacher during the school year. After adjusting for pre-test scores, students in the intervention classes performed significantly better than students in the control group on a test covering Year 4 Mathematics curriculum content. In contrast to Clarke et al.'s 2014 findings, Andersson and Palm (2017) discovered no statistically significant differences for procedural tasks or those requiring problem solving and reasoning. However, they note that professional development requires extended time and considerable expert support to advance teachers' formative assessment practices.

Often curriculum materials and/or instructional resources accompanied the teacher professional development program. Phelan et al. (2011, 2012) devised an intervention to compare formative assessment in conjunction with professional development and a Teacher Handbook containing instructional resources to a control group that only received formative assessments. The implementation comprised eight algebra lessons over an academic year, though there was no attempt to control for what the teachers did outside those lessons. Results indicated no statistically significant main effect for the intervention. Randel et al. (2016) examined the impact of the Classroom Assessment for Student Learning (CASL) teacher professional learning program developed by the South Carolina Department of Education. Teachers also received a textbook, DVDs, ancillary texts and a learning team facilitator handbook. Results indicated that teachers who participated in CASL enhanced their knowledge of assessment practices and used formative assessment strategies to ask questions about students' current knowledge, where they would like to improve and how they could do so. However, there was no difference found in student achievement for the experimental classes.

Other studies did report significant gains for students' mathematical achievement, but they include a range of intervention activities. Konstantopoulos, Miller and van der Ploeg (2013) implemented a randomised experiment design to examine the impact of interim assessment programs on Mathematics and Reading achievement. While they found the treatment effects were positively significant in some instances, there was no attempt to monitor how teachers used the assessment programs. In a study by Bond and Ellis (2013), teachers taught different Mathematics topics to the experimental and control groups (probability/statistics and area/perimeter, respectively) and the researchers applied a post-test only design, so it is difficult to determine the treatment effect. Sumantri and Satriani (2016) also did not undertake any pre-test of student participants nor did the authors provide sufficient detail about how formative tests were scored. Abu-Hamour and Mattar (2013) reported a statistically significant effect for students in their Curriculum-Based-Measurement in Math Computation (M-CBM) program but did not provide details about the program or how it was taught. Menesses and Graham (2009) compared reciprocal peer tutoring (RPT, where pairs of students take turns role-playing tutor and tutee) and non-reciprocal peer tutoring (NPT, in which one student in the pair always tutors their partner). Tutors provided positive and corrective feedback to tutees on responses to basic number fact questions over a three-minute period. The authors found RPT students and NPT tutees scored significantly higher on the post-test than the control group, though the inclusion of progress-monitoring charts and extrinsic rewards such as lollies, stickers and small toys may have been a contributing factor. Baten, Praet and Desoete (2017) investigated the impact of computer-assisted interventions (CAI) in Kindergarten classes. They found significant gains in counting and comparing numbers for children from the intervention classes. However, the control group CAI focused on reading skills, so gains reported for number skills in the intervention classes might be related to the additional time spent developing them.

Some other studies showed a positive impact for teachers' formative assessment practices but did so with the substantial involvement of highly trained experts or the researchers themselves. James and Folorunso (2012) trained research assistants to teach classes in their study on the impact of feedback and remediation. Axtell, McCallum and Bell (2009) used graduate students enrolled in a special education program. Clarke et al. (2011) employed 'interventionists' who were district employees while in the Bryant et al. (2011) study, tutoring sessions were taught by trained doctoral or master's student research assistants. Carlson, Borman and Robinson (2011) analysed elementary school students' achievement in Mathematics and Reading, finding a positive effect on Mathematics achievement, although not for Reading. However, the intervention relied on consultants conducting monthly training sessions for teachers and working with district and school leaders to review the assessment data and identify problem areas at each school. Wong (2017) conducted a study to investigate the effects of self-assessment training on students' perceptions of self-assessment but taught the intervention students herself, which may have influenced her results.

3.2.3.4 Summary

Results of the lower quality studies are varied. For instance, Clarke et al. (2014) reported statistically significant gains for students' conceptual understanding but not for procedural fluency, and the students in Andersson and Palm's (2017) study made statistically significant gains overall, but not for procedural tasks or those requiring problem solving and reasoning. Carlson, Borman and Robinson (2011) found a positive effect on Mathematics achievement, although not for Reading, and while teachers in the study by Randel et al. (2016) developed their knowledge and implementation of formative assessment practices, there was no significant improvement in students' achievement.

These studies were identified as lower quality because they did not moderate for factors such as the provision of teacher professional learning, curriculum and instructional materials and lesson scripts, or the impact of trained experts. Hence, it is not possible to ascertain the extent to which any reported student achievement gains were specifically due to the impact of the formative assessment practices employed. Stronger research designs in these studies could, for example, have provided control teachers with a similar degree of professional development on formative assessment teaching strategies to eliminate professional development as an alternative possible cause. Consequently, the most favourable interpretation of these studies is that they provide some (mixed) evidence about the impact of formative assessment practices on students' mathematical achievement but offer little confirmation as to why these practices worked or the feasibility of replicating them.

3.3 Part B: Online formative assessment in Mathematics: Overview of studies

Based on the initial review of the literature, 20 papers were identified. Following an in-depth review to confirm appropriateness for inclusion, seven were deemed unsuitable for a variety of reasons:

- The focus of the study was on use of dynamic manipulatives not dynamic formative assessment.
- The study focused on alternative methods of scoring multiple choice questions but did not result in feedback to the student.
- The study focused on accommodations for students with special needs during summative assessments.
- There was no assessment of outcomes (rather a focus on how students interact with features in different environments).
- Tools were used as supplementary Mathematics tuition with no details on the nature of the assessment or feedback.
- The focus was on teacher mathematical content knowledge rather than formative assessment or assessment for learning.
- The focus was on a technological tool to support learning but with no focus on assessment for learning or formative assessment.

Thirteen studies were retained and included in the detailed review provided below.

3.3.1 Location of studies

Studies were conducted in diverse locations. These included: Europe (The Netherlands: Faber et al. 2017; Germany: Rakoczy et al 2019); USA (Chappell et al. 2015; Koedinger et al. 2010; Rochelle et al. 2016; Irving et al. 2016; Polly et al. 2017, 2018); and Asia (Taiwan: Hsiao et al. 2017; Tsuei 2017; Wu et al. 2017; Yang et al. 2016; Thailand: Wongwatkit et al. 2017).

3.3.2 Sample characteristics

The studies predominantly focused on students in Grades 2–9. The majority focused on typically achieving children with just two studies specifically targeting low-achieving children receiving support for learning (Chappell et al. 2015; Tsuei 2017). Sample sizes ranged from small-scale interventions focused on just two classes (e.g. Tsuei 2017; Wongwatkit et al. 2017) to very large-scale studies making use of nationally collected data (e.g. Polly et al. 2017).

3.3.3 What is being measured – learning, gaps, specific difficulties, reasons for difficulties?

In virtually all cases, student learning was the key variable assessed, either in grade-appropriate general Mathematics ability or specific curriculum units. Some interventions also used assessments in a diagnostic fashion to determine the focus of continuing instructional activities.

3.3.4 Who is the feedback to – the learner, the teacher or software?

For online or digital tools, feedback is most often given to the learner and the system. For example, in studies using adaptive tools, performance on a task by the learner is fed back into the system, which then, on the basis of pre-determined criteria decided by domain experts and teachers, provides feedback to the learner and directs them towards adaptive assignments. In cases where this intervention has not been pre-determined, the learner's performance is fed back to the teacher (and sometimes to the learner). The teacher then makes a decision about the next instructional component and tasks. In synchronous peer or online tutoring, there is continuous feedback from the tutor to the tutee and vice versa.

3.3.5 Type of feedback to the learner

Learner feedback has a broad range. It can include none or indirect feedback (where feedback from the assessment is directed to the teacher) through to explanation and specific activities.

3.3.6 Type of feedback to the teacher

In some cases, the teacher was given a student profile delivering a clear indication of areas requiring additional instructional focus but was not provided with additional instructional activities that should be undertaken. Even in cases where the teacher was provided with instructional resources to support learning, it was impossible to clearly differentiate how teachers adapted their instruction, or how the quality of their implementation impacted students' achievement and interest in Mathematics. In some cases, there was no immediate feedback to the teacher.

3.4 Evidence of impact on student learning progress/outcomes

3.4.1 Highest quality studies

The reviewed studies used a robust experimental design and appropriate analytical techniques. Importantly, with well-designed control conditions and control of other confounding variables (e.g. amount of instruction, topics covered, equivalent time on task in intervention and control conditions), treatment effects can be specifically attributed to the formative assessment intervention.

3.4.1.1 Formative assessment and feedback using adaptive computerised tools

In all examples presented, discipline experts undertook a *task analysis* prior to intervention to determine the skill hierarchy and relationships between skills in specific areas of mathematical knowledge (e.g. a unit focused on addition and subtraction of fractions with different denominators, or a unit focused on circle area). The algorithm for computerised dynamic adaptive assessment is based on this skill hierarchy for the knowledge domain. The upper-level skill with the most links is administered to a student. The status of the response is checked by the system. If a student answers the item correctly, then it is inferred the student also knows its prerequisite concepts. When students respond incorrectly to an item, they receive instructional prompts (designed by discipline experts) with different levels based on the number of times students answered incorrectly. Thereby learning is based on the nature of students' misconceptions. In other words, each student receives a different set of instructional prompts according to their needs. When no prompts remain, the student receives direct instruction on that item.

Wu et al. (2017) examined the use of a computerised dynamic adaptive assessment in a sample of 118 fifth-grade Taiwanese students. Prior to the intervention, students attended five periods of fractions instruction. They then completed a pre-test, participated in the intervention during one period, and then completed the post-test. Students from six classes were randomly assigned to three groups: a dynamic individualised assessment and individualised instruction group (group 1), an individualised instruction group (group 2) and an instruction group (group 3). Groups 1 and 2 received computerised individualised instruction based on expert knowledge structure. The students in group 1 received prompts according to which options they chose, while students in group 2 received direct instruction. In group 3, skills were taught by teachers in sequence, based on the group report of the pre-test. For students in groups 1 and 2, teachers were given a student profile providing clear indication of areas requiring additional instructional focus. Teachers were not provided with additional instructional activities to be undertaken. All three groups showed significant improvement from pre- to post-test. Taking into account pre-test performance, group 1 students performed significantly better than students in groups 2 and 3 (which did not differ from each other). This result confirms the benefit seen in group 1 is specific to the use of prompts as formative feedback, rather than due to general individualised instruction (because the group (2), which received individualised direct instruction, did not perform significantly better than the group (3), which received group instruction).

Wongwatkit et al. (2017) used a similar paradigm in a study of 63 Thai sixth graders during a learning activity lasting 150 minutes. The experimental group participated in a formative assessment-based personalised web learning system (equivalent to Wu et al.’s (2017) group 1). When students failed to correctly answer an item, the system showed hints and supplementary material/tasks to further guide their learning rather than providing them with correct answers. This means they were encouraged to find the correct answers at their own pace. The control group completed the same lesson in a conventional web-based learning system without the formative assessment (equivalent to Wu et al.'s (2017) group 2). In addition, students completed questionnaires to assess their learning style (visual vs. verbal) and perceptions of the learning system. Students then received material and prompts matched to their reported learning style (e.g. diagrams versus a written narrative).

The formative assessment-based personalised web learning system had significantly greater positive effects (M= 7.89; SD = 1.99) on the visual-style students’ learning achievements than the conventional web-based learning system (M = 6.45; SD = 2.01). No significant difference was found between the learning achievements of verbal-style students who used the two learning approaches. Students with a visual learning style (M = 7.89; SD = 1.99) benefited more than those with a verbal learning style (M = 6.50; SD = 2.47) when learning with the formative assessment-based, personalised web learning system. Students in the experimental group, and particularly those in the visual learning group, showed significantly higher perceived usefulness (useful for improving learning performance), perceived assistance to learn (degree to which a student believed the formative assessment activities, learning paths and supplementary materials/tasks in the system was personalised and appropriate for him/her to learn), and material layout satisfaction than those in the control group. Children who reported a preferred verbal learning style were not exposed to the visual learning materials, so it is impossible to definitively state that only students with a stated preference for visual learning would benefit. It is possible that all students will benefit more from visual learning materials compared to learning materials presented in a narrative format.

Hsiao et al. (2017) conducted a quasi-experimental study with 153 Taiwanese seventh grade students using the Problem-solving Assessment, Diagnosis, and Remedial Instruction (PSADRI) system. The intervention lasted three weeks and focused on a linear equations unit. The authors designed the contents of the training system in accordance with grade-appropriate competency indicators. Students in the experimental group and the control group learned the same content and carried out the same activity procedures with different teachers. The training system was used for assessment, diagnosis and teaching guided mathematical problem-solving in the experimental group, whereas the control group participated in traditional instruction. The experimental and control groups differed significantly at both pre- and post-test (experimental group was better). Controlling for pre-test score, students in the experimental group scored significantly higher at post-test compared to the control group. The authors also examined scores on the four individual problem-solving ability assessments. For problem translation and integration, students with lower scores at pre-test showed greater benefit from traditional instruction, while children with higher scores showed more benefit from PSARDI. Overall, problem solving was higher in the experimental group regardless of pre-test ability. No difference was found between the experimental and control groups on solution planning and monitoring. Students in the experimental group displayed more positive perceptions toward Mathematics learning compared to the control group.

3.4.1.2 Computerised assessment with feedback from the teacher

Rakoczy et al. (2019) conducted a cluster randomised field trial involving 620 German ninth-graders from 26 classes. Prior to intervention, all teachers were introduced to subject-specific content and provided with maths tasks for a thirteen-lesson teaching unit on Pythagoras' Theorem. Teachers were also trained to use a didactic approach that fostered students' ability to apply maths tools to real-world problems. For the formative assessment condition, teachers were given additional training on assessing student performance and instructed in providing written process-oriented feedback. At the end of the fifth, eighth and eleventh lessons, teachers asked students to complete tasks on a diagnostic and feedback tool. After the lessons, they assessed students' solutions, added individualised, process-oriented feedback on students’ strengths and weaknesses, and recommended future strategies. The authors developed a list of cognitive processes and operations based on cognitive task analyses needed to solve the respective diagnostic task (e.g. identify catheti and hypotenuse in a right-angled triangle). Each process and operation could be fed back as a strength or weakness depending on whether a student had mastered it or not. A strategy or hint on how to continue was provided for each process and operation that had not been mastered. At the end of the feedback section, students were asked to complete another similar task applying strategies provided in feedback.

Results indicate that students perceived feedback more useful when teachers had employed the diagnostic and feedback tool as they had been instructed in teacher training during the formative assessment condition. Moreover, students reported being more confident about their achievement on a forthcoming test and tended to show greater interest in the test topic. Achievement, however, did not differ between the control and intervention conditions. There was also no indirect effect of formative assessment on achievement via self-efficacy or perceived usefulness. The direct and indirect effect together did not result in a statistically significant total effect of formative assessment on achievement.

3.4.1.3 Computerised tools to assist with synchronous in-class formative assessment and feedback

Irving et al. (2016) conducted a three-year randomised crossover trial involving 3,589 US students enrolled in Algebra I classes. The trial used classroom connectivity technology (CCT), in which classroom wireless communication systems connected student handhelds (in this case, graphing calculators) with the teacher’s computer. Using a quick poll and learning checks, teachers sent questions to each handheld device for a student response. Resulting data was displayed publicly for whole-class review. Teachers could also use screen capture to get a snapshot of each student’s calculator screen for review and discussion. Each of these components gave teachers data to make instructional decisions and give students immediate feedback. Using an activity centre, teachers could also involve students in discovery lessons through display and interaction with a coordinate system.

Cohort 1 teachers were trained to use CCT during a one-week residential course followed by professional development. In the first year, Cohort 1 teachers used CCT while Cohort 2 teachers used graphing calculators (the control condition).

Cohort 2 teachers then received CCT training and began using the technology in the second year. The study continued with the same teachers over the three years, but with different student groups each year. There was one control group each year (Cohort 2, year one) and five comparison treatment groups (Cohort 1, years one,two and three and Cohort 2, years two and three). Analyses compared achievement within the five treatment groups with the control group. Controlling for teaching experience and pre-test algebra scores, analyses revealed two (out of five) significant treatment effects in favour of CCT (effect sizes 0.30 and 0.20). The three remaining comparisons were not statistically significant (effect sizes 0.23, 0.24, 0.13, $p > 0.15$). There was a concern that pre-tests may have been administered late in some of the treatments groups, resulting in over-inflated pre-test scores and causing an unfair adjustment when controlled for in post-test scores analysis. The same analysis including only teacher experience as a covariate revealed three of four comparisons were significant, with treatment groups achieving higher algebra post-test scores than the control group. The authors indicate that medium effect sizes are relatively rare in national randomised control studies. They also note the intervention should be seen as CCT plus professional development because simply providing CCT equipment was previously found ineffective. CCT increases learning opportunities by acting as a mediating tool to produce classroom environments that support student examination and pattern analysis, supporting collaborative work and justification of mathematical generalisations. CCT also facilitates immediate student and teacher feedback and promotes productive classroom discourse.

3.4.1.4 Summary

The benefits of formative assessment for typically achieving students from Grades 5 to 9 offer mixed results. All studies reviewed focus on targeted units of assessment within Mathematics rather than general assessments of mathematical ability; this appears to allow for specific feedback to students based on a clearly defined hierarchy of skills. Benefits appear greater where feedback is provided instantaneously and more frequently, as in the case of Wu et al.’s (2017) adaptive program, and when assessment and feedback are embedded within classroom activities (Irving et al. 2016). There was no benefit to achievement where feedback was provided by the teacher at a minimal number of points during a longer learning session, as in Rakoczy et al. (2019).

With regard to changes in teacher practice, in some cases it is not clear what information is fed back to the teacher (Hsiao et al. 2017; Wongwatkit et al. 2017), or how feedback is subsequently used to inform instructional activities (Wu et al. 2017). Rakoczy et al. (2019) acknowledge they were unable to differentiate how teachers implemented the formative assessment intervention and how the quality of their implementation impacted students' achievement and interest in Mathematics. In Irving et al. (2016) teachers received immediate feedback on student learning but the lack of adaptive programming placed limits on the individualised pace, level and frequency of skill practice undertaken.

In three studies assessing motivational factors (e.g. self-efficacy, perceived usefulness of feedback, interest), students in formative assessment conditions showed more positive ratings than students in control conditions. In one study (Rakoczy et al. 2019), this did not provide a direct route to impact on achievement. In the other two studies, this mediation was untested (Hsiao et al. 2017; Wongwatkit et al. 2017). Further research in this area should place greater emphasis on investigating cognitive and motivational processes that explain how formative assessment impacts learning.

3.4.2 Lower quality studies

These lower quality studies used a robust experimental design and appropriate analytical techniques. However, treatment effects cannot be clearly and specifically attributed to formative assessment intervention.

3.4.2.1 Computerised assessments with feedback provided by the teacher

Two US studies examined the efficacy of ASSISTments, a web-based Mathematics cognitive tutor. Like the adaptive systems described above, ASSISTments scaffolds problems into requisite skills and knowledge components. If a student incorrectly answers an original item or requests help, the first scaffold is automatically generated. Once in scaffold tutoring, the student completes a series of scaffolds for that item. Teachers are also provided timely, organised feedback on student work, helping them decide on the next unit of assessment for individual students or whether to adapt instruction to the entire class.

Rochelle et al. (2016) conducted a randomised field trial with 2,850 US seventh-graders incorporating ASSISTments in regular homework over a single school year (engagement with ASSISTments averaged 14 hours). The treatment group showed higher post-intervention scores (controlling for pre-intervention mathematical ability and other demographic variables). The effect of ASSISTments intervention was greater for lower performing students than for higher performing students. The intervention also incorporated teacher professional development to increase teachers' readiness to use ASSISTments. Target practices included: (1) encouraging students to rework problems they got wrong initially (and to enter revised answers), (2) focusing attention on homework problems students did not answer correctly, (3) reviewing correct solution processes for problems students found difficult, and (4) discussing common wrong answers to address underlying misunderstandings. Teachers could then personalise ASSISTments for individual students, groups of students or the whole class. For example, they could assign additional practice to particular students (e.g. by assigning skill builders), enter other problems or give hints to existing problem sets. Students directly benefited from the feedback and hints they received while doing their homework, regardless of whether their teacher's behaviour was different during classroom instruction. However, teachers could also use ASSISTments-generated reports to adapt their teaching based on student work. The authors indicated additional data analysis would be undertaken, including looking at the number of homework items completed, the extent to which changes in student learning were mediated by changes in teacher behaviour, and analysing and integrating additional data from observations, interviews, surveys and instructional logs. There is preliminary indication that teachers did change their behaviour but, due to the complexity of analysing and integrating the multiple qualitative and quantitative data sources, comprehensive analysis of this issue awaits subsequent research effort.

In the second study, Koedinger et al. (2010) conducted a year-long study with 1,240 US seventh graders. The treatment group showed higher seventh grade Mathematics scores (controlling for sixth grade mathematical achievement), but this effect was specific to students receiving special education. Typically achieving students in the treatment condition did not attain significantly higher post-intervention scores compared to typically achieving students in the control group. The assignment to treatment versus control group was not random because the assignment to treatment condition was based on whether schools had sufficient access to computers. The authors were uncertain if their results were caused by ASSISTments due to the nature of quasi-experimentation and potential selection bias.

Faber et al. (2017) examined the use of a digital formative assessment tool – Snappet – over a five-month intervention window with 1,808 Dutch Grade three students attending 79 schools. Students using Snappet (experimental condition) completed assignments on their own Snappet tablets with instructional content and assignments comparable to traditional instruction (control condition). Randomisation to treatment versus control was at the school level. Students opened assignments from their Snappet start screen and received simple feedback (correct/incorrect) immediately on completion. Based on performance, students were then directed to their adaptive assignment (decided by the teacher). Teachers followed student progress on their own dashboard where they could also (1) preview assignments in each lesson and select assignments for students, and (2) monitor progress of individual students (e.g. performance on a specific learning goal compared to other learning goals), the entire class (e.g. comparison of performance with other classes using Snappet), or the lesson (e.g. number of assignments each student completes and accuracy of student answers). Teachers could also select quiz functions and instruction videos. In the experimental condition, teachers attended an introductory lesson on how to integrate Snappet into classroom lessons. They were given the option to follow additional lessons about interpreting and using feedback for differentiated instruction. All teachers had access to consultation with a coach.

Using a standardised Mathematics achievement test (employed across many Dutch schools), students in the experimental condition showed significantly better performance at post-test (after controlling for pre-test ability). The difference between control and experimental groups in mean achievement growth was highest for the top 20% performing students. Note that lower performing students did show benefit from using Snappet but a comparison to control children at the same ability level showed their level of benefit was less than higher achieving children. Faber et al. (2017) argue higher achieving students benefit from doing more (and more challenging) adaptive tasks, which does not happen in a business-as-usual classroom where instruction is targeted at low- to average-ability students. Consequently, higher performing students do not typically have the opportunity to engage in more-challenging work.

Faber et al. (2017) is one of the few studies to conduct classroom observations examining changes in teacher instructional practice after Snappet feedback. Unfortunately, this is only briefly mentioned, noting that observation scores indicated teachers did not use teacher feedback to a great extent. Teachers reported a better overview of lesson progress and students' understanding of the assignments, allowing them to respond more quickly and purposefully to students' questions. Overall, it is unclear whether the intervention resulted in better differentiated classroom instruction by the teacher. The only other effect found was that completion of more assignments (based on log file data) was associated with higher student motivation, but the direction of this relationship is unclear. The authors acknowledge higher motivation could result in more Snappet use and subsequently higher achievement. They do not report whether there is a relationship between motivation and maths performance in the control group and note the mechanism for improvement in the experimental group is unclear. It is possible that the greater change in performance for students in the experimental group was due to completing more assignments because no control group data exists for the number of tasks completed during the same period. Therefore, Faber et al. (2017) cannot definitively say that greater benefits for the experimental group are due to formative feedback. Ideally, the control group would also receive digital intervention to complete the same number of assignments but without the formative feedback component.

3.4.2.2 Teacher professional development to support computerised formative assessment and feedback.

Two US studies featuring students from Kindergarten to Grade 2 focused on using a generative internet-based tool (Assessing Mathematics Concepts–AMC Anywhere) formative assessment system (Polly et al. 2017, 2018). Both studies focused on how teacher professional development led to making better use of AMC Anywhere. Teachers in the treatment conditions were provided with up to 80 hours of professional development during which they learned about formative assessment and how children develop number sense. Teachers also learned how to use AMC Anywhere data to select and modify instructional activities from Developing Number Concepts (DNC) curricular resources. By contrast, teachers in the control group were given a 30-minute introduction to AMC Anywhere. In Polly et al. (2018; n = 5,302 in Kindergarten and Grade 1), all teachers were asked to conduct an assessment of counting skills a minimum of three times per school year (and up to a maximum of 20 times). There was negligible effect on the treatment condition across four counting tasks. In analyses predicting the intercept and slope of final counting skills, there was no effect on the treatment condition. Achievement was predicted by the number of assessments given (irrespective of treatment condition), where the more a teacher assessed the class using the AMC formative assessment system, the better students performed. A similar study by Polly et al. (2017; n = 13,567 in Kindergarten through Grade 2) indicated that growth in number-sense achievement was again predicted by the number of assessments administered and not by the treatment condition. In both studies, the authors highlight they provide no objective data regarding how teachers adapt their instruction based on data fed back by AMC Anywhere.

3.4.2.3 Peer and online tutoring

Tsuei (2017) examined i-GMath, a synchronous peer-tutoring system on tablet devices. The system was designed to provide virtual Mathematics manipulatives representing the student problem-solving process. It incorporated a reward scheme to help motivate low-achieving students to learn Mathematics. Teachers assigned mathematical problems using a learning management subsystem, and helping tools were developed to facilitate students' peer-tutoring behaviours – that is, a tutee could seek help by asking, 'Please indicate the error,' 'Give a hint' or 'Please demonstrate the solution.' The tutor could then use a number of tools to provide feedback to the tutee.

Thirty-four third grade Taiwanese students receiving learning support for Mathematics participated in the study. The experimental group worked face-to-face in dyads using i-GMath while the control group worked face-to-face in dyads using a collaborative learning strategy. Weekly tests were administered to both groups over a seven-week period. A significant interaction effect revealed the i-GMath group showed significantly greater increases in achievement from week one to week seven, compared with the control group. Based on an analysis of the help-seeking sequence and tutor feedback, the author argues that requiring tutees to select a help function that corresponded to their needs forced them to communicate these needs to their tutor. These requests facilitated tutors' use of the various helping behaviour tools. Tsuei argues that, for low-achieving students, help-seeking behaviour leads to a much greater increase in the chance of success than trying to resolve the error alone. Unfortunately, Tsuei provides no analysis of student interactions in the control group. Furthermore, other aspects of the i-GMath system may have contributed to the increase in Mathematics ability, for example, the motivation associated with gaining rewards.

Yang et al. (2016) examined a reciprocal peer-tutoring-enhanced mathematical communication (RTPMC) system, designed for supporting third grade Taiwanese students' mathematical creations and reciprocal peer-tutoring activities (n = 51 from two classes). Students participated in 13 activities during one semester. The outcome variable assessed was mathematics communication ability, not actual achievement. The learning activity involved four sub-activities: creating, reciprocal peer-tutoring, revising and staging. The control variable was the daily learning approach. In other words, both groups had the same Mathematics learning time in the same one-to-one self-learning environment. However, the control group practised Mathematics by teacher-led instruction for solving various word problems while the experimental group participated in RPTMC activity to solve related word problems chosen by the teacher and researchers. The RPTCM group showed significant improvement in total mathematics communication score whereas the control group showed no significant improvement. All three sub-domains of maths communication saw improvement, suggesting sufficient practice on finding solutions and explaining them through writing and drawing and verbal forms may assist students in expressing their own mathematical concepts and understanding others' mathematical thought. It remains to be determined if the ability to better communicate mathematical ideas translates into better Mathematics achievement. Again, with no analysis of the nature of teacher-led instruction and the interactions between teachers and students in the control condition, it is impossible to pinpoint the exact RTPMC features that resulted in a positive impact on mathematics communication ability. However, students' ability to better communicate their own mathematical understanding may provide formative feedback to the teacher in terms of the follow-up instructional support required.

Chappell et al. (2015) examined the influence of online tutoring for low-achieving US students in Grades 6–8. Focus EduVation (FEV) provide online interactive tutoring for K–12 students, delivered in a synchronous one-to-one environment using chat, instant messaging and a virtual whiteboard. Prior to tutoring, each student completed a diagnostic assessment and program, and school personnel completed individualised learning objectives and a learning plan. The latter served as a foundation for online tutoring in a differentiated engaging environment where skills were enhanced through sharing curricular materials, practice problems and visuals, and graphic features. Communication and collaboration between students and tutors were improved. Tutors completed a log of feedback provided to learners, including accessing prior knowledge, modelling, explaining steps, identifying process and operation errors, scaffolding using questions/prompts, and guided practice using multiple explanations and representations of target concepts.

Comparisons between tutored and non-tutored students in the same school revealed no significant differences in post-test scores (controlling for pre-test scores and multiple demographic criteria), although within-group effect sizes for pre- to post-intervention scores were higher for tutored students (d = 0.95) compared to non-tutored students (d = 0.24). Tutored students in a second school (but no within-school non-tutored comparison) showed great gains (d = 1.47). Note that students in School 2 undertook an average of 23 hours tutoring compared to 14 hours in School 1. Untutored children did not spend an equivalent amount of time (equivalent to tutoring hours) engaged in mathematical activities, so it is impossible to separate out change in performance based on formative assessment and feedback versus spending more time in mathematical activities.

3.4.2.4 Summary

The results from these studies again present a mixed picture regarding the impact of formative assessment interventions. Of those reporting significant benefits, some isolate those benefits to specific groups of children (low achieving, high achieving). Other studies do report overall beneficial effects but lack a sufficiently large sample size to look at effects for specific groups of students. Unfortunately, these studies lack detail regarding control group activities and/or cannot differentiate effects due to formative assessment versus effects due to time spent on Mathematics activities. Ideally, these studies would include a digital intervention control group where the same number of assignments were completed as experimental groups, but without the formative feedback component.

Studies focused on professional development supporting teachers to make better use of formative assessment tools report no advantage for students of those teachers who completed the professional development, with the main conclusion being students who completed more assessments showed greater gains in Mathematics achievement. Future studies that focus on professional development need to consider student learning in conjunction with intensive examinations of teacher instructional practices. This would help to better understand if and how teachers are leveraging assessment data to make data-based instructional decisions and support their students' mathematical understanding. Simply providing professional development, much like simply providing technology, does not automatically equate to adapted assessment and instructional practices in the classroom. Again, results suggest that frequency and embeddedness of formative feedback may be critical, where assessment and feedback at minimal points during the year is insufficient to provide timely feedback to students on specific learning tasks vital to their success.

4 Chapter 4: Review results – Research on effective formative assessment practices in Reading

4.1 Summary

The evidence for the impact of formative assessment on student learning showed disappointing results for reading achievement. We did not find reading studies where effects could be unambiguously ascribed to formative assessment. High levels of intensive and sustained professional development for teachers addressing all aspects of formative assessment (from selecting assessment tools to translating results and evidence-based differentiated instruction) are required before formative assessment practices in reading instruction show robust positive effects on student performance.

4.1.1 The evidence for the effectiveness of particular tools and resources

- No particular tool can be singled out as effective at this point.
- Many of the reviewed studies included some technology/software component. There is mixed evidence for the effectiveness of technology in improving student learning outcomes.

4.1.2 Implications for effective implementation of formative assessment practices

- It is likely that formative assessment implementation is successful only if it includes a significant professional development component.
- Professional development needs to include assessment and pedagogical content knowledge components, and both theory and practical examples.

4.1.3 Chapter overview

This chapter examines studies exploring formative assessments in Reading. It is divided into four sections: Part A explores formative assessment in reading employing progress-monitoring techniques alone; Part B investigates formative assessment in Reading employing progress-monitoring techniques combined with teacher professional development; Part C examines formative assessment in Reading employing interim assessments; and Part D explores the impact of the large-scale US ISI/A2i project on formative assessment in Reading**.**

The initial search identified 15 studies that met the search criteria, including a focus on improving reading skills. A senior Reading researcher read through all studies and excluded ten from the final pool. Studies were excluded if they:

- did not include identifiable formative assessments (Aydemir et al. 2013; Kingsley et al. 2015; Swapna et al. 2017; Wheldall et al. 2017)
- included a comprehensive intervention that changed the entire curriculum (Fantuzzo, Gadsden & McDermott 2011)
- added significant amounts of additional instructional time not available to the control group (Tyler et al. 2015)
- compared using e-portfolios (with feedback as one component of many) to not using e-portfolios (Abrami et al. 2013; Meyer et al. 2010)

- included an intervention in which formative assessments played a minor role (Albers & Hoffman 2012; Baker et al. 2018)
- focused on English language learners (Ponce et al. 2018).

The main reason for excluding those studies (with the exception of Ponce et al. where the sample did not come from our target population) was that the effects they reported could not be plausibly attributed to the possible formative assessment practices included. To some extent, this criticism also applies to most of the studies included and reviewed, as none present results that can be clearly attributed to formative assessments. We will return to this issue in the concluding discussion, but the immediate implication is that we were unable to organise the chapter according to quality of studies. Instead, we will review the remaining studies plus two groups of studies not initially identified, loosely grouped under the headings progress monitoring, progress monitoring and professional development, interim assessments, and ISI/2Ai.

4.2 Part A: Progress monitoring

Progress monitoring, and particularly the use of curriculum-based measurement mazes (CBM mazes; e.g. Deno, Mirkin & Chiang 1982), has an established history in Reading research (for reviews of earlier studies, see Fuchs & Fuchs 1986; for examination of their psychometric properties, see Stecker, Fuchs & Fuchs 2005; Ardoin et al. 2013). In general, most of the studies focused on low-achieving students and their results suggest progress monitoring can lead to better student performance (e.g. Fuchs et al. 1992), with the effect typically enhanced when progress monitoring is combined with sufficient support for teachers in interpreting the assessment results and differentiating their instruction accordingly (e.g. Espin et al. 2017; Fuchs et al. 1992; Zeuch, Förster & Souvignier 2017).

In the included studies, two from the same German research group examined the impact of additional online progress monitoring to students' reading fluency and comprehension. Förster and Souvignier (2014) investigated the effects of learning-progress assessments (LPA – their term for progress monitoring) and of LPA combined with student goal setting (LPA-G) on typical Year 4 students' Reading achievement, motivation and self-concept. Students in the LPA and LPA-G groups completed eight online reading assessments over a six-month period and their performances on standardised reading tests were compared to those of a control group receiving their usual Reading instruction. Online assessments were CBM mazes in which students read texts onscreen.

The first part of the test was a typical maze task in which every seventh word had been replaced with a choice between the correct word and two distractors. The students' task was to choose the correct words to fill the gaps as quickly as possible. The time needed to read the text was recorded as 'reading rate' and the number of correct selections was recorded as a measure of reading accuracy. In the second part, students answered 16 multiple-choice comprehension questions (comprising eight text-based recall and eight knowledge-based inference). At the end of each assessment, teachers received information on each student's reading rate, reading accuracy, and text-based and knowledge-based reading comprehension. LPA-G group students further set themselves a goal on how well they would perform before each of the eight LPA tests, reflected on their goal achievement afterwards, and considered the reasons for their perceived success or failure.

The results indicated the LPA group showed significantly greater growth in Reading achievement (as assessed by standardised tests) than the LPA-G or the control group (*d*s = 0.27 and 0.24, respectively), whereas the latter two did not differ. Students in the LPA group also improved significantly more over the eight online assessments than students in the LPA-G group. Further, LPA-G group's reading motivation and reading self-concept ratings were poorer at post-test than at pre-test and both differences were significant compared to the control group, with LPA group falling in between. Therefore, simple progress monitoring had a positive effect on Reading achievement but, when combined with the possibly more demanding goal-setting approach aimed at increasing self-regulated learning, the effect was reversed.

In the second study from the same group, Förster, Kawohl and Souvignier (2018) investigated the effects of learning progress assessments and differentiated instruction for Years 3 and 4 classrooms. The intervention combined (1) frequent assessment-based information about student progress in Reading (exactly as above), and (2) instructional materials for teachers to differentiate Reading instruction based on two established interventions (repeated reading and reciprocal teaching). Experimental group students completed the computerised CBM maze Reading assessments every three weeks and the results were used to indicate whether they should focus on practising reading fluency or comprehension strategies. Teachers were provided with materials to adapt instruction to the identified needs. The control group received their typical Reading instruction and had no access to additional assessments or instructional materials.

The results showed that, at the end of Year 3, the experimental group showed better reading fluency ($d = 0.30$) than the control group, and the difference remained at the end of Year 4 ($d = 0.31$). Growth in reading comprehension was not significantly different between the two groups. Additional analyses suggested the reading comprehension intervention (reciprocal reading) was implemented less frequently and with lesser fidelity compared to the reading fluency intervention (repeated reading), and this may have contributed to the lack of impact on comprehension. It remains to be seen if a more thorough professional development for teachers would make the reading comprehension intervention more frequently used and effective. We should also note that while reciprocal reading is widely used, the evidence base for it is relatively thin and it may not have been an optimal choice for reading comprehension intervention.

Hall et al. (2015) reported a related study with Years 6–8 US students that specifically contrasted online and offline CBM. While this study does not allow calculating an effect size for using formative assessments, it is interesting because of the specific comparison made. Both groups used an online software package called Strategic Reader (developed by CAST but currently unavailable at their website, cast.org) that provided students with a digital reading environment and multiple support features, such as text-to-speech translation, a dictionary, a glossary, text highlighting, embedded reciprocal teaching questions and an online forum for student-to-student and student-to-teacher discussions. The tool also includes CBM measures that assess oral reading fluency, accuracy, comprehension and comprehension strategies (summarising, predicting, questioning and clarifying – from reciprocal reading) similar to software used in the German studies. The program administers the measures and automatically calculates and displays the results to both teacher and student. All scores are displayed in graphic and tabular formats, allowing students and teachers to see how their performance has changed over time and determine whether progress is adequate. The teacher can view both individual student results and class results and can adjust the number of support features available to students.

In the study, Treatment 1 teachers and students used Strategic Reader for 11–12 weeks with all CBM features. In Treatment 2, Strategic Reader with offline CBM included progress monitoring using a traditional offline paper-and-pencil structure that required teacher administration, scoring and graphing. All teachers attended a two-day workshop before the study began. Thus, the only difference between the two conditions was whether the CBMs were administered online or offline, with offline administration significantly increasing teacher workload. Perhaps unsurprisingly, results indicated that, compared to Treatment 2 teachers, teachers in Treatment 1 viewed student data three times more frequently, designed interventions five times more frequently, and modified student supports 40 times more frequently. Student performance was assessed with standardised reading comprehension tests pre- and post-intervention. Unfortunately, the reported data and statistical analyses performed do not allow calculation of the effect of intervention on students. However, the provided means indicate somewhat better progress for students in Treatment 1 than in Treatment 2, and that the difference between the conditions was larger for students with learning difficulties than for students without learning difficulties.

In the final progress-monitoring study, Simmons et al. (2015) focused on students receiving remedial Reading instruction. The authors used data from a series of experimental studies examining the effectiveness of the Early Reading Intervention (ERI – a packaged reading intervention program for struggling early readers). Simmons et al. were specifically interested in whether adding ongoing assessments and instructional adjustments to the conventional ERI implementation would impact the intervention's effectiveness. Conventional ERI implementation includes four content mastery assessments, one at the end of each unit, but students typically stay in their groups throughout the year irrespective of assessment results. The adjusted ERI included an additional four curriculum-embedded measures (CEMs), one at the mid-point of each unit to evaluate students' mastery of skills taught during that period. Grounded in the theory of mastery learning, CEMs are formative assessments of recently taught content or skills designed to provide information to guide instructional modifications for individual students. In Simmons et al. (2015), research team members met with teachers after each assessment to make instructional adjustments (either no adjustments, repeat content or accelerate progression) based on student data.

The results indicated that, when compared to propensity-matched controls, students whose program was either decelerated with additional repetition of poorly learned content or accelerated due to quick mastery of content did significantly better at the end of the program across a variety of reading measures, with effect sizes varying from small to medium for the decelerated group and medium to large for the accelerated group. Therefore, this study indicates we can expect better learning outcomes when formative assessment leads to meaningful changes in instruction.

In summary, these progress-monitoring studies provide cautious support for formative assessment, particularly online formative assessment. However, it is likely the impact would depend on what supports (online or in person), are provided to teachers for making effective instructional adjustments and whether those adjustments are based on solid evidence, as exemplified by repeated reading in Förster, Kawohl and Souvignier (2018) and mastery learning in Simmons et al. (2015).

4.3 Part B: Progress monitoring combined with professional development

The two progress-monitoring studies reviewed in this section included a more explicit focus on professional development. Witmer et al. (2014) focused on primary students' comprehension of informational texts. They assigned volunteer teachers randomly to experimental and control groups, with teachers in the experimental group given ongoing professional development on how to administer and interpret Years 1 and 2 classrooms results from the Concepts of Comprehension Assessment (COCA). Teachers in the experimental group administered the COCA to a subset of six children in their classrooms at the beginning, middle and end of the school year, whereas researchers administered COCA to additional students and all control students.

The COCA is an individually administered test designed to measure Years 1 and 2 students' knowledge and skills for comprehending informational text. It is intended to help inform instruction. Four subscales – Vocabulary, Text Features, Graphics in the Context of Text and Comprehension Strategies – include items measuring students' skills and knowledge for comprehending informational text. Text and questions are read to students to facilitate accurate measurement of knowledge and skills for comprehending informational text separate from their decoding skills.

Witmer et al. (2014) reported that, while the two groups were equal at the beginning of the study, middle- and end-of-year COCA scores were significantly higher for the experimental group than for the control group. To further validate the results, the researchers used a writing assessment to assess the transfer of skills. These results also indicated a significant – albeit smaller – effect favouring the experimental group. In addition, experimental group teachers reported changing their instruction on the basis of assessments and rated COCA professional development as having a positive impact on student learning. Witmer et al. (2014) interpreted their results to support the notion that professional development in the administration and use of COCA data can have a positive effect on student learning of related skills.

Brookhart, Moss and Long (2010) also examined the impact of professional development on formative assessment on the performance of students attending remedial Reading classes in primary schools. In this study, six teachers attended a year-long professional development program that focused on formative assessment but provided no specific tools to use. Another group of remedial Reading teachers (n = 11) functioned as the control group. The impact of professional development on experimental teachers' formative assessment practices was loosely documented as improved while the impact on students was assessed with standardised tests. In these analyses, students of the six teachers attending professional development were treated as the formative assessment group and students of the 11 control teachers were treated as the comparison group, with Kindergarten and Year 1 students receiving different age-appropriate assessments. The results showed no impact of formative assessment professional development on Kindergarten children's learning, and a small but significant effect on Year 1 children's learning. As the professional development given to teachers emphasised reflective professional inquiry, the results could be attributed to either formative assessment practices or a more general effect of professional development.

To summarise, the above results broadly align with a summary of the CBM research provided by Stecker et al. (2005), indicating that training teachers to monitor student progress and make associated instructional decisions can have a positive effect on student learning. It is also possible that the second component was not sufficiently present in the Brookhart, Moss and Long (2010) study, reflected in the minimal effects.

4.4 Part C: Interim assessments

The completed searches for this review failed to locate studies that may be relevant to assessing formative assessment arguments because of different terminologies used. The previous meta-analyses our search strategy followed would not have included them either. While, strictly speaking, these studies do not allow conclusions about the effectiveness of formative assessment practices *per se*, they do provide important information for those planning similar kinds of projects under the formative assessment framework.

The first set of additional studies are those focused on large-scale interim or benchmark assessments that are increasingly common components of data-based decision-making practices in North American school districts. Interim assessments are periodic diagnostic assessments typically administered three or four times during the school year to help teachers use evidence to differentiate instruction and make better instructional decisions, often in preparation for year-end summative assessments used as accountability measures. The rationale behind these assessments is that, if objective data on student performance is frequently available for teachers, it can be used to better understand their students' learning and adjust instruction early, prior to year-end accountability tests. Some interim assessment software, such as mCLASS (https://www.amplify.com/ programs/mclass/), provide tools to assess literacy and numeracy growth from K to Year 3 and allow teachers to monitor individual student progress more frequently by including in-between assessment tasks (e.g. CBM mazes). In this approach, teachers receive whole-class and individual progress reports regularly throughout the school year and can further choose to assess students more frequently when concerns arise. Therefore, an interim assessment can be indistinguishable from the progress-monitoring approaches reviewed above, although typically the frequency of assessments would be lesser in interim assessment programs and the content of assessments is closer to the curriculum content than is sometimes the case with more skills-focused progress monitoring.

Early quasi-experimental studies of interim assessments produced inconclusive findings. Henderson et al. (2007) examined pilot data for Massachusetts' quarterly benchmark assessments and found no statistically significant or substantively important differences between program and comparison schools. Quint, Sepanik and Smith (2008) used an interrupted time-series design to investigate the impact of Boston's Formative Assessments of Students Thinking in Reading (FAST-R) on the Massachusetts' State Test and the Stanford Achievement Test. Their results were generally positive, but the differences were not statistically significant. In a large-scale cluster randomised experiment, May and Robinson (2007) evaluated the impact of Ohio's Personalized Assessment Reporting System (PARS) on the Ohio Graduation Tests (OGT). PARS includes repeated assessments and provides online reports on test outcomes. The authors compared Year 10 student achievement between 51 treatment and 49 control schools during the pilot year. They found PARS did not have a significant impact on students' achievement in OGT. However, among students who did not pass OGT the first time, the PARS group was more likely to attempt the test a second time and received significantly higher scores than students in the control group.

Four recent large-scale randomised controlled trials have compared interim assessment programs to business-as-usual controls. Carlson, Borman and Robinson (2011) examined the impact on student achievement of an intervention developed by the Center for Data-Driven Reform in Education (CDDRE). The CDDRE intervention is a data-driven decision-making process that emphasises instructional change based on quarterly benchmark/interim assessment results (based on CDDRE-designed 4Sight assessments). CDDRE consultants work with participating districts to implement quarterly student benchmark assessments and provide district and school leaders with extensive training on interpreting and using the data to guide instruction. Carlson, Borman and Robinson (2011) analysed data from a multistate, district-level cluster randomised experiment to investigate CDDRE's potential benefits. The sample included more than 500 schools in 59 school districts across seven US states. The first year of experimental results indicated statistically non-significant (g =0.14) positive effects on Reading scores across Years 3–8. Unfortunately, data was not examined separately across different years.

A follow-up study by Slavin et al. (2013) investigated CDDRE's impact over a four-year period on more than 600 primary and middle schools. Multilevel models were used to analyse its impact on Year 5 and Year 8 Reading achievement. The results showed that positive effects accumulated at the school level across the four years of implementation for Year 5 students. Effects were statistically significant after four years (d =0 .49) but not after one, two or three years. For Year 8 students, effects were significant in the first (d = 0.26) and second (d =0.23) years but not the following years. Slavin et al. further examined reading programs used in schools and concluded that positive effects were driven by those adopting an evidence-based reading program. Schools that did not use evidence-based reading programs did not differ from control schools. In other words, observed effects probably represent an interaction between more-frequent assessments and teachers' ability to adjust instruction within the confines of an evidence-based program.

Cordray et al. (2012) examined the impact of Measures of Academic Progress (MAP, https://www.nwea.org/the-map-suite/) interim assessments on Reading achievement. MAP is a widely-used, commercially-available program incorporating computer-adaptive assessments and professional development in differentiated instruction. In Cordray et al. (2012), Years 4 and 5 classes in 32 elementary schools across five Illinois school districts were randomly assigned to treatment and control conditions (with one year per school assigned to treatment and the other to control). The analyses included more than 170 teachers and nearly 4,000 students. The study found MAP was implemented with moderate fidelity, but MAP teachers were not more likely to differentiate instruction than their non-MAP colleagues. Further, researchers found no statistically significant differences in reading achievement at either year on the Illinois State Achievement Test or on the MAP composite score (all effect sizes were effectively zero).

Finally, Konstantopoulos et al. (2013, 2016) have published two analyses of Indiana's Diagnostic Assessment Tools' (DAT) impact on student performance. DAT was rolled out in 2008, making Indiana the first US state to commit to technology-supported interim assessments. DAT consists of two commercial products aligned to Indiana's curriculum and academic content standards. For Years K–2, the mCLASS assessment conducts Reading assessments one-on-one. A student performs tasks while the teacher records characteristics of their work. Teachers are guided through the assessment process by an interface and can immediately view results and compare them to prior performance. At any point, teachers are able to monitor individual student classroom progress using short one-on-one tasks and can see those results graphically linked to previous results. For Years 3–8, Indiana selected CTB/McGraw-Hill's Acuity assessments. These consist of 30- to 35-item multiple-choice online tests to be completed within a class period. The Acuity assessments feature two types – diagnostic and predictive – with most schools selecting only one type. Diagnostic Acuity focuses on identifying specific students' instructional needs. Predictive Acuity forecasts student performance on Indiana's state summative test. Both types allow teachers to construct progress-monitoring assessments from banks of aligned items. Instructional resources – packaged student exercises to practise skills or explore others – are available and may be assigned to students directly from Acuity's computerised report displays.

In the first Indiana study, Konstantopoulos, Miller and van Ploeg (2013) used first-wave data from a large-scale (500 schools with 220,000 K–8 students) school-level cluster randomised experiment to examine the impact of DAT on Mathematics and Reading achievement. A train-the-trainer model was used when DAT was introduced to the first set of schools, with one to four teachers from each volunteering school received two to three days of training before the school year started. Teachers trained in the summer received a supply of materials to train their colleagues and were expected to conduct two to three training sessions at their schools during the first six months of the program. Results indicated the treatment effects were positive, but not consistently significant. Treatment effects were smaller for younger students (i.e. Years K–2), and larger in upper years (i.e. Years 3–8). Significant treatment effects were reported, especially for Year 3 and 4 Reading scores.

In a follow-up study, Konstantopoulos et al. (2016) used a second wave of data that included over 30,000 students and 70 schools across Indiana. The authors randomly assigned 36 schools to the treatment condition and 34 to the control condition (who did not get access to DAT). In this school-level randomised experiment, results from K to 2 revealed students in treatment schools were at a disadvantage compared to those in control schools. Years 3–8 students showed no effects. The authors speculated that the lack of ongoing professional development and specific suggestions about differentiated instruction may have contributed to negative effects in a changing assessment context.

In summary, US states have spent millions of dollars on the development and implementation of interim assessments that have failed to show consistent educational benefits. There are naturally many potential reasons for this, including the sparse nature of the assessments and their apparent connection to statewide accountability assessments that may affect teacher attitudes negatively (see e.g. Keuning, Van Geel & Visscher 2017). As the implementation of interim assessment programs has usually been large scale, it is also likely that, when professional development was offered to the schools, it has been assessment-tool focused and not sufficient for change in instructional practices. To achieve meaningful instructional change, principals and teachers need additional skills and knowledge. It is questionable that, for example, a train-the-trainer model focused on assessments increases schools' in-house competencies enough to result in more effective instructional adjustments. Interim assessments can identify areas for improvement but only teachers can implement the instructional strategies and practices that will bring about targeted improvements.

4.5 Part D: ISI/A2i project

ISI/A2i is an ambitious large-scale US project by Connor and Morrison (e.g. Connor et al. 2007, 2013; Connor & Morrison 2017) focused on improving K–3 student reading skills. ISI/A2i uses continuous Reading assessments and interpretative software to deliver instructional guidance to teachers. While this project does not refer to formative assessment or assessment for learning, its general approach is very much aligned with formative assessment as defined in this review. We will first describe the intervention in more detail due to its unique combination of continuous assessments and instructional guidance to teachers. We will then review the evidence collected across a series of studies, conducted mostly by Connor and Morrison (see review in Connor & Morrison 2017).

The Individualizing Student Instruction (ISI) intervention has three major components. The first comprises the Dynamic Forecasting Intervention (DFI) model and algorithms that compute recommended amounts (min/day) and types of literacy instruction (four categories derived by crossing teacher-managed versus student-managed and meaning-focused versus code-focused) based on each student’s unique vocabulary, decoding and comprehension skill profile. The DFI algorithms are essentially reverse-engineered regression equations based a series of correlational studies (e.g. Connor, Morrison & Katch, 2004; Connor, Morrison & Petrella, 2004; Morrison & Connor 2002) predicting student outcomes on the basis of earlier performance levels and instruction provided. Instead of solving for student outcomes, the algorithms solve for amounts and types of instruction required for children to read at or above grade level by the end of the school year.

The A2i software displays DFI algorithm outcomes for teachers. Its key component is a classroom view displaying the recommended type of instruction amounts for every student. It also suggests homogeneous skill-based learning groups. Teachers then select the number of groups they want to use and can also change students’ suggested groups. A2i keeps track of assessment information for all students in the class. It also keeps separate individual student information pages that include progress-monitoring graphs and other key information to help teachers interpret assessment results. A2i indexes the core curriculum to the four types of instruction, allowing teachers to use this information when planning lessons in the program and adjust existing teaching materials to meet individual students’ learning needs. More information and the commercialised A2i software are available from United2Read (http://www.united2read.org/learningovations/).

ISI/A2i implementation studies have included substantial professional development. In most studies, teachers attended a half-day workshop to learn about ISI intervention and how to use A2i software before the beginning of the school-year. This was followed by monthly communities of practice meetings in school teams to discuss A2i and how to implement ISI/A2i to deliver A2i-recommended amounts of evidence-based Reading instruction to each student. Bi-weekly classroom-based support was provided during the literacy block to further support implementation, troubleshoot and model evidence-based practices. In summary, despite the A2i software’s highly automated nature, relatively intensive professional development was deemed necessary to promote proper software use and recommended evidence-based code- and meaning-focused instructional practices.

Evidence for ISI/A2i comes from a series of randomised controlled trials. In the first study (Connor et al. 2007), Year 1 students across ten schools were randomly assigned to ISI/A2i or a delayed treatment control (which received A2i and professional development the following year). Results indicated schools using ISI/A2i showed greater gains in passage comprehension, controlling for vocabulary and socioeconomic status. Moreover, results suggested the more teachers used A2i, the greater the difference between the treatment and control conditions. A second study with Year 1 students using the same design produced similar results (Connor et al. 2011). Moreover, when the number of minutes per day each student spent in the four types of instruction was computed and compared to the recommended amounts, the effects of ISI/A2i increased as the difference between the amount of instruction students received and A2i recommended amounts decreased.

In the next set of studies, Connor et al. (2013, 2014) randomly assigned teachers within schools to either the ISI/A2i condition or an alternative treatment condition (Mathematics) and provided the same amount of professional development to both groups. In Connor et al. (2013), Year 1 students whose teachers were assigned to the ISI/A2i condition showed greater gains in word reading and passage comprehension compared to students in the Mathematics condition. In Connor et al. (2014), no treatment effects were found for Year 2 students. Based on classroom observations, researchers determined that although implementation was fine, the DFI algorithms used in Year 2 were flawed. When these were calibrated to fit the new Year 2 observation data, the authors found a significant effect of ISI/A2i on students' reading skills (both word reading and comprehension) compared to the Mathematics comparison group (Connor et al. 2013). The same study also showed effects on Year 2 students when compared to the Mathematics condition, similar to Connor et al. (2011) where the comparison group received vocabulary intervention.

In all the above studies, the comparison intervention was, at best, attention control and the observed effects cannot be assigned explicitly to the use of ISI/A2i (and by extension, to the included formative assessment practices). Al Otaiba et al. (2011), however, conducted an RCT with Kindergarten classes in which teachers within schools were randomly assigned to ISI/A2i or to an alternative professional development control that showed them how to individualise literacy instruction. The alternative professional development control had no access to A2i. Results showed students whose teachers used A2i made significantly greater gains in early reading skills compared to those whose teachers received literacy professional development but not A2i. This study provides the strongest evidence for the role of software as an active ingredient in ISI/A2i.

These studies generally yielded effect sizes (*d*) of between 0.2 and 0.4 (0.5 in Al Otaiba et al. 2011). Connor et al. (2013) were interested in the possible accumulation of effects across Year 1 to Year 3. They assigned Year 1 classes randomly to ISI/A2i or a Mathematics condition. When students entered Year 2, their classes were again randomly assigned to conditions. This process was repeated in Year 3, resulting in groups of students who had (1) received ISI/A2i across all three years, (2) only in one or two of the three years, or (3) not at all. Growth-curve modelling showed the more years students spent with teachers using ISI/A2i, the better their Reading performance at the end of Year 3 compared to students who had attended comparison classes in all three years. ISI/A2i students in all three years were a full year ahead in their reading skills ($d = 0.77$). Further, Connor et al. (2013) noted that Year 1 attendance appeared to be necessary but not sufficient to support greater reading skill by the end of third grade.

In the field of Reading research, the ISI/A2i research program by Connor and Morrison is unparalleled both in terms of the number of studies and of the technology focus. The studies can be criticised for their choice of attention or wait comparison conditions; with the possible exception of Al Otaiba et al. (2011), none compared ISI/A2i to anything resembling the 'gold standard' or best evidence-based practice. Furthermore, the contrast in these studies was never between formative assessment and no formative assessment, and thus they do not allow attribution of observed effects to the included formative assessment practices. This criticism, however, applies to all the studies reviewed here. We will return to this point below.

4.5.1 Discussion

The above studies examined a wide variety of students from K to Year 8, both with and without learning difficulties. Most of the studies were conducted in the US, where both formative (in particular, progress monitoring) and summative accountability assessments have been in the forefront since the 2002 *No Child Left Behind Act*. Further, the studies examined a wide variety of general reading skills, with disciplinary skills the focus in only one study (Witmer et al. 2014). For the most part, the studies examined current performance levels of typically developing students, with assessment information limited to three aspects of reading skills at the most (e.g. fluency and comprehension, or vocabulary, decoding and comprehension). Despite limited foci, without exception the selected foci seem to align with what we know about reading development and how to assess it. In other words, the Reading studies cannot be criticised for using inappropriate or outdated task or cognitive models. When instructional guidance was provided to teachers, for the most part it also seemed to be based on evidence-based practices. In summary, the theoretical foundations of the above studies appear solid, but the results are best described as disappointing.

Before speculating on reasons for less than impressive results, we should note some issues with the study designs and reporting. First, we were unable to locate any Reading studies in which the effect could be unambiguously attributed to formative assessment. This raises the question about previous meta-analyses that have reported variable effectiveness estimates for formative assessments in Reading.

For example, Klute et al. (2017) reported an average effect size of 0.22 calculated across 12 effects from seven studies. While some of their studies were published before 2000 (almost all of these studies focused on using CBM with struggling readers and are summarised in Stecker et al. 2005) and therefore not included in this review, others were the same studies we ruled out because they did not provide valid information on the effectiveness of formative assessments (e.g. Abrami et al. 2013; Meyer et al. 2010). This raises the question: How generalisable are arguments made on the basis of meta-analyses if one research group deems a study to be about formative assessment while another does not? Clearly, the pool of studies is very diverse, and we must be cautious about any generalisations.

The second design, or perhaps reporting, issue we would like to raise is the lack of information on assessment practices and instruction provided to control groups. This information was often missing from reports altogether, leaving open the option that control participants were also enjoying formative assessment benefits, just not the ones examined. It is clear that, if a study wants to establish the effectiveness of formative assessment practices, control group practices require description.

For example, when classes are randomly assigned to formative assessment intervention and wait-control conditions, potentially effective practices continue uninterrupted in the wait-control classes while formative assessment classes experience disruption, albeit a potentially positive one. Additionally, it is no longer acceptable to compare reading intervention to mathematics intervention (attention control) and fail to report on reading instruction in control classrooms. Better yet, as we already have extensive information on best practices in Reading and Mathematics, the value of a new program needs to be established against the best existing alternatives tested, accepted and adopted by teachers.

Design issues aside, what factors could explain the disappointing results and lack of robust effects of interventions that seemingly increase formative assessment practices? We have already alluded to some possible factors above, such as professional development and teacher support. It seems clear teachers need professional development in administering the assessments, interpreting results and translating information obtained into effective instructional adjustments. This last step is not a trivial matter, as evidenced by Connor et al.'s 2014 highly developed evidence-based algorithm failure (see also Ardoin et al. 2013). Most of the studies above reported some professional development for teachers, but it was often unclear whether that only included assessment procedures or also extended to interpretation of results and instruction. In general, teachers' ability to differentiate instruction can be limited (e.g. Fuchs & Vaughn 2012). Unless this problem is addressed, no amount of additional assessment information will lead to better instruction. Against this background, it is informative that when schools were either already using evidence-based programs that limited the instructional choices to those with a higher likelihood of impact, or the researchers (Simmons et al. 2015) or software (Förster, Kawohl & Souvignier 2018; ISI/A2i studies) provided guidance in implementing evidence-based instruction, the results tended to be better. It seems that pedagogical content knowledge is as critical as assessment knowledge for positive effects to emerge. In particular, this may have been a problem for interim assessment programs that reported limited professional development.

A related issue is teacher workload. To simplify, the more time required to administer, score and interpret assessments, the less time is left for modifying instruction (Hall et al. 2015; Al Otaiba et al. 2011). Perhaps there is also an underlying commitment issue, such as if formative assessments are seen as an additional task to be completed rather than an integral part of instruction. Whatever the root cause, it is likely that the less workload is increased, the more impact formative assessments can have. Perhaps this realisation is behind the pervasiveness of software-driven formative assessments in recent studies; formative assessment developers are clearly trying to make their approaches intuitive, as automated as possible and easy to embed in daily practices.

To conclude, it seems clear that high levels of intensive and sustained professional development for teachers addressing all aspects of formative assessment (from selecting assessment tools to translating results and evidence-based differentiated instruction) are required before formative assessment practices in Reading instruction show robust positive effects on student performance.

5 Chapter 5: Review results – Research on effective formative assessment practices in Writing

5.1 Summary

Evidence indicates formative assessment has a strong impact on student writing quality, particularly when cognitive strategies and self-regulation around writing choices are supported by assessment tools. Formative assessment also has a positive effect on student's understanding of how to make grammatical choices when supported by metalinguistic conversations with a teacher.

5.1.1 Evidence for the effectiveness of particular tools and resources (including online tools)

The Using Sources Tool was useful. It aided in facilitating teachers' evaluation and strategising.

5.1.2 Implications for effective implementation of formative assessment practices

We found explicit teaching of cognitive strategies related to writing choices for planning, revising and/or editing improves Writing outcomes. Self-regulation, often characterised by explicit teaching of writing strategies with democratic student participation in the assessment process (self-assessment and goal identification), also indicates improvements in Writing outcomes.

There is a significant effect on students' ability and understanding of grammatical choices when teachers conduct metalinguistic conversations during Writing lessons. Furthermore, when students understand how their writing choices affect meaning, they produce clearer and more complex work. Writing choice lessons should, therefore, be contextualised.

5.2 Chapter overview

The following chapter examines studies exploring formative assessment approaches to improve writing in K–12 classrooms. Of the seven studies identified in the original database search, two (Abrami et al. 2013; Meyer et al. 2010) were excluded because they did not isolate the role of formative assessment. These studies compared instructional approaches with and without the use of e-portfolios, with feedback as an intervention component.

We retained five writing studies to include in this review. Of these, two were conducted in the US (Campbell & Filimon 2018; Gallagher, Arshan & Woodworth 2017), two in the Netherlands (Faber & Visscher 2018) and one each in Australia (Fletcher & Shaw 2012) and Canada (Meyer et al. 2010).

5.3 Overview of studies

5.3.1 Immediate feedback digital assessment tool for spelling (Fletcher & Shaw 2012)

Writing interventions demonstrate the potential value of using formative assessment to improve students' learning and writing performance. For example, Fletcher and Shaw (2012), used a student-directed assessment process (SDA) to understand students' engagement and learning outcomes when provided with opportunities to identify their own learning goals and determine assessment criteria through planning templates that facilitate self-reflection on learning progress. The mixed methods study was conducted at a primary school in the Northern Territory in response to the 2008 introduction of the National Assessment Program in Literacy and Numeracy (NAPLAN). Results show a statistically significant difference between teacher-directed and student-directed groups.

5.3.2 Strategy-focused writing instruction: guided feedback and peer feedback (Campbell & Filimon 2018)

Campbell and Filimon (2018) discuss the effects of strategy-focused writing instruction on argumentative essay writing. Their pedagogical approach encourages the provision of teacher-guided feedback and student engagement in peer feedback practices. Data analyses collected from 47 linguistically diverse seventh-grade students in the US reveal students' overall writing performance increased significantly from pre-test to post-test in Evidence and Elaboration and Conventions of Standard English, but not in Purpose, Focus, and Organization.

5.3.3 Peer assessment and self-assessment practices, on the development of self-regulation and motivation in writing (Meusen-Beekmana, Brinke & Boshuizen 2016)

The studies described above explored the effects of formative assessment as part of a large teaching intervention. By contrast, Meusen-Beekmana, Brinke and Boshuizen (2016) investigated the direct effect of formative assessment on the development of self-regulation and motivation in writing among sixth graders in the Netherlands, including peer assessment and self-assessment practices.

Peer assessment and self-assessment processes involved students' active contribution to setting criteria for evaluation, using a checklist to monitor progress. In addition, students could self-reflect about the task and the requirements for improvement following feedback. Results showed a significantly better effect on student writing self-regulation and motivation.

5.3.4 Learning through writing (Klein & Rose 2010)

Writing is mainly characterised as 'learning to write' among the four reviewed educational interventions. However, it is also seen as 'writing to learn', and as a means to extend and deepen students' knowledge across subject matters (Graham & Perin 2007, p. 31). In a study by Klein and Rose (2010), writing was applied as a tool for enhancing students' learning of content material. The study aimed to develop instructional design models to improve students' ability to learn through writing. It consisted of two phases, focusing on argument writing and explanation writing. The instructional design's goals and elements comprised frequent writing in the content area, teaching cognitive strategies about argument and explanation writing, providing feedback to students, evaluating and revising for content learning, and assessment design to support self-evaluation.

A series of post-test activities showed the experimental group had a significantly greater ability to learn while writing. They demonstrated greater argument genre knowledge, explanation genre knowledge and explanation text quality, when compared with the control group. However, it is difficult to decipher the direct contribution formative assessment had on student learning because of the intervention's complexity and number of simultaneous changes.

5.3.5 Teacher professional learning and the use of tools in the delivery of assessment (Gallagher, Arshan & Woodworth 2017).

When the US District of Columbia adopted new college and career-ready standards in English language, the Arts and Mathematics, the greater emphasis on argument writing meant teachers sought more support (Gallagher, Arshan & Woodworth 2017). The study's findings showed professional development and a 'Using Sources Tool' helped teachers assess student work, strategise which skills to focus on in applied instruction (Gallagher, Arshan & Woodworth 2017), and had positive significant effects on teacher practice and student source-based argument writing.

5.3.6 Discussion

Two key findings from these studies relate to improvement in Writing outcomes through formative assessment: (1) explicit teaching of cognitive strategies for planning, revising and/or editing, and (2) self-regulation with centralised democratic student participation in the assessment process. These findings are consistent with Graham and Perin's (2007) work on the most effective writing interventions. Self-regulated strategies (often characterised by explicit teaching of writing strategies and involving self-assessment processes and goal identification) are an effective approach for teaching writing strategies. Furthermore, Myhill, Jones and Wilson (2016) and Newman and Myhill (2016) provide large-scale empirical evidence that, when teachers use language knowledge to have metalinguistic conversations (metatalk) during Writing lessons, there is a significant effect on students' understanding of how to make grammatical choices. Myhill, Jones and Wilson's (2016) intervention highlights the importance of teaching about writing choices in context.

In Australia, Mackenzie, Scull and Bowles (2015) analysed 500 texts from 250 early writers. They showed students produced clearer and more complex texts when they understood how their writing choices affected meaning in their work. The findings from this review indicate there is a strong impact on the quality of student writing when cognitive strategies and self-regulation around writing choices are supported through formative assessment.

6 Chapter 6: Review results – Research on effective formative assessment practices in Science

6.1 Summary

Formative assessment in Science is most effective at improving student learning outcomes when it involves regular, targeted and embedded feedback linked to learning progressions. Embedded formative assessment within a curriculum-based learning sequence has the ability to promote deep understanding when it provides feedback to students about how to collect and analyse data about conceptual change.

6.1.1 Evidence for the effectiveness of particular tools and resources

Formative assessment in Science using online tools is most effective when it involves:

- targeted, timely and relevant student feedback related to the concepts/content being studied
- opportunities for students to use the tools as frequently as desired
- comments and feedback that are scientifically accurate and follow an established learning progression.

6.1.2 Implications for effective implementation of formative assessment practices

- Elaborated conversations/discussions are needed to elicit student preconceptions.
- Informative and motivating feedback supports learning for students with poor language proficiency (Year 3).
- Self-assessment using a combination of scripts and rubrics has the potential to improve student performance and achievement in Years 7–8 Science.
- A combination of group discussion, feedback from peers and feedback from teachers has the potential to improve Year 9 student Science performance.

6.2 Chapter overview

This chapter examines studies exploring formative assessments in Science. It is divided into two sections: Part A explores formative assessment in Science in general, and Part B looks specifically at online formative assessment in this domain.

Learners bring a range of incorrect or partially incorrect conceptual understandings and explanations about everyday scientific phenomena to the classroom. Unless these ideas are uncovered, explored and explained during learning experiences they endure and confuse future concept acquisition (Duit & Treagust 2003). To support learning in Science, it is critical that individual preconceptions are identified at the beginning of a study period, that repeated assessments are used to monitor learning progressions, and that individual feedback on current conceptual understandings are given and matched with corresponding learning activities to challenge specific misconceptions.

The literature suggests one way to achieve this is through effective formative assessment. This chapter reports findings of a review of international empirical research exploring formative assessment practices in K–12 Science contexts over the last ten years. The review focuses on the evidence base for the impact of formative assessment on teaching practice, student learning progress and outcomes in Science, the use and impact of tools and resources that support teachers' professional judgements of learners' needs, and the implementation of formative assessment practices in Science classrooms K–12.

6.3 Overview of studies

Based on the initial literature search, 17 Science papers were identified. All were initially reviewed to confirm appropriateness for inclusion; seven were deemed unsuitable because the studies did not focus on formative assessment (Anderson & Barnett 2013; Chase & Klahr 2017; Fagella-Luby et al. 2016; Olakanmi 2017; Thoron & Rubenstein 2013; Uzezi 2017) or used inappropriate methods to analyse the data (Bartholomew, Strimel & Yoshikawa 2019).

Of the remaining ten international studies (see Table 4), two were conducted in Asia: one in Singapore (Soong et al. 2010) and one in Taiwan (Wang 2010); three in Europe: one each in Germany (Decristian et al. 2015), the Netherlands (Vogelzang & Admiraal 2017) and Spain (Panadero, Tapia & Huertas 2012); and four in North America: one in Canada (Resendes et al. 2015) and three in the United States (Terrazas-Arellanes et al. 2018; Yin et al. 2015; Zhang & Misiak 2015). Of these studies, three were identified as having a 'high-quality' research design (Resendes et al. 2015; Wang 2010; Yin et al. 2015), with two using online digital technologies (Resendes et al. 2015; Yin et al. 2015). Seven studies were identified as 'low quality' (Decristian et al. 2015; Panadero et al. 2012; Soong et al. 2010; Terrazas-Arellanes et al. 2018; Vogelzang & Admiraal 2017; Zhang & Misiak 2015; Zucker, Kay & Staudt 2014) with three using online technologies (Soong et al. 2010; Terrazas-Arellanes et al. 2018; Zucker, Kay & Staudt 2014).

High-quality studies used a robust experimental design and appropriate analytical techniques where treatment effects were clearly described and could be specifically attributed to the formative assessment intervention. Low-quality studies also used a robust experimental design and appropriate analytical techniques, but the treatment effects lacked transparency and therefore the necessary details to reproduce findings. Treatment effects were either incompletely described and/or could not be attributed to the formative assessment intervention. This highlights the need for more-carefully designed formative assessment studies in the future, so that the contribution of formative assessment to student learning can be identified.

The majority of the studies focused on specific Science content, such as photosynthesis (Wang 2010), lactic acid (Vogelzang & Admiraal 2017), growth and change in animals (Resendes et al. 2015) or on broader scientific concepts such as floating and sinking (Decristian et al. 2015; Yin et al. 2015). Some studies also reported on scientific or instructional processes, such as metacognition (Vogelzang & Admiraal 2017), self-regulation and self-efficacy (Pandero et al. 2012) and peer feedback (Resendes et al. 2015; Vogelzang & Admiraal 2017**).** Two studies had a particular interest in examining the effects of interventions for low-performing students (Wang 2010; Terrazas-Arellanes et al. 2018) or students whose first language was not English (Terrazas-Arellanes et al. 2018). Some studies compared different assessment types, such as group discussion, feedback from peers and feedback from teachers (Vogelzang & Admiraal 2017) or the impact of grading/feedback approaches on student performance (Zhang & Misiak 2015). Yin et al. (2015) investigated the effect of formative assessment on learners through a scaffolded series of cognitive conflict episodes. Other studies investigated specific formative assessment programs or tools (such as Resendes et al. 2015; Wang 2010; Soong et al. 2010; Terrazas-Arellanes et al. 2018; Zucker, Kay & Staudt 2014).

6.4 Part A: Formative assessment in Science

This section details findings from high (n = 1) and low (n =5) quality studies. It includes those where students are engaged in formative assessment activities during Science lessons.

Table 4: Summary of Science papers reviewed.

Author/s and year of publication	Country in which study was conducted	Student age group	Topic	Group/ individual feedback (G/I)	Online or not (Y/N)	High/ low quality study (H/L)
Decristian et al. (2015)	Germany	Year 3	Floating and sinking	I	N	L
Panadero et al. (2012)	Spain	Years 9–10	Describing landscapes	I	N	L
Resendes et al. (2015)	Canada	Year 2	Growth and change in animals: birds and salmon	G	Y	H
Soong et al. (2010)	Singapore	Year 9	Air pressure and temperature	I	Y	L
Terrazas-Arellanes et al. (2018)	United States	Years 6–8	Wide range	I	Y	L
Vogelzang and Admiraal (2017)	Netherlands	Year 9	Lactic acid	I and G	N	L
Wang (2010)	Taiwan	Year 6	Photosynthesis	I	Y	H
Yin et al. (2015)	United States	Year 6	Floating and sinking	I and G	N	H
Zhang and Misiak (2015)	United States	Year s7–8	Magnetism, electricity, astronomy	I	N	L
Zucker, Kay and Staudt (2014)	United States	Years 8–9	Motion	I	Y	L

6.4.1 High-quality study

6.4.1.1 Formative assessment to promote cognitive conflict (Yin et al. 2015)

Yin et al. (2015) embedded cognitive conflict into a series of formative assessments aligned to the topic 'why things float and sink'. The study measured conceptual change and achievement in Year 6 (n= 52) students in a US charter school. Students were randomly assigned to a control or experimental group. Both groups were taught about sinking and floating by the same teacher using identical curriculum materials and activities, with 12 investigations over 12 weeks. Feedback to the experimental group focused on how the investigations were conducted, interpretation of the data collected by the class and how that data provided evidence to support their conceptions. Control group students did not receive structured experiences aimed at addressing misconceptions. In weeks four, seven and ten, experimental students participated in a 'predict–observe–explain' (POE) assessment activity followed by a challenge question. Overall, there was a significant effect on conceptual change and achievement in the experimental group compared with the control group – they scored significantly higher on diagnostic conceptual items, displayed fewer misconceptions when answering short-answer questions, and scored higher on a performance assessment.

Year 6 students who experienced embedded formal formative assessment (that is, feedback on how to collect and analyse data about conceptual change) within a curriculum-based learning sequence on floating and sinking demonstrated, on average, greater conceptual change than those who did not. The overall treatment effect on conceptual change was significant after controlling for pre-test misconception score (Wilks' l = 0.77, F (3, 43) = 4.30, p = 0.01, partial etasq = 0.23). Useful information from the study for teacher practice includes: (1) embedding formative assessments to create an evolving series of challenging activities with feedback promoting cognitive conflict, (2) designing formative assessments that align with expected learning progressions, and (3) guiding students to discover the weakness of their misconceptions through discussion and investigations.

6.4.2 Low-quality studies

6.4.2.1 Formative assessment (FA) versus scaffolding instructional discourse, peer-assisted learning and inquiry-based instruction (Decristian et al. 2015)

Decristian et al. (2015) conducted a large-scale study (n = 1,070) on Year 3 German students' conceptions of floating and sinking. They compared the use of three instructional approaches: scaffolding instructional discourse (SID), formative assessment (FA) and peer-assisted learning (PAL) with standard inquiry-based Science instruction (the control group). Fifty-four teachers in 39 schools participated in the study. Teachers in the formative assessment group received information regarding the design and use of diagnostic tasks to elicit students' preconceptions and evaluate their current levels of conceptual understanding. Furthermore, teachers learned how to provide informative and motivating feedback to their students. Controlling for pre-test, students in the formative assessment group scored significantly higher in the post-test than students in the control group. SID and PAL interventions yielded non-significant results. The researchers did not report on the specific strategies used by teachers to elicit students' preconceptions or provide examples of 'informative and motivating feedback' given to students. The absence of these details means the study cannot be replicated and translation of treatment effects to school classrooms cannot occur.

Year 3 students who received informative and motivating formative assessment when learning about floating and sinking scored higher mean conceptual understanding levels than students in a control group. Controlling for pre-test, students in the formative assessment group scored significantly higher in the post-test than students in the control group (beta = 0.24, SE = 0.12, p <0.05), with non-significant results for the other interventions. Formative assessment was particularly beneficial for students with poor language proficiency and teachers learned how to provide informative and motivating feedback to their students.

6.4.2.2 Impact of self-assessment on learning, self-regulation and self-efficacy (Panadero et al. 2012)

Panadero et al. (2012) compared the effects of two different self-assessment tools (rubrics and scripts) on self-regulation, learning and self-efficacy in 120 Years 9 and 10 students across two Spanish schools. Rubrics are self-assessment tools with three characteristics: a list of criteria for assessing the important goals of the task, a scale for grading the different levels of achievement and a description for each qualitative level. Scripts offer specific questions structured in steps to follow an expert model of approaching a task from beginning to end. They are designed to analyse the process being followed during a task, although they can also be used to analyse the final outcome. The researchers used a complex experimental design with a 2 x 3 x 2 structure, resulting in three between-group independent variables: (1) type of instructions (oriented to process or to performance), (2) presence or absence of self-assessment tool (control vs. rubric vs. script), and (3) feedback (oriented to process or to performance). Variables (2) and (3) relate to formative assessment and are therefore of most interest for this review. Each category had only ten participants, limiting confidence in the statistical results. The rubric categories and script ‘prompts’ focused on visible evidence in the landscapes, with the latter providing steps to be followed for successfully describing features in each category. Performance feedback abruptly stated what was missing, while process feedback explained why the missing feature was important and what it was. Use of the two self-assessment tools increased learning over the control group but the effects of either the scripts or rubrics could not be determined. The only significant effect on learning was the interaction between self-assessment tool and occasion. Process-oriented feedback increased self-efficacy more than performance-oriented feedback.

Years 9–10 students using scripts and/or rubrics for self-assessment were found to have learned more than a control group. Using scripts enhanced self-regulation more than rubrics. The only significant effect on learning was on the interaction between self-assessment tool and occasion (F (2, 108) = 7.85, p b 0.001; η2= 0.127). That is, both rubric and script conditions outperformed control on all three trials.

6.4.2.3 Impact of group discussion, feedback from peers and teachers (Vogelzang & Admiraal 2017)

Vogelzang and Admiraal (2017) conducted an action research project including 69 Year 9 students across two classes in a single Netherlands school. The study sought insights into formative assessments’ effects on achievement during a context-based Chemistry course on lactic acid and polymers. The formative assessment group participated in three types of activity: group discussion, feedback from peers and feedback from teachers. The control group answered posed questions in class as normal. The control group teacher did not draw special attention to the nature of the questions and the students were not explicitly asked to provide feedback to each other. In the formative assessment groups, students discussed questions and were asked to write down their answers individually. They provided feedback to each other and received feedback from the teacher (40 minutes). Feedback was provided to groups and, if necessary, to individual students; it focused on both students’ subject matter understanding (cognitive level) and their learning strategies (metacognitive level). Professional learning was provided for the teacher as researcher.

The intervention teaching condition experienced a more explicit and interactive form of learning with an emphasis on providing feedback to each other when compared to the regular condition. The conditions were subsequently switched, and the entire first round procedure repeated for the condensation polymers topic. The replication confirmed formative assessments had a significant effect on student achievement. However, the reported findings do not determine which formative assessment strategy was important, or if the specific combination the study used created the effect. Vogelzang and Admiraal (2017, p. 155) do, however, allude to 'intriguing discussions' emerging 'between students, between students and teacher and between teachers'.

Year 9 students who participated in a number of formative assessment activities (group discussion, feedback from peers and feedback from teachers) experienced a significant effect on their achievement levels in knowledge tests related to the topic. A two-way repeated measures ANOVA found, for the first run (lactic acid), the formative assessment group scored significantly higher than the control group on the post-test (F(1.56) = 36.93; $p < 0.001$; η2 = 0.397; Cohen's d = 1.62). For the second run (polymers), the formative assessment group also scored significantly higher than the control group (F(1.56) = 12.15; $p < 0.001$; η2 = 0.178; Cohen's d = 0.93).

6.4.2.4 Impact of grading/feedback approaches on student performance (Zhang & Misiak 2015)

Zhang and Misiak's (2015) study investigated the impact of three different grading methods on US Year 7 and 8 students' (n=136) Science achievement and motivation: (1) point based, (2) rubric based, and (3) rubric-plus-written-feedback based. The authors gave details of each rubric and rubric-plus-feedback condition but failed to provide details on the point-based condition. Therefore, we cannot know whether the point-based teacher used formative assessment strategies in their lessons, thus influencing reported findings. In the rubric-only condition, students were graded based on rubrics that communicated learning goals and criteria associated with a standards framework. The rubric also included a system for students to track their current level of development using a four-point scale: (1) basic, (2) approaches standard, (3) meets standard, and (4) exceeds standard. In the rubric-plus-written-feedback condition, students received additional written feedback from teachers along with a rubric rating. The written feedback included a suggestion or question that allowed students to continue their standard development. All students in the standard-based system were taught by the same teacher in an inquiry-based educational setting using the same instructional tasks and assessments. Control classes used a points-based grading system and were taught by a different teacher.

Zhang and Misiak (2015) found that student in both both Years 7 and 8 in the rubric-plus-written-feedback group performed significantly better on both achievement and motivation measures than those in the points-based and rubric-only groups. The rubric-only group performed no better than the points-based group, underscoring that written feedback may be vital for effective formative assessment results. Year 7 and 8 students who received teacher feedback through rubric-plus-written-feedback performed significantly better on both achievement and motivation measures than students who received scores-only or rubric-only feedback. Both Years 7 and 8 students in the rubric plus feedback group performed better than the rubric-only group ($p = 0.01$), and better than the points-only group ($p < 0.01$). The rubric-only group compared with the points-only group showed no significant difference ($p = 0.08$).

6.5 Part B: Formative assessment using online tools

This section details findings from high (n = 2) and low (n = 3) quality studies. It includes those in which students were provided with feedback through online tools during or outside class activities.

6.5.1 High-quality studies

6.5.1.1 Use of group-level feedback tools to support their meta-discourse (Resendes et al. 2015)

Resendes et al. (2015) conducted a two-year study with 42 Year 2 students in a Toronto laboratory school to explore students' ability to engage in productive discussion about knowledge building using group-level feedback tools to support their meta-discourse. The control class (C, n = 21) received normal Year 2 instruction while, the following year, the same teacher used treatment protocols to teach the incoming Year 2 class (E, n = 21). Both classes experienced wide-ranging teacher-facilitated student-driven practices, but class E interacted with (1) an online vocabulary assessment tool (presented in word clouds), and (2) an epistemic-discourse moves tool (presented in bar graphs reporting the frequency of use of knowledge forum scaffolds). There were two treatment groups in class E – those who only used word clouds and those who used both word clouds and frequency graphs. Study participants had previous experience with knowledge-building practices from Kindergarten and had used an online 'Knowledge Forum' from Year 1, where they posted ideas, questions and insights with teachers providing guidance as needed. Student groups interpreted word clouds and bar graph 'visualisations' about their knowledge-building capabilities and engaged in discussions (meta-discourse) with peers and their teacher when learning about birds and salmon. Findings resulted in significantly greater scientific accuracy and more elaborate explanations in students' subsequent online work. The study was unable to identify individual effects of tools and teacher–student discussion. The general conclusion was that students as young as age seven are capable of productive meta-discourse when supported with online tools and teacher feedback, although not independently.

Positive learner outcomes were further seen when a classroom teacher and researchers co-developed two formative assessment visualisation tools to support Year 2 students' choice of vocabulary and frequency-of-use knowledge forum scaffolds to promote metacognitive thinking. Students demonstrated significant improvements in their domain-specific vocabulary, repertoire of discourse moves, scientific understanding, epistemic complexity of ideas and interpersonal connectedness in online discourse. The teacher also provided verbal support as students developed their meta-discourse capabilities; that is, at all stages of learning, the teacher was directly involved with students' thinking and learning. Using tools while learning also meant students received feedback on their learning progress. The Epistemic Discourse Moves tool (Group B) was better than the Comparative Word Clouds tool (Group A), which was better than no tool (control group) on most lexical measures ($p < 0.05$; number of words written, number of domain words, number of unique domain words, percentage of domain words above grade level). However, there was no significant difference in the number and percentage of academic terms employed in the students' thousand most frequently used words. Post-hoc tests (Tukey's HSD) indicated that experimental groups A and B performed better than the control group on scientific-ness ($p < 0.01$, Cohen's d = 0.59) and epistemic complexity ($p< 0.05$, Cohen's d = 0.39), but there was no significant differences between Groups A and B for student notes on scientificness and epistemic complexity.

6.5.1.2 Multiple-choice web-based dynamic assessment system (Wang 2010)

Wang (2010) investigated how effective using a multiple-choice, web-based dynamic assessment system was compared with a normal web-based test on Year 6 Taiwanese students' (n = 116) understanding of plant photosynthesis. All students received two weeks e-learning instruction materials presented in a series of 'main points' via figures, tables and animated images, with 30 multiple-choice questions presented in one of two modes: (1) correct or incorrect response (scores-only), or (2) feedback or hints with opportunities to amend responses (graded-prompts). During the instruction period, each group could access e-learning materials via school computers, and thereafter whenever they chose. The scores-only group was given access to all the multiple-choice questions at once and could attempt to answer questions as many times as they wished. The graded-prompts group could attempt each question in turn up to three times and thereafter the question would be unavailable. Hints were posed in a sequence of increasing explicitness, beginning with general hints. For example, responding incorrectly to the question: 'When can plants do photosynthesis?' by selecting the answer, 'Only when there is no light can plants do photosynthesis' would generate a hint such as, 'Think about it. What do plants need when they do photosynthesis?' Student learning of photosynthesis outcomes in the graded-prompt group was significantly better than in the scores-only group, and the strategy was found to be particularly effective for students with low-level prior knowledge.

Year 6 students worked in school and at home with online materials developed by specialist high school Biology teachers. Students received feedback through graded-prompts or scores-only while learning about photosynthesis. The web-based dynamic assessment tool was found to be particularly effective for learning photosynthesis knowledge concepts by learners with low-level prior knowledge. The classroom teacher was not involved in the development or use of the tool, so this study supports primary students' learning rather than supporting teachers' professional judgement or teachers' scientific pedagogical practices. Data was analysed as follows: 2-way ANCOVA (prior knowledge x treatment, pre-test as covar): treatment main effect: (F1, 109 = 235.974, p < 0.01); prior knowledge: (F2, 109 = 8.020, p < 0.01); the interaction was also significant F(2,109 = 4.185, p < 0.05).

6.5.1.3 Capturing problem-solving discourse using online discussion logs (Soong et al. 2010)

Soong et al. (2010) researched the impact of different 'revision strategies' on Year 9 Singaporean Physics students to reveal and address misconceptions and misunderstandings. The small-scale study (n = 21) split participants into three groups (n = 7): control, experimental and after-school physics tuition. Students in the experimental group used online discussion logs with an anonymous peer feedback about posed problems following class instruction. Teachers analysed the problem-solving discourse for student misconceptions and then designed targeted prescriptive lessons to address any issues. While the authors found students in the experimental group significantly improved their understanding of the physics concepts covered, it was unclear what the 'control' and 'after-school Physics tuition' groups were actually doing, or what students were being tested on in their pre- and post-tests.

Teachers analysed the problem-solving discourse of Year 9 Physics students' online discussions (with an anonymous peer) to design targeted prescriptive lessons to address any issues. Students in the experimental group significantly improved their understanding of physics concepts. Teachers benefitted by improving their understanding of student conceptual learning progressions and designing appropriate learning opportunities to support student concept acquisition. Data was analysed as follows: T-tests on gain scores: control compared with tutoring – not significant, t = 0.672, p = 0.514; control compared with collaboration – significant, t = —3.20, p < 0.01; collaboration compared with tutoring – significant, t = —4.89, p < 0.01.

6.5.1.4 Online corrective and explanatory feedback (Terrazas-Arellanes et al. 2018)

Terrazas-Arellanes et al. (2018) conducted a large-scale US study over three years on 2,303 Years 6–8 students (typical versus disability versus English learners) with 71 teachers in 13 schools. The study aimed to compare the effects of Science teaching using online multimedia with 'normal' teaching. Online interactive resources included two out of five activities that could be categorised as 'formative assessment': corrective and explanatory feedback was provided to students about their scientific knowledge through interactive quizzes, and students had access to dialoguing opportunities and idea exchange via peer-to-peer interactions, small group discussion and purposeful forum activities. Unfortunately, the study design did not address the independent effects of these activities and it was unclear whether control teachers were using formative assessment in their 'normal' teaching.

The Years 6–8 students who interacted with online multimedia, which included formative assessment strategies, significantly deepened their Science knowledge. Students in the treatment condition improved scores an average of 16.7 percentage points compared to 5.7 percentage points in the control condition. Overall, the treatment had a significant effect (F(df = 1,29) = 16.8, p < 0.001); effect size d = 0.65). Interactions between disability status and condition, and English language status and condition were not significant, suggesting that both subgroups improved relative to controls (which was a main effect of learning disability, as they had lower scores than other students (F(df = 2,041) = 5.6, p = 0.018).

6.5.1.5 Customising lessons to students needs using SmartGraphs (Zucker, Kay & Staudt 2014)

Zucker, Kay & Staudt (2014) investigated the impact of a free web-based software named SmartGraphs specifically designed to help students overcome misconceptions regarding graphs. The large-scale US study was conducted over two years – the first year comprised 2,000 Years 8 and 9 students in 91 classes with 35 teachers while the second year comprised 1,700 Year 8 and 9 students with 29 teachers. All three authors work for the online software developer, which provides targeted hints in response to student answers. The scaffolding effectively creates lessons customised to student needs because only students who need help see scaffolding on any given page. Scaffolding can be in the form of written hints, equations or visual markers on a graph or table, none of which were detailed in the paper. Using SmartGraphs activities that focus on the motion of objects as a supplement to normal instructional activities in physical Science classes resulted in statistically significant learning gains for students. A significant issue with this study is the absence of details about key formative assessment prompts. However, it is possible to review prompt types directly online (https://smartgraphs-activities.concord.org/activities/225-african-lions-modeling-populations/student_preview/), where they appear to follow a 'learning progression' sequence (i.e. participants are 'stepped' through a series of increasingly complex questions and opportunities to interact with hypothetical graphical data related to a scenario).

Years 8 and 9 students using SmartGraphs software activities that provide targeted hints in response to student answers achieved significant learning gains. The total gain in scores was significantly higher in the experimental groups (t = –2.669, df = 1684, p = 0.008).

Using a scaffolding tool to help interpret graphs was not beneficial to the students.

7 Chapter 7: Review results – Research on effective formative assessment practices in The Arts

7.1 Summary

The evidence for the impact of formative assessment on student learning in The Arts shows mixed results. Students showed more positive ratings on motivational factors (e.g. self-efficacy, perceived usefulness of feedback, interest) when engaged in formative assessment compared to those in control groups. Future research should place greater emphasis on the investigation of cognitive and motivational processes, explaining how formative assessment impacts learning.

7.1.1 The evidence for the impact of formative assessment on student learning in this domain

- There are few rigorous experimental design studies of the impact of formative assessment on student learning in The Arts.
- Studies note improvement in achievement across Arts areas when formative assessment was present as regular, specific, individualised, on-the-spot feedback.
- Teachers observed improvements in engagement and motivation, increased self-evaluation, self-regulation and the ability to provide valuable peer feedback.

7.1.2 Evidence for the effectiveness of particular tools and resources (including online tools)

- MyeDance was the only online tool cited. Students found it particularly useful when the feedback provided was both a rating and a comment – both based on a rubric provided by an expert and teacher.

7.1.3 Implications for effective implementation of formative assessment practices

- The teacher plays a significant role in providing accurate, valuable and timely feedback on performance/production tasks.
- Teachers need well-developed pedagogical content knowledge (PCK) to analyse assessment data, deliver appropriate feedback and implement evidence-based interventions.
- Teachers need ongoing professional development to learn how best to target the students' individual needs and subsequently improve student achievement/outcomes in Arts learning.
- Regular feedback and criteria-based rubrics are key to improving student learning outcomes.
- Assessing creativity remains problematic. A variety of assessment tools and processes are needed, including self-, peer and teacher feedback, and constant collaboration between students and teachers to maximise Arts classroom learning.

7.2 Chapter overview

Formative assessment is very familiar to Arts teachers, particularly in relation to performance and production tasks where ongoing feedback forms the basis of the learning process. Research conducted in this area primarily investigates the use and quality of feedback, and the use of rubrics to guide feedback. The tension inherent in assessing creative expression remains a major issue in assessment task design and measuring student outcomes. Teachers continue to resist pressure to assign a numerical value to a work of art, a dance, a theatrical piece or a musical composition, due to the subjective and aesthetic variance of both producer and viewer.

The initial search for literature on formative assessment in The Arts yielded a small number of research papers and we found very few research studies used an experimental design. Of the five studies located, only two aligned with the search criteria, although these reported on the same study: *Arts Achieve* (Chen et al. 2017; Chen & Andrade 2018).

Arts Achieve was a large-scale study conducted in New York City, designed and evaluated by experts from Metis Associates. It aimed to obtain empirical evidence that formative assessment could have a positive impact on student learning in The Arts when implemented with purpose and care. The study involved 75 'high needs' schools selected from across New York City, and used a stratified randomised design. Each art form (dance, music, theatre/drama, visual arts) and education level (elementary, middle, high school) were used in the stratification. Participating schools were randomly assigned to either control or treatment conditions. Teachers in the treatment condition received training to use criteria-referenced formative assessment. Only those students whose teachers demonstrated high fidelity in implementing the treatment were included in the data.

The assessment tasks created specifically for this study included multiple-choice questions, short responses, essays and cloze passages. They were designed to 'authentically measure students' conceptual understanding, literacy, application of knowledge, and analytical and performance skills relevant to each art form' (Chen et al. 2017, p. 302). Tasks were examined and revised several times to ensure internal consistency and reliability before pre- and post-tests were administered to all students. Some assessment items for elementary and middle school visual arts and high school dance were found to have low reliability because tasks measured several different dimensions within the one assessment. The authors note that 'internal consistency assumes unidimensionality' and yet 'multidimensional' assessments target 'authentic variety of knowledge and skills' (Chen et al. 2017, p. 303). Additionally, no student data from high school music and drama (theatre arts) could be used due to low fidelity from all participating teachers in these two specific art forms.

Chen et al. (2017) report on the overall research project and whether or not there was evidence that formative assessment had a direct or causal impact on student learning (performance) in The Arts. They state that, despite a range of measures put into place, the research had a number of limitations, particularly reliability and consistency across art forms and educational levels. While they offered three examples of the type of tasks used (e.g. multiple choice, short response), they offered very little information about how these were implemented as formative assessment. There was also scant information about how tasks were marked or graded, with the exception that they were derived from criteria presented in relevant curriculum documents, specifically the New York City Department of Education *Blueprints for Teaching and Learning in the Arts* and Common Core State Standards in English Language Arts. This study seems to involve a large number of facets (including demographic considerations and connections across subject areas), that make it difficult to claim the formative assessment implemented was responsible for the small increase in students' performance. Nevertheless, the authors present evidence that formative assessment 'has a small, positive effect' (Chen et al. 2017, p. 310) when students are given clear criteria for success and when self- and peer feedback is used.

Using a dataset from the same *Arts Achieve* project, Chen and Andrade (2018) further examined fifth-grade student achievement in theatre arts (drama), with an aim to 'improve student achievement in the Arts by providing support to Arts teachers on the use of balanced assessment and the integration of technology in instruction and assessment'

(Chen & Andrade 2018, p. 312). However, little or no information is provided about the actual tasks administered, the training teachers undertook in the treatment condition, or the technology used. The treatment condition did have a positive impact on student achievement, although it was small and only related to performance tasks. Again, data from treatment groups was only included if deemed high fidelity (that is, teachers had satisfactorily applied 'task-specific criteria, peer and self-assessment, and opportunities to revise in order to improve work and deepen learning,' Chen & Andrade 2018, p. 317). This is consistent with a typical approach to theatre performance work, where students know what they need to do, receive feedback about their performance from the teacher and peers, reflect on their own performance (self-assessment) and are given time to revise their work for improvement.

It is worth noting that the multiple-choice, short-answer and other written tasks administered yielded no evidence that the provision of criteria-referenced formative assessment had any positive effect on student achievement in these areas. The authors also highlight that the project applied several interventions, which confused results and made it problematic to claim a causal effect between formative assessment and student achievement (Chen & Andrade 2018).

Due to the scarcity of experimental studies, we adjusted the search criteria to capture qualitative studies about how formative assessment has been used in Arts education settings, employing more-specific terms such as dance, drama, music and visual arts. Two main themes emerged from this adjusted process: (1) feedback is an essential feature of formative assessment in each of the art forms and some forms of feedback are more effective than others, and (2) as a common tool to assess student work, the rubric has a powerful impact on student progress when used effectively. Only one study used an online tool.

Hsia, Huang and Hwang (2016) present their findings from a quasi-experimental study undertaken in Taiwan with 100 college dance students (average age = 19.98 years). The research focus was the impact of different forms and combinations of feedback, rather than the formative assessment task completed by students. This process was mediated through the use of *MyeDance,* an online program designed for giving feedback based on the *iRubric–Dance performance evaluation rubric,* designed by Campus (2014). The pre-test involved students learning the group dance code, rehearsing, video recording the dance and uploading it to *MyeDance.* Over a subsequent two-week period, students used the program to provide peer feedback. Students were divided between three experimental groups and delivered feedback on all dance video recordings according to their assigned learning mode: Group 1 (n = 29) gave feedback in the form of peer comments; Group 2 (n = 32) provided a peer rating; and Group 3 (n = 39) provided both a peer comment and a peer rating. As part of the post-test, after feedback was received students were given one week to rehearse and revise their group dance performance. This stage also involved the completion of 'Likert-type scale' questionnaires (Hsia, Huang & Hwang 2016, p. 62) related to self-efficacy and motivation, and interviews to collect qualitative information. Content analysis was applied to peer feedback based on an existing 'behavior and a skills analysis tool called the Teacher feedback observation system' (Hsia, Huang & Hwang 2016, p. 62).

The study considered a significant number of elements, including the value of peer feedback, its impact on the learner, self-efficacy and motivation. It also considered the quality of feedback, whether good or poor, and the impact this has on the receiver. It consequently questioned students' ability to provide effective feedback. Overall, findings demonstrated that feedback in the form of peer ratings (provided by experimental Group 2) improved group performance skills, whereas mixed feedback (peer rating plus peer comment) improved individual performance skills. The study saw a positive correlation between motivation, self-efficacy and pre-test dance performance skill, suggesting that feedback motivated and empowered students to improve their performance and achievement in dance.

Peer-feedback was also used in a US study involving six jazz dance students aged 12–17 years (Quinn et al. 2016). This study used a non-concurrent multiple-baseline across-participants design and targeted the use of auditory peer feedback. All students arrived at the dance studio 30 minutes prior to start and stretched for five to ten minutes, then performed specified dance skills four times each (Quinn et al. 2016, p. 372). Dancers were grouped in pairs for the intervention and all participants were given 30 minutes training, which described the intervention and practised using a ‘clicker’ that indicated correct performance. Each skill had to be performed correctly at least three times, or a demonstration was provided. The ‘click’ acted as auditory feedback and no comments accompanied it. The student providing feedback was trained to click only when a skill was performed correctly and not to make any other comments about incorrect or inaccurate elements. Results demonstrated ‘auditory feedback increased the percentage of correct dance movements for each individual receiving the feedback’ (Quinn et al. 2016, p. 376). Peer feedback was particularly effective. Students were also asked to complete a Likert scale questionnaire to ascertain the intervention’s social validity. Overall, responses indicated a positive attitude to the intervention and an improvement in the dance skills of those receiving feedback.

As stated previously, the application of feedback is a common theme that emerges in literature about formative assessment in The Arts. Many studies comment on the efficacy of feedback, particularly in relation to performance or production tasks. Some studies discuss the quality of feedback as well as how it is received by students and the impact this has on student learning progress.

Denis (2018) analyses literature on formative assessment in music across educational levels – from elementary to high school. He suggests music teachers traditionally engage in formative assessment because they constantly ‘listen critically to student performance/practice, make judgements and provide feedback’ (Denis 2018, p. 21). This type of feedback is individual, specific and ‘conducted during learning’ (Denis 2018, p. 21). The study discusses definitions of assessment, why it matters and how it is enacted, grappling with the ongoing conflict associated with assessing creativity. The subjective and democratic nature of feedback in The Arts means that assessments are in danger of being based on non-music criteria, such as participation, behaviour and observation, and the study argues this undermines the credibility of music as a legitimate area of study. It also describes the various assessments – rubric, checklist, portfolio and rating scale – generally applied to performance tasks while noting music teachers do frequently use written assessments (quizzes, exams and worksheets) to assess theoretical knowledge (vocabulary, music theory, notation, music reading and literacy skills), which is very different and more straightforward than assessing performance and creativity. Denis outlines various challenges impeding music assessment procedures, including disagreement between teachers about ‘music curricula and end goals of instruction’ (Denis 2018, p. 24), impediments related to time, support and class size, and the ‘lack of transparency in grading’ practices (Denis 2018, p. 25). The study further highlights the few standardised measures used in research, such as the Torrance Tests of Creative Thinking (TTCT) and Amabile's (1983) Consensual Assessment Technique (CAT) but acknowledges concerns surrounding reliability. Its general conclusion is that music assessment relies on the practitioner planning meaningful tasks aligned with learning objectives but cautions that care must be taken to ensure what is considered formative assessment is not actually frequent or constant summative assessment.

Ferm Almqvist et al. (2017) discuss similar considerations, reiterating the difficulty music teachers have when assessing creativity and the problem of creating authentic yet measurable criteria-based assessments that potentially disregard the significance of individual learning needs and progression in-situ. Two examples are presented: one each from Sweden and Norway. Each teacher provided extensive written feedback following the first stage of a music task and students used this feedback to complete the task’s remaining elements.

Ferm Almqvist et al. warn of the danger of using assessment as the ‘central didactical tool for the enhancement of students’ academic achievement’ (Ferm Almqvist et al. 2017, p. 6) as they reflect on the work of Torrance (2007) in the UK and apply his ‘theory of criteria compliance’ (Ferm Almqvist et al. 2017, p. 4) to the Scandinavian context. They discuss quality in terms of both student outcomes and the teacher’s work, as well as the tension between autonomy and accountability, arguing ‘the assessment of artistic work can hardly follow standard measures’ (Ferm Almqvist et al. 2017, p. 12). They suggest conformist, standardised approaches to assessment weaken music learning because they fail to do ‘sufficient justice to the subject matter of music’ (Ferm Almqvist et al. 2017, p. 12) by undermining the diversity inherent in subjectivity. The difficulty of assessing creativity has meant that creative tasks are reduced to a checklist of skills and knowledge, thereby reducing the legitimacy of music as a subject. The pressure on teachers to demonstrate evidence of student achievement has resulted in constantly assessing students, which has misrepresented assessment *for* learning and turned it into assessment *as* learning. Teachers run the risk of tailoring tasks to comply with requisite criteria rather than addressing individual student needs. This approach disregards the ‘uniqueness of learning experience’ and the ‘qualities that surface in the *situations* of learning’ (Ferm Almqvist et al. 2017, p. 12), that impacts teacher planning quality. The authors maintain that assessment *for* learning has already become assessment *as* learning, where all activity becomes assessment and disconnected from the learner.

In Lithuania, Kazragytė and Kudinovienė (2018) present another perspective on feedback as formative assessment. Consistent with other research, they maintain assessment in The Arts is difficult and research on the topic is rare. They also propose formative assessment is simply information a teacher must use to adjust their lessons to better facilitate student learning – a natural part of the teaching/learning exchange. This view is supported by Goh and Walker (2018), who describe formative assessment as 'bridging the gap' between where a student is at and where they need to be. Kazragytė and Kudinovienė’s (2018) study is based on a series of observations of Arts lessons (music, dance, fine arts, theatre) where protocol – formative assessment – was present. No information about tasks was provided; however, the teacher gave oral feedback at the time of the art-making tasks. The authors therefore emphasise the relationship between formative assessment and a ‘teacher’s metacognitive abilities, needed to envisage pupils’ achievements and plan them’ (Kazragytė & Kudinovienė 2018, p. 218).

Like many authors, Kazragytė and Kudinovienė advocate the significance of the teacher in planning and administering effective formative assessment aligned with student capacities and needs. Like Ferm Alqvist et al. (2017), Kazragytė and Kudinovienė (2018) note the difficulty of establishing ‘concrete assessment criteria’ for creative tasks and how this can result in a ‘lack of validity in the assessment’ (p. 221). When present, formative assessment is correlated with a positive class atmosphere, positive teacher–student relationships and positive student learning. When formative assessment was reported as ineffective, it was found the teacher had not included it as part of the lesson protocol. Formative assessment was particularly effective when used throughout a lesson: as students are told about what was expected of them in the task; during or after the performance of the task as they received teacher feedback; and afterward, when students learned what they had achieved. Once again, the efficacy of formative assessment remains largely in the hands of the teacher. Applying formative assessment in the ways presented here is further supported by similar research, such as Schmid (2012), Schuler (2011), Staley (2015) and Mills (2009). These authors reiterate that formative assessment is a daily practice in any dance, drama, music and visual arts classroom. It takes various forms but is always based on immediate feedback. Mills (2009) discusses the use of portfolios, Schuler (2011) confirms the nature of assessment and providing students with information for success from the beginning, and Schmid (2012) examines using data and evidence in dance education to best target student needs and create quality assessment tasks that ‘yield accurate information about student mastery of the intended content or skill’ (Schmid 2012, p. 78).

A study by Lin (2013) also supports this view. Conducted in Taiwan with primary school students (Years 1, 3 and 5), it focuses on how to create drama assessments that fulfil the formative assessment traditions of Arts education and the summative requirements of curriculum and reporting, essentially balancing assessment *for* and assessment *of* learning. Participant teachers devised assessment tasks and accompanying rubrics in collaboration with the researchers over an extended period of time. Tasks and rubrics were revised numerous times and tested for consistency. Activity-based, discussion-based and writing-based tasks were used at the teacher's discretion. While the study did not discuss formative assessment's impact on student learning, it did consider the issues surrounding development of tasks and rubrics related to teachers involved. In particular, it found teachers' experience, preferences and interpretations of the rubric all played a significant role in assessment construction and revision. Teacher development, therefore, becomes an essential ingredient in building effective formative assessment.

The Finnish education system has long been regarded as one of the world's best, supported by consistently high PISA rankings. According to Juntunen (2017), this high standard results from teacher autonomy and equity of education, underpinned by high-quality teacher education that encourages professional freedom. The study involved 4,792 high-school music students (aged 15–16 years) and their teachers. It acknowledges music teachers have more autonomy than other teachers because the music curriculum is very open. The study had three task stages: (1) a written test, (2) an advanced written test, and (3) production tasks (singing, playing, composing). Assessment tasks were analysed for validity and reliability and Cronbach's alpha was used to establish that tasks had internal uniformity (Juntunen 2017, p. 7). The teachers created tasks in line with national requirements, but it is noted these were summative assessments and no formative assessment took place. Students did not receive feedback or the chance to revise or change their work. Instead, the study focused on teacher autonomy and teachers' ability to plan assessment tasks according to curriculum criteria and national education requirements. Like Ferm Alqvist et al. (2017) and Kazragytė and Kudinovienė (2018), Juntunen questions the appropriateness of standardised assessment in music, stating 'assessment in music is complex and even disputable' (Juntunen 2017, p. 11) and that the assessment tools available do not 'measure' what are considered integral aspects of music learning.

Mastrorilli, Harnett and Zhu discuss the professional development of teachers in the *Arts Achieve* project mentioned earlier. They propose that, when teachers are actively involved in professional development, 'their art instruction will improve' (Mastrorilli, Harnett & Zhu 2014, p. 2). They advocate professional learning communities in which teachers can collaborate and support each other to improve practice and student learning. The study found that, after a year of Arts learning, students' Arts achievement had improved, but there was little evidence to show this was a result of teacher professional development in formative assessment. In fact, teacher growth was not evident from the measures used by evaluators.

Other research also highlights teacher significance. Scott argues that music teachers have a responsibility to provide 'valid and reliable information' (Scott 2012, p. 31) in relation to their students' performance. She discusses assessment *of* learning, assessment *for* learning and assessment *as* learning, differentiating these in terms of the nature of a task, including whether a task (1) is done to, for or by the student, (2) is active or passive, (3) is teacher or student-centred, (4) is competitive, collaborative or personal, and (5) is controlled by administrative bodies or helps students learn (Scott 2012). She outlines the variety of tools designed to help teachers provide valuable feedback and guidance to improve student understanding, performance, and achievement, such as checklists, rating scales, rubrics, portfolios, narrative reports, rehearsal logs and questioning. A Swedish study by Andersson used a phenomenological approach to analyse the reflections of secondary school dance teachers on their assessment practices in an attempt to establish a generalised 'essence' of dance assessment for learning. Teachers described the various forms of feedback used, including verbal and written correction and encouragement, body contact and documentation. All feedback was individualised and very specific to what a student required in the moment. Students were expected to change their dance technique in response to teacher feedback while the teacher was expected to change their practice to support student needs and 'different capacities' (Andersson 2018, p. 284). Teacher feedback and student self-assessment used a rubric based on dance syllabus criteria.

According to Bartlett and McPhail, ‘feedback is regarded as one of the most significant pedagogical acts a teacher undertakes’ (Bartlett & McPhail 2016, p. 63). However, feedback can reflect a lack of teacher expertise in certain study areas. The New Zealand study observed how high school music students responded to formative teacher feedback about music composition, part of which involved understanding how students used feedback to change their compositions. Two teachers were observed giving students feedback regularly across a five-week period. Feedback was given verbally to individual students while looking at their composition; it was both detailed and descriptive. Discussions between teacher and student occurred when teacher expertise did not extend to a stylistic choice made by the student. After students explained their choice(s), ‘ideas for enhancement of the piece were discussed’ (Bartlett & McPhail 2016, p. 71). Where teacher expertise in a specific musical genre was lacking, they modified their feedback to focus on different elements of the piece. The study found students accepted teacher feedback and applied it to their compositions, improving the quality of the final work considerably.

By contrast, Goh and Walker (2018) discuss effective feedback in terms of how teachers give feedback and how students receive and respond to feedback. This Australian study found that participating Year 7 students responded negatively to teacher feedback, potentially because they perceived task feedback as personal criticism. The authors argue context is always important and generates self-regulation, a key consideration for goal setting and for achievement. Similarly, Sicherl Kafol, Kordeš and Holcar Brunauer (2017) discuss a study set in Slovenia involving Year 8 music students and a composition task. The authors concluded that assessment is an ongoing process based on feedback, which also helps students overcome negative perceptions of assessment.

Another theme to emerge from the literature is using rubrics in formative assessment. In most cases, rubrics are used to drive feedback. Andrade, Hefferen and Palma discuss a study in South Brooklyn involving students from Years 3–8 (elementary and middle school) as part of the *Artful Learning Communities* project. They focus on the co-creation of a rubric for assessing a visual arts exercise. Initially, specialist art teachers resisted the rubric, claiming ‘art cannot be assessed’ (Andrade, Hefferen & Palma 2014a, p. 34); however, the researchers convinced them of the assessment’s value by suggesting that ‘authentic artistic processes, such as setting goals, assessing one’s own work and revising’, were all part of the assessment process and a natural part of ‘any creative endeavor that involves rehearsal and redoing’ (p. 35). The study reports two examples where teachers and students developed rubric assessment guidelines to self- and peer assess and engage in discussions about their work to reflect and modify. Participants could choose whether to change their product based on feedback provided. The study suggests students were successful in applying the rubric to peer feedback and were able to self-assess using the rubric’s criteria. The researchers attributed several results to this work, including students becoming their own teachers and demonstrating the ability to self-monitor, self-assess and self-teach. In addition, teachers saw increased engagement and quality of student work. The authors discuss this same study in a chapter about formative assessment in The Arts (Andrade, Hefferen & Palma 2014b), grappling with the tension between ‘preserving the core features’ (p.126) of each art form, allowing for the creative decisions of developing artists, and systemic requirements for measuring and reporting ‘success criteria’ (p. 140).

As previously mentioned, Hsia, Huang and Hwang present findings from an online feedback system, *MyeDance,* for the assessment of group and individual dance skills achievement, mainly using Campus’ (2014) evaluation rubric, which is based on four specific categories: ‘choreography, technical skills, performance skills and rhythm’ (Hsia, Huang & Hwang 2016, p. 61). In this example, the teachers added two more rubric items relating to group performance: ‘teamwork and originality/creativity’ (p. 61). Researchers found the rubric helpful for students when constructing and understanding peer feedback and interpreting a peer rating. Participants used the rubric and feedback information to adjust their performances. The authors conclude rubric use prompted teachers and students to consider task criteria more purposefully, which led to a greater understanding of the process requirements and expectations.

Based on data gathered from the same *Artful Learning Communities* project discussed earlier, Andrade et al. (2015) highlight the use of co-created rubrics in Brooklyn middle school dance classes. Once again, the rubric was used to provide explicit criteria to guide feedback, and students successfully employed it for self- and peer assessment. The study used a range of feedback combinations, including self-, group and peer assessment feedback, reflection and revision, all guided by the criteria-referenced rubric. Teachers saw an improvement in students' knowledge, skills and ability to make meaningful revisions to their performances. They also noted an improvement in students' collaborative skills. Researchers observed a change in teacher practice to one more 'student-centred, formative and focused on frequent, goal-oriented feedback to students' (Hsia, Huang & Hwang 2016, p. 49).

The reviewed research refers to performance or production tasks – activities forming the basis of Arts education across dance, drama, music and visual arts, and which have traditionally demonstrated a formative approach to assessment and learning. In this context, skills and knowledge are 'assessed' incidentally as they are applied to a performance or artwork. Despite the ongoing tension of assessing creativity, when rubrics and feedback are used regularly as part of the learning process research demonstrates an improvement in student Arts achievement. While feedback forms vary (encompassing self-, peer and teacher feedback, written and verbal feedback), it always accompanies opportunities for students to revise their work and, at times, for teachers to adjust their practice to better facilitate learning.

Additionally, research shows professional development for teachers also improves student learning and achievement in the Arts.

8 Chapter 8: The optimal conditions for the effective implementation of formative assessment practices

8.1 Summary

This chapter considers the optimal conditions for the effective implementation of formative assessment practices. Environmental conditions and teacher level each play a key role.

8.1.1 Educational system structures and supports

The research highlights the need for:

- school leaders who understand formative assessment, can provide a rationale for its use and create a supportive and non-threatening environment modelling effective use of assessment data for teachers
- high-quality professional development and effective support for formative assessment implementation (see below)
- a commitment to providing teachers with regular and protected meeting time for meaningful examination of assessment practices
- leaders who can establish a schoolwide formative assessment culture with vision and expectations for assessments, and a school/classroom climate promoting trust, mutual respect and cooperation
- teachers to promote a classroom philosophy that regards mistakes as opportunities to learn and encourages honest reflection
- school leaders to strategically align expertise and resources to support teachers' learning about effective formative assessment practice
- the development of communities focused on improving formative assessment practice
- decentralised organisational structures and distributed leadership in schools, so accountability pressures on teachers do not lead to unintended impacts on instructional and assessment practices, helping to build a broader base of engagement, expertise and a greater sense of shared vision and ownership.
- increased focus on assessment literacy in initial teacher education and in-service teacher PD at a tertiary education level.

8.1.2 Teacher level knowledge, skills and beliefs

- Deep, flexible pedagogical content knowledge (PCK) is required so teachers can break down critical concepts, find appropriate entry points for all students, and redesign instruction to match students' understandings and misconceptions as evidenced in formative assessment.
- Teachers require both assessment knowledge and data literacy to effectively implement formative assessment.
- Teachers' pedagogic preferences and orientations influence their use of computer-based formative assessment.
- Teachers' attitudes and beliefs about teaching and learning has direct impact on implementation quality. Teachers with a focus on the learning outcomes of individual children are more likely to implement formative assessments compared to teachers focused on covering the curriculum.

8.1.3 Optimal use of technology

When considering the optimal technology use for supporting formative assessment, studies show a mixed picture regarding the effectiveness of formative assessment interventions in different disciplines. The evidence for technology use depends on various factors, such as specific groups of children (low achieving, high achieving), study sample size, experimental design and tool selection. The same general principles for effective formative assessment in general also apply to online assessment. Assessments need to be based on a valid model of task components, and the prerequisite cognitive and learning skills underlying successful progression. Interventions need to be evidence-based and aligned with validated learning progressions for the targeted concept or skill.

- Regarding the type of computer-based feedback provided to students, elaborated feedback with prompts is generally more effective than feedback that only recognises errors or provides correct answers.
- The following tools were found to be potentially useful:
  a. The Arts – MyeDance
  b. Mathematics – computerised dynamic adaptive test system, diagnosis and formative assessment-based personalised web learning system, classroom connectivity technology (CCT), problem-solving assessment, diagnosis and remedial instruction (PSADRI) system
  c. Reading – ISI/A2i
  d. Writing – Using Sources Tool
  e. Science – web-based dynamic assessment system and visualisation tools (e.g. word clouds, bar graphs).
- It is vital that teachers have the requisite knowledge and skills to use formative assessment hardware and software. In addition to the digital technology provided, teachers need professional development to administer assessments, interpret results and translate information obtained into effective teaching instructions.

8.2 Features of effective formative assessment professional development

Few studies examined how the characteristics of professional development programs on formative assessment affect teacher practices and student achievement. Rigorous evidence linking professional development to teacher and student outcomes is lacking (Schneider & Randel 2010). Research on high-quality professional development points to general attributes that improve teachers’ learning experiences:

- Professional development should be grounded in specific subject matter and increase teachers' content understanding and the diverse ways to teach it.
- Teachers’ attitudes to formative assessment can create barriers for implementation, especially when ideas and practices are incompatible with participating teachers.
- Brief interventions, such as short-term product-oriented workshops, are less likely to change practice effectively.
- Long-term process-oriented professional development with ample opportunities for collaboration, feedback and discussion appears to be more effective for successfully changing teachers’ classroom assessment practices.
- Strong school leadership is also associated with higher fidelity formative assessment professional development implementation by teachers.
- Professional learning that is sustained, collaborative work-embedded and situated within school needs is preferred over one-day workshops or formally presented interventions.
- Professional development is most effective when teachers engage actively in instructional inquiry in the context of collaborative professional communities that are focused on instructional improvement and student achievement.

- Continuous support is necessary for sustained application of evidence-based practice. Teachers need follow-up and support for new ideas and strategies to be effectively implemented.
- Collective participation by multiple teachers and/or students sharing their development work together at individual school sites or across networks is an important factor in the FAPD success.

8.2.1 Teacher knowledge, skills and attitudes

- Deep and flexible pedagogical content knowledge (PCK) is important for teachers to break down critical concepts, find appropriate entry points for all students, and redesign instruction to match students' understandings and misconceptions, as evidenced in their formative assessment.
- Teachers' assessment knowledge is insufficient for successful formative assessment implementation. Teachers require an understanding of assessment theory and research, and of how to translate these into concrete classroom practices.
- Using computer-based formative assessment is dependent on teachers' pedagogic preferences and orientations.

8.3 Chapter overview

Black and Wiliam's (1998b) seminal review suggests that, when teachers effectively utilise formative assessment strategies, student learning can increase significantly. However, the authors also acknowledge 'poverty of practice' among the teachers they observed, in that few fully understood how to implement classroom formative assessment. This assertion has since been repeated multiple times (e.g. Lysaght & O'Leary 2017). Despite the two decades since the initial observation, we still lack a strong research base on possible causes and how to best support teachers in implementing effective formative assessment practices. It has been suggested initial teacher education courses include too little focus on assessment in general, and the resulting gaps in assessment knowledge then impact teachers' ability to implement formative assessments (e.g. Impara, Plake & Fager 1993; Lingam & Lingam 2016). However, others advise formative assessment is difficult to understand and implement even for very experienced teachers (e.g. Luttenegger 2009), shifting the focus to insufficient continued professional development on assessment (Stiggins 1991). In survey after survey, teachers identify their assessment knowledge as insufficient (e.g. Noll 1955; Impara, Plake & Fager 1993; Wayman, Cho & Johnston 2007). Approaches to teaching and learning (e.g. Rakoczy et al. 2008), a school culture focused on high-stakes summative exams (e.g. Bramwell-Lalor & Rainford 2015), curriculum coverage (e.g. Box, Skoog & Dabbs 2015) and lack of teacher collaboration (e.g. Weinbaum 2009) are also mentioned as potentially preventing effective formative assessment implementation. By contrast, a constructivist approach, assessment knowledge, pedagogical content knowledge and supportive administration appear to facilitate implementation of high-quality formative assessments (Brink & Bartz 2017; Goertz, Olah & Riggan 2009; Sabel, Forbes & Flynn 2016).

This section of the review explores two sets of optimal conditions for the effective implementation of formative assessment practices. Part A looks at environmental conditions – including education system structures and supports – and school-level organisational and cultural factors. Part B considers teacher-level factors – including teacher knowledge, skills, attitudes and beliefs.

8.4 Part A: Environmental conditions

8.4.1 Education system structures and supports

Very few studies consider the larger context within which teachers operate and its impact on formative assessment practices. As a general statement, all formative assessments must align with curricula that outline the expected learning to be assessed. The Australian Curriculum and the state-level syllabi integrating it emphasise formative assessment and feedback (van der Kleij, Cumming & Looney 2018) and thus advocate formative assessments. As such, current education policies at national and state levels widely support effective implementation of formative assessment practices. At the same time, the visibility of summative assessment practices (such as NAPLAN), may undermine the role of formative assessment unless high-quality professional development and effective supports for implementation are in place, which may not currently be the case (van der Kleij, Cumming & Looney 2018).

Finally, Hopfenbeck, Florez Petour & Tolo (2015) suggest formative assessment implementation in schools was more successful when there was trust between the school district, school leaders and teachers. If schools and teachers view a formative assessment program as a top-down control of what is happening in schools, implementation is less likely to be less successful. According to Egan, Cobb and Anastasia (2009), a key element for their project's success was the district's commitment to providing teachers with regular meeting time and protecting that time for meaningful examination of assessment practices (see also Goertz, Oláh & Riggan 2009; Means et al. 2009). Similar practices by school leaders are also identified as facilitating formative assessments. These are discussed in further detail below.

8.4.2 School-level factors: Leadership and school culture

Various school-level organisational and cultural factors are identified as barriers or facilitators of formative assessment implementation. In this section, we discuss the leadership and school culture that may influence teachers' use of formative assessment and other data for instructional improvement. We consider professional development in the following section.

Studies focusing on school leaders generally acknowledge that leaders are the prime movers and change managers in creating new school cultures around assessment practices: they set the vision, outline the change process and create a supportive non-threatening environment (e.g. Deneen et al. 2019). For example, teachers interviewed by Brink and Bartz (2017) identified that administration support for formative assessment was essential for creating a cultural shift from summative to formative assessments. The leadership's role includes modelling appropriate uses of assessment data for teachers, providing a rationale (or theory of action) for using formative assessments, strategically aligning expertise and resources to support teachers' learning about how to use assessment results, and deploying resources to cover a range of data-related functions (Young 2006). Therefore, the first requirement for school leaders is that they understand formative assessment themselves (Love & Crowell 2018; Moss, Brookhart & Long 2013).

Several aspects of school culture are noted as important because classroom cultures tend to mirror organisational culture (e.g. Birenbaum et al. 2009). School leadership should establish a schoolwide formative assessment culture with vision and expectations for assessments (Havnes et al. 2012; Sach 2013). Further, leaders need to facilitate formative assessment use by scheduling time for professional development, collaboration among teachers and preparing assessments, feedback and instructional responses (Lee, Feldman & Beatty 2012; Ní Chroinín & Cosgrave 2013). Several studies suggest that collaboration is an essential focal point for a school culture that encourages formative assessment. Teachers working collaboratively and engaging in communities of practice tend to have better long-term success

(Birenbaum, Kimron & Shilton 2011; Kay & Knaack 2009; Lee 2011). A school climate of trust, mutual respect and cooperation is associated with higher quality formative assessment practices compared to a climate of mistrust, stress and competition between teachers. In a study of implementing assessment for learning in high schools, Weinbaum (2009) finds pre-existing school culture difficult to change, concluding the process of building a trusting team in which teachers feel safe discussing their assessment and instructional practices is as challenging as improving their understanding of assessment for student learning.

Classroom atmosphere is also important, with positive and trusting classroom relations between teacher, students and peers associated with better assessment practices. A classroom philosophy that regards mistakes as opportunities to learn and encourages honest reflection allows students to perceive critical feedback as constructive instead of judgemental (Harris & Brown 2013; Havnes et al. 2012; Newby & Winterbottom 2011; Rakoczy et al. 2008).

Finally, it may also be that teachers working in schools with a more decentralised organisational structure engage in better formative assessment practices than teachers whose schools are more centralised (Heitink et al. 2016). This may reflect the accountability pressures teachers feel in centralised organisations and their impact on instructional and assessment practices (Aschbacher & Alonzo 2006; Birenbaum, Kimron & Shilton 2011; Sach 2013). If teachers feel pressure to cover all curriculum content and prepare students for high-stakes exams, they may not see formative assessments as a viable instructional choice (Box, Skoog & Dabbs 2015). An alternative explanation is that distributed leadership can help build a broader base of engagement and expertise, and a greater sense of shared vision and ownership (Copland 2002; Young & Kim 2010). In general, the accountability and high-stakes examination culture of many school systems may be counterproductive for formative assessments (Berry 2011; Bramwell-Lalor & Rainford 2015). This could, for example, explain the disappointing results from large-scale interim assessment projects designed to use formative assessments to improve performance in high-stakes exams (see the Reading chapter for details).

8.4.3 Using technology effectively to support formative assessment

Technology can be a considerable support in learning and teaching when it offers the ability to provide formative assessment during the teaching and learning process. Digital technology use in education not only changes the types of assessment, but also the entire learning experience. The following discussion illustrates the benefits of technology-enhanced learning with evidence-based practices in different domains. In general, it is more popular to use digital technologies in the Mathematics discipline, compared with Reading, Writing, The Arts and Science. Researchers have generated the following technology-use conditions to support formative assessment effectively in schools:

1. Using adaptive computerised tools to provide students with different types of instruction prompts based on their understanding of Mathematics enhances student learning (Hsiao et al. 2017; Wongwatkit et al. 2017; Wu et al. 2017).
2. Elaborated feedback with prompts are generally more effective than feedback that only recognises errors or provides correct answers, especially in Science (Soong et al. 2010; Wang 2010).
3. It is more beneficial for students if specific feedback is provided instantly and frequently (e.g. Hsiao et al 2017; Wongwatkit et al. 2017; Wu et al. 2017). This argument appears to e apply to all discipline areas.
4. Web-based formative assessment (e.g. e-learning) can motivate students to learn actively. Formative assessment is more successful if students can learn whenever and wherever they want. Due to the online nature of the materials, students can learn in and after class. All students freely used the e-learning environment and web-based assessment to study Science-related topics (Wang 2010).
5. Technology use enables incorporation of scaffolding feedback within Physics activities (Zucker, Kay & Staudt 2014). Scaffolding can be in the form of written hints, equations or visual markers on a graph or table. The system enables

targeted hints that can respond to specific student answers, effectively creating lessons customised to student needs.

6. When using technology to support formative assessment, it is vital that teachers have the requisite knowledge and skills for computer-based assessment hardware and software (Heitink et al. 2016). In addition to providing digital technology, teachers need professional development to administer assessments, interpret results and translate information obtained into effective teaching instructions (see chapter on Reading).
7. When teachers apply digital technologies to formative assessments, the technology should be as automated as possible, and easily embedded in daily practices. Otherwise, teachers may be unwilling to incorporate technological components into their busy teaching schedule.
8. For teachers to use technology effectively in instructional activities, professional development should focus on how teachers utilise assessment data to make instructional adjustments. General training does not guarantee successful implementation of technology-enhanced classroom teaching practices (see chapter on Mathematics).

8.4.4 Limitations

The following circumstances indicate that digital technology use does not always contribute to better learning outcomes. In some circumstances, teachers prefer traditional instructional practices:

1. Feedback provided by Mathematics teachers at a minimal number of points during a longer learning session results in no improvement (Rakoczy et al. 2019).
2. Even if teachers can master the technology for class use, formative assessments supported by technology may not guarantee students receive more feedback if a teacher does not prefer using digital tools for instruction (Tolo, Chan & Hopfenbeck 2018).
3. Digital technologies work particularly well in the Mathematics and Science domains. However, application of technology in Reading, Writing and The Arts lacks evidence.
4. The lack of information on assessment practices and instruction provided to control groups throughout the studies makes it difficult to isolate the impact of online formative assessment on student learning.

8.5 Professional development

8.5.1 Studies addressing formative assessment professional development

The literature search identified 13 studies about formative assessment professional development (FAPD) that met the inclusion criteria for this review (Andersson & Palm 2017; Chen et al. 2017; Fantuzzo, Gadsden & McDermott 2011; Gallagher, Arshan & Woodworth, 2017; Randel et al. 2016; Reddy, Dudek & Lekwa 2017; Roschelle et al. 2010; Shechtman et al. 2010; Smit et al. 2017; van den Berg, Bosker & Suhre 2018; Witmer et al. 2014; Yin et al. 2015; Phelan et al. 2012). The studies varied in time, location and teacher participant types.

Despite efforts at sourcing a range of global contributions, the literature had a wholly North American–European bias. Contributions were mainly sourced from the US, where nine studies were based in diverse state and district locations (including Colombia, Hawaii, New York, New Jersey, Ohio and Texas), or high-need, rural districts across ten states (Chen et al. 2017; Fantuzzo, Gadsden & McDermott 2011; Gallagher, Arshan & Woodworth 2017; Randel et al. 2016; Reddy, Dudek & Lekwa 2017; Roschelle et al. 2010; Shechtman et al. 2010; Witmer et al. 2014; Yin et al. 2015). European contributions to the literature included three studies based in the Netherlands (van den Berg et al. 2018), Sweden (Andersson & Palm 2017), and Switzerland (Smit et al., 2017).

Table 5 shows the number of participating teachers ranged from a very small sample in some studies with numbers down to 11 (Witmer et al. 2014), to larger studies with sample sizes of more than 300 (Gallagher, Arshan & Woodworth 2017; Randel et al. 2016; Roschelle et al. 2010; Shechtman et al. 2010). Teacher participants across most studies were employed within primary education (Andersson & Palm 2017; Randel et al. 2016; Reddy, Dudek & Lekwa 2017; Smit et al. 2017; van den Berg et al. 2018; Witmer et al. 2014); primary and pre-K preparation work (Fantuzzo et al. 2011); or primary and secondary settings (Chen et al. 2017). Four studies included teachers from middle- and high-school settings only (Gallagher, Arshan & Woodworth 2017; Roschelle et al. 2010; Shechtman et al. 2010; Yin et al. 2015). Eight studies included Mathematics teachers (Andersson & Palm 2017; Roschelle et al. 2010; Shechtman et al. 2010; Smit et al. 2017; van den Berg et al. 2018; Yin et al. 2015). Five studies featured teachers of other subjects, including general education or multi-literacies (Fantuzzo et al. 2011; (Andersson & Palm 2017; Randel et al. 2016; Reddy, Dudek & Lekwa 2017); English language Arts-related literacies only (Gallagher, Arshan & Woodworth 2017; Witmer et al. 2014); or the Arts (music, visual arts, theatre, dance; Chen et al. 2017). Teachers taught from Pre-K (Fantuzzo et al. 2011) to multi-grade cohorts inclusive of high-school grades (Chen et al. 2017).

8.5.2 Positive impacts of FAPDs across multiple measures for students, classes and teachers

Formative assessment professional development (FAPDs) studies included in this review report a range of positive impacts for students and teachers. Table 6 lists all articles reporting at least *some* positive impacts for FAPDs, including those claiming largely neutral impacts overall (e.g. some report particularly positive outcomes for the teacher-level rather than other levels, Randel et al. 2016). While positive FAPD impacts are found across all student, class, teacher and coach level measures, it is important to note that some researchers (e.g. Schneider & Randel 2010) *question the rigour of evidence* linking professional development to teacher and student outcomes. These limitations are discussed when we examine the features of evidence-based, high-quality FAPD programs.

Student-level impacts: The positive impacts of FAPDs most commonly involved improvements to student-level (individualised student) subject content achievement in tests for the topic area assessed (Chen et al. 2017; Fantuzzo et al. 2011; Roschelle et al. 2010; Shechtman et al. 2010; van den Berg et al. 2018), with neutral findings for particular achievements for Mathematics students (Randel et al. 2016; Shechtman et al. 2010; van den Berg et al. 2018). The EPIC FAPD proved one of the more successful programs for Mathematics (Fantuzzo et al. 2011). Student-level positive impacts also included improvements in student subject content skills and knowledge demonstrations (Gallagher, Arshan & Woodworth 2017; van den Berg et al. 2018; Witmer et al. 2014). One study showed improvements to student knowledge of assessment concepts, including self-assessment (Witmer et al. 2014).

Class and school-level impacts: FAPDs are also linked to class-level improvements (whole-group) in subject content achievement (Andersson & Palm 2017; Fantuzzo et al. 2011), and class-level subject content skills and knowledge demonstrations (Witmer et al. 2014). One study showed improvements to student- and class-level formative assessment feedback, including improvements to both individual and class efforts at peer- and self-regulation through formative assessment use, and sense of efficacy in formative assessment use (Smit et al. 2017). Where measured, there was no improvement to school-level Mathematics achievement in a statewide test (Randel et al. 2016). A further research limitation is that some factors impacting student learning were not controlled for in these studies, including the self-regulation skills of individually motivated high-achieving students (Dignath & Buttner 2008). However, as several studies had higher student participant numbers (in the thousands), concerns over the factor's influence is somewhat mediated.

Teacher and coach levels: The studies suggested FAPDs could also lead to improvements at the teacher level. These included improvements in teachers' specialised instructional knowledge (Gallagher, Arshan & Woodworth 2017;

Shechtman et al. 2010) and content knowledge (Yin et al. 2015), with greater gains in English (Gallagher, Arshan & Woodworth 2017) than Mathematics where effects may be heavily mediated by other instructional factors (Shechtman et al. 2010). FAPDs also led to reported improvements in teachers' use of formative assessment activities in interdisciplinary contexts (Fantuzzo et al. 2011; Reddy Dudek & Lekwa 2017). FAPDs improved teachers' reported satisfaction with, or attitude to, FAPD and their own formative assessment intervention efforts, including diagnostic and feedback efforts (Fantuzzo et al. 2011; Reddy Dudek & Lekwa 2017; Smit et al. 2017; Yin et al. 2015), perceived formative assessment knowledge (Randel et al. 2016; Yin et al. 2015), self-efficacy (Yin et al. 2015) and reported teacher perceptions and evaluations of formative assessment interventions' overall success (Reddy Dudek & Lekwa 2017; Smit et al. 2017; Yin et al. 2015). Teachers reported an increase in the frequency of student involvement in classroom assessment (Randel et al. 2016). Coaches also reported improvements to teacher performance and achievement of formative assessment based on classroom observations (Reddy Dudek & Lekwa 2017). In terms of peer- and self-assessment, studies suggest teachers can have a biased perception of the effectiveness of their approach and can, at times, pay inadequate attention to strategies for building student self-regulation and self-efficacy. This is a concern given the importance of these skills for successful peer- and self-assessment (Randel et al. 2016; Smit et al. 2017).

8.5.3 Features of an evidence-based high-quality professional development program on formative assessment

The specific requirements for a successful professional development on formative assessment have not yet been established (Andersson & Palm 2018). While there are several guidelines emanating from professional development literature in general, we have very little information on how much time is required, what the necessary content is, how much outside expertise is required, and to what extent answers to these questions depend on the contexts and assessments themselves. While some studies point to the positive impact of FAPDs (see above), few have examined *how* characteristics of professional development programs on formative assessment impact teacher practices and student achievement – rigorous evidence linking professional development to teacher and student outcomes is lacking (Schneider & Randel 2010). Heitink et al. (2016), for example, identified only two studies showing a direct link between professional development and student achievement (Aschbacher & Alonzo 2006; Phelan et al. 2012). Our review identified a few more (Andersson & Palm 2017a & 2017b; Witmer et al. 2014), indicating a plausible link but one that cannot be attributed to any specific professional development characteristic. As such, we cannot recommend an evidence-based high-quality professional development program on formative assessment that would be sufficient to change teaching practices and student achievement (see also Andersson & Palm 2018).

However, research on high-quality professional development points to general attributes that improve teachers' learning experiences, including intensity, subject-matter specificity and collaboration (Corcoran, Shields & Zucker 1998; Desimone, 2009; Garet et al. 1999). The following discussion focuses specifically on formative assessment professional development and examines what is known about the content, format and intensity of such programs.

Table 5: Teachers included in formative assessment professional developments, by study

Study (researchers, year)	Location	# Teachers	Schooling level	Subject	Grades
(Andersson & Palm 2017)	Sweden	22	Primary	Mathematics	4
(Chen et al. 2017)	USA	75	Primary, Middle, High school	Arts (Music, Visual Arts, Theatre, Dance)	Multi
(Fantuzzo et al. 2011)	USA	80	Pre-K, Primary	Integrated Literacy and Numeracy Head Start	Pre-K–K
(Gallagher et al. 2017)	USA	329	Middle, High school	English language arts	7–10
(Phelan et al. 2012)	USA	36	Primary	Mathematics	6
(Randel et al. 2016)	USA	231	Primary	Mathematics	4–5
(Reddy et al. 2017)	USA	89	Primary	General education	K–5
(Roschelle et al. 2010)	USA	285	Middle school	Mathematics	7–8
(Shechtman et al. 2010)	USA	181	Middle school	Mathematics	7–8
(Smit et al. 2017)	Switzerland	45	Primary	Mathematics	5–6
(van den Berg et al. 2018)	Netherlands	34	Primary	Mathematics	4–5
(Witmer et al. 2014)	USA	11	Primary	Reading	1–2
(Yin et al. 2015)	USA	32	Middle school	Mathematics	7

Table 6: Overall finding on impact of formative assessment professional development for teachers, by study

Study (researchers, year)	FAPD hrs/ duration	FAPD focus	Level measures	Overall FAPD impact
(Andersson & Palm 2017)	122 hr/1term	Integrated formative assessment strategies	Class level achievement	Positive
(Chen et al. 2017)	Unspecified/1 year	Arts achieve criteria-referenced formative assessment (rubric formative assessment, teacher-, peer- and student-led formative assessment, revision).	Student level achievement	Positive
(Fantuzzo et al. 2011)	50–90 hr/ 2years	Epic program including formative assessment training in curriculum-based assessments, and learning community model of PD based on distributed leadership (reciprocal t&l)	Student- and class-level achievement, teacher-level team reporting on activities' use and satisfaction	Positive
(Gallagher et al. 2017)	90 hr/2years	National writing project’s college-ready writers program (crwp), professional development paired with supporting curricular resources and a standards-based formative assessment tool/ rubric	Student-level source-based argument writing, teacher-level argument-writing instruction practice	Positive

Study (researchers, year)	FAPD hrs/ duration	FAPD focus	Level measures	Overall FAPD impact
(Phelan et al. 2012)	Unspecified/4 meetings	Initial meeting: teachers given advice on how to look at and use student data to gather information on student understanding and to modify instruction Follow-up meeting: teachers had the opportunity to look at student assessment data from within their district	Student-level achievement	Positive
(Randel et al. 2016)	Variable/ 2 years	Classroom assessment for student learning (CASL), a classroom and FAPD program	Student level and school level achievement, teacher level self-reporting on FA use, FA knowledge and FA student involvement	Neutral–positive
(Teddy et al. 2017)	2 hr/4–9 weeks	Data-driven classroom strategies coaching (CSC): identifying teachers' practice needs, goals, plans and progress to improve use of evidenced-based classroom-level management practices including formative assessment with a validated observation instrument; and visual performance feedback	Coach-level teacher observation, teacher-level self-reporting on activities' use, and perceived intervention success	Positive
(Roschelle et al. 2010)	3–7days/ 3years	The simcalc approach, which integrates an interactive representational technology, paper curriculum, and teacher professional development	Student-level achievement	Positive
(Shechtman et al. 2010)	4–5days/2 years	Content knowledge, use of curriculum materials and/or planning specifically how to use the materials in formative assessment	Student-level achievement and teacher-level specialised math instructional knowledge	Neutral–positive (mostly neutral)
(Smit et al. 2017)	1 day + 9 lessons/2 years	Using rubrics to teach and assess progress in mathematics	Student- and class-level FA feedback, peer-/self-regulation and efficacy. Teacher-level diagnostic skills, FA feedback, peer-/self-assessment and beliefs	Positive–neutral (mostly positive)
(Van den berg et al. 2018)	Unspecified/1 year	Daily and weekly cycles of assessing students' mastery of learning goals (goal-directed instruction, assessment and immediate instructional feedback)	Student- and class-level achievement	Neutral–positive (mostly neutral)
(Witmer et al. 2014)	14.5 hr/1 year	Administration And use of COCA comprehension testing data	Student- and class-level skills and knowledge demonstrations	Positive

Study (researchers, year)	FAPD hrs/ duration	FAPD focus	Level measures	Overall FAPD impact
(Yin et al. 2015)	100 hr/3 years	Formative assessment and texas instruments navigator (nav) technology processes	Teacher-level formative assessment perceived knowledge, self-efficacy, attitude and evaluation	Positive–neutral (mostly positive)

8.5.4 Content

Formative assessment practice is difficult. In the past, it has been identified as one of the weakest aspects of teacher practice (Assessment Reform Group 1999). Using the assessment information to plan subsequent differentiated instruction is especially challenging (Heritage et al. 2009; Schneider & Meyer 2012) as it relies strongly on pedagogical content knowledge that teachers may or may not have. Further, as is true for many instructional practices, teachers' ability to problem-solve when an activity does not work as planned may depend on their understanding of the activity's theoretical underpinnings – in other words, on having deep pedagogical content knowledge.

When assessing their successful professional development program, Andersson and Palm (2018) reported that their participants identified the availability of activities directly usable in the classroom and the lectures on theoretical background underlying these activities as important characteristics. According to the participating teachers, a mixture of theory and practice provided a structure within which they could process their learnings. They also appreciated the value of usable activities for pedagogical reasons, as it allowed them to experiment in classrooms, observe the value of formative assessment activities on student learning, and get comfortable with using activities in daily teaching. All but one participant named the almost immediate theory-to-practice link as an essential characteristic of professional development that helped them change their practice.

Brink and Bartz' (2017) three in-depth case studies show that, when teachers are provided with specific information about formative assessment through staff development, they become more positive toward such assessment, and their implementation skills are greatly improved. Staff development had an especially positive impact on teachers' understanding and skill sets for individualising instructional practices – a critical component of pedagogical content knowledge.

In turn, Yin et al. (2015) examined the importance of general assessment knowledge on implementing formative assessments using networked classroom technology. They compared two versions of professional development – technology then assessment and assessment then technology – finding teachers preferred to be taught about assessments first. While both groups implemented formative assessments equally well after two years, the assessment-first group reached that level more quickly than the technology-first group.

Finally, teachers' attitudes towards assessments can provide a real challenge for implementation (Borko et al. 1997; Keuning, Van Gell & Visscher 2017). After noticing participating teachers were likely to ignore ideas and practices incompatible with their own, Borko et al. concluded if they were to 'embark on another staff development effort, we would build in explicit attention to beliefs as well as practices' (1997, p. 27).

In summary, high-quality professional development needs to be grounded in specific subject matter and increase teachers' content understanding and the diverse ways to teach it. Teachers must know what they are teaching and what they can reasonably assess. They must know the content and processes well enough to understand the basis of students' potential misconceptions. They also need to broadly understand assessment to know how to get to those misunderstandings. While it is possible that understanding the theory and practice behind successful formative assessment is enough to create a positive attitude towards the assessments, it may be necessary to include scaffolded practice before teachers are completely comfortable with changing their day-to-day practice. Many interventions reviewed in this document offered professional development for content knowledge and the learning process as well as assessment particulars, deepening teachers' capacity to use assessments formatively within their subject-matter contexts.

8.5.5 Format and duration

While no studies experimented with the professional development format (with the exception of sequencing in Yin et al. 2015), it appears likely some form of process-oriented professional development with ample opportunities for collaboration, discussion and learning from more knowledgeable others (experts, administrators, coaches or other teachers) is needed. Product-oriented short-term professional development workshops are less likely to change practice effectively. For example, Andersson and Palm's (2018) teacher participants highly valued their process-oriented and relatively long professional development. Their program was more extensive than most, in that it included 24 weekly six-hour meetings at a university over one term plus another 72 hours for reading, planning and reflection. A typical weekly meeting included a lecture on the theory of formative assessment and its research base, examples of concrete activities for its implementation in the classroom, group discussions focused on planning new formative assessment activities, and discussions about experiences from the previous week's implementation. In brief, the spiralling structure of the professional development itself exemplified good formative assessment practices, with ample time for learning from more knowledgeable others, feedback and collaboration.

However, it is also likely that the length of Andersson and Palm's program was an important contributor. Klinger, Volante and Deluca (2012) examined the impact of a shorter, three-part process-oriented professional development on teachers' evolving conceptions of classroom assessment. This consisted of three half-day sessions spread over a four- to eight-week period. Each session focused on specific aspects of formative assessment and combined practical examples, readings, videos and discussions to help participants explore classroom assessment issues and practices. In between sessions, teachers were expected to implement what they learned within their classrooms. Klinger, Volante and Deluca (2012) noted that teachers continued to struggle with understanding the theoretical foundations of formative assessments and how to further develop their assessment practices after the professional development. While they valued the professional learning community created, it is likely a much more prolonged effort is required than what the study provided. Smaller FAPD hours and briefer intervention durations are also cited as problematic in a number of other studies (Randel et al. 2016; Smit et al. 2017; van den Berg et al. 2018). Klinger et al. (2012) concluded that while current models of professional development may be more aligned with principles of effective professional learning, genuinely changing teachers' classroom assessment practices may require a much more prolonged effort than most projects offer.

8.5.6 School leadership and quality coaching

Sustained and proactive school leadership by principals, departmental heads and/or school-based expert coaches is required for successful FAPD implementation (Andrade & Cizek 2010; Sanzo, Myran & Caggiano 2014; Wiliam 2007; Yin et al. 2015). Strong school leadership is also associated with higher fidelity FAPD implementation by teachers (Moss, Brookhart & Long 2013). School leadership could potentially be coached to perform formative assessment leadership key tasks, including, for example, seeking out grade-level and class-level patterns across cohorts to understand if stronger or weaker formative assessment results emerge from particular pedagogies or conditions (Andrade & Cizek 2010; Sanzo, Myran & Caggiano 2014). Leaders can also be trained to diagnose and 'treat' key issues by mimicking the role of 'coach' in FAPDs after their completion – potentially addressing problems by, for example, arranging team teaching and mentoring opportunities, teacher-swaps for particular content areas across a grade cohort or for specific classes, or other intervention responses refocusing teaching efforts towards meeting standards and rubrics as needed (Andrade & Cizek 2010; Reddy, Dudek & Lekwa 2017; Sanzo, Myran & Caggiano 2014). These leadership efforts can occur over several years to identify both short-term and long-term patterns for different departments and teachers, including the variable strengths and any needs for further FAPDs.

8.5.7 External expertise

How experts are used may be as critical for successful professional development as whether they are used. In traditional professional development models, experts are typically used to deliver one-time workshops (or less frequently a series) aimed at increasing teachers' knowledge and skills in implementation, then leave individual teachers with little or no continuing support (e.g. Penuel et al. 2007; Wylie & Lyon 2009). Therefore these approaches are premised on the notion that introduction of new ideas is sufficient for professional learning and, consequently, teaching practices and student learning. The expert's 'role is to convey that knowledge in a clear, concise manner; the learner's role is to absorb it' (Osterman & Kottkamp 2004, p. 14).

Research over the past 20 years has repeatedly illustrated that such approaches have little effect. Instead, professional learning that is sustained, collaborative, work-embedded and situated within school needs is preferred, especially over one-day workshops or formally presented interventions (Garet et al. 2001; Gersten et al. 2010; Huberman & Miles 1984). The argument is that professional development is most effective when teachers engage actively in instructional inquiry in the context of collaborative professional communities, focused on instructional improvement and student achievement (Wei et al. 2009). However, this success is conditional on teachers being able to understand assessment theory and research, and how to translate the into concrete classroom practices (Robinson et al. 2014). In this model, the expert's role is to enhance theoretical understanding, help teachers design and test formative assessment practices and, perhaps most importantly, support ongoing changes in instruction (Speck & Knipe 2005). Continuous support may be necessary because application of new knowledge into an ongoing practice on the fly is a cognitively complex problem-solving activity (Marzano & Kendall 2007) that teachers are initially asked to engage in on top of their existing workload.

Against this background, it is not surprising that Andersson and Palm's (2018) participants highly valued both the expert's theoretical content and support during the professional development process. Their expert was a university researcher with expertise in formative assessment and Mathematics (all participants were Mathematics teachers). Other studies have found that school-district or school-based experts in dedicated professional development roles can be a key factor in helping teachers make connections between assessment results and instructional actions (Goertz, Olah & Riggan 2009; Means et al. 2009), but they can also undermine change (Keuning, Van Geel & Visscher 2017). Experts can model concrete classroom techniques, and their interactions with teachers can create professional accountability for changes in classroom practice (Lee & Wiliam 2005). Teachers need follow-up and support if new ideas and strategies are to be effectively implemented (Klinger, Volante & Deluca 2012).

In summary, it seems expert involvement in a continuous and cyclical manner is likely to be more beneficial than not involving them or involving them in content delivery only. However, the extent to which successful professional development requires the presence of external experts is not clear, and it is likely to be modulated by the presence and effectiveness of teacher collaboration in schools.

8.5.8 Collaboration

Teachers learn from each other, especially when they are able to share their experiences while exploring authentic assessment and instruction problems occurring in their classrooms. How this learning is organised and supported will probably have a large impact on the effectiveness of any professional development program that aims to change classroom practice. Collective participation by multiple teachers and/or students sharing their development work together at individual school sites or across networks was seen as an important factor in FAPD success across the studies reviewed for this report. Collective participation featured in several of FAPDs studied and was consistently discussed as a positive factor in aiding formative assessment implementation (Chen et al. 2017; Fantuzzo et al. 2011; Gallagher, Arshan & Woodworth 2017; Randel et al. 2016; Reddy, Dudek & Lekwa 2017). Indeed, Table 7 shows participating teachers who were lone representatives from their school cited a lack of collective participation by other school staff as an issue mediating the impacts of FAPDs, and in some cases even sought to overcome this barrier by bringing on board school leadership, team-teachers, school-based peers or local teacher networks (particularly in van den Berg, Bosker & Suhre 2018). FAPDs that lacked 'coherence with instructional context' and onsite 'active learning for teachers' had mainly neutral impacts overall (Shechtman et al. 2010; van den Ber , Bosker & Suhre 2018)

Practice-centred collaboration, often in the form of school-based professional learning communities, was reported in several studies as a critical ingredient of effective implementation of formative assessment practices (e.g. Accardo & Kuder 2017; Hargreaves 2013). Teachers benefit from conversations with colleagues about formative assessment and teaching, and collaboratively solving shared problems and dilemmas (Birenbaum, Kimron & Shilton 2011; Feldman & Capobianco 2008; Lyon & Leahy 2009). Proponents of professional learning communities argue they promote the pedagogical knowledge and skills critical for correctly diagnosing students' states of understanding (explaining the gap between observed and intended outcomes) and making the right instructional adjustments to close the gap. Research on teachers' knowledge points to a common deficiency in pedagogical knowledge, especially with regards to differentiating instruction to close the gap (e.g. Heritage et al. 2009). Hargreaves (2013) notes, however, that benefits associated with professional learning communities (her term was 'teacher learning communities') are compromised when imposed on teachers, not accommodated sufficiently within other school commitments, meetings are too directive and inflexible, and when emphasis is solely on practice at the expense of theories. It appears critical that professional learning communities cover pedagogical content knowledge, including differentiated instruction. This last point may require input from within and beyond the school to support teachers' theoretical and practical learning. Hargreaves concludes both formative assessment and learning communities require sustained critical reflection among their participants to succeed. Assessment data is not formative unless teachers make use of the information for instructional practice or program design. Therefore, to the extent that teachers' joint efforts underpin this critical step of bridging data analysis and instruction-related decisions, collaborative structures may be a key lever in changing how teachers develop and refine their repertoire. Collaboration and access to others' instructional expertise may be particularly valuable for novice teachers by helping them make sense of assessment results and learn what can be done with them (Young 2006).

Lyon and Leahy (2009) identify four processes that help learning communities develop better assessment practices: collaborative problem solving; joint customisation of existing techniques and creation of new techniques; shared examples of positive feedback from students, teachers and administrators; and commitment to the group.

It is tempting to argue that collaboration is desirable – even necessary – to building capacity for incorporating formative assessment results into instructional decision-making and practices. At this point, however, we know little about how collaborations should be structured, what content should be covered, whether norms of trust and learning can be developed simultaneously or must be in place for teachers to begin fruitful discussions, and what additional supports teachers need to leverage the time they spend together. We have argued above that input from experts is probably beneficial, if not necessary. We have also noted successful collaborations require time. Andersson and Palm (2018) suggest two types of time are needed: time for learning and practising new knowledge and skills, and time for regular meetings spread out over the term (with time in between to experiment).

8.6 Professional development conclusion

For over 30 years, a great deal of research noted the limited benefits of some of the most popular and traditional forms of professional development such as workshops, conferences and using guest speakers to introduce new ideas (e.g. Darling-Hammond et al. 2009; Fullan 1993; Wylie & Lyon 2009). Traditional forms of professional development require minimal time commitments and famous speakers add great appeal to annual reports. They are problematic because they treat teachers as recipients of frequently de-contextualised knowledge that must then be acted on – and the 'act on' part (the only part that matters) is then left to individual teachers, often with little or no follow-up and support (e.g. Hargreaves 2007; Penuel et al. 2007). Many studies reviewed models of professional development more closely aligned with effective learning principles (e.g. Andersson & Palm 2018; Klinger, Volante & Deluca 2012), but the professional learning opportunities provided in many projects reviewed frequently appear severely lacking in terms of time commitment, breadth of content and opportunities for reflection. This may partly explain why many well-designed interventions produce disappointing results.

Table 7: Methodological and other limitations in FAPD studies

Study (researchers, year)	**Reported method**	**Limitations noted by researchers**
(Andersson & Palm 2017)	Randomly sampled evaluation of an FAPD on a unity of different, integrated strategies.	Low participant number. Analysis at class level only. Not controlled for teachers' adjustment of teaching based on collected evidence of student learning, self-regulated learning, collaborators and other whole-school efforts. Several teachers cited broad spread in students and students disturbing learning environments as aggravating conditions. Researchers want longer duration.
(Chen et al. 2017)	Large-scale experimental study with a control group.	FAPD hours unspecified. Some missing data for individual participants. Low internal consistency of the visual arts measures for elementary and middle schools and high school dance. Inter-rater reliabilities for some tasks were low. Researchers call for random assignment at the student level.

Study (researchers, year)	Reported method	Limitations noted by researchers
(Fantuzzo et al. 2011)	Development and random field trial comparison of an integrated curriculum with wait-listed control group (delayed FAPD).	Various teachers/classes/students dropped out of the study over the years. Teacher-level team reporting on activities' use and satisfaction may be subject to a 'party-line' approach as it lacks localised internal anonymity and may relate to view of success/achievement in the eyes of others in place of employment.
(Gallagher et al. 2017)	District-randomised controlled trial with a control group.	Various teachers/classes/students dropped out of the study over the years, more than 40%, moving districts/jobs/careers. Teachers were replaced by their classroom replacements in many cases, late joining of project thus common.
(Randel et al. 2016)	Impact study of classroom assessment for professional development program in classroom and formative assessment under real-world/'natural' conditions (no researcher involvement).	Various schools withdrew from the study at orientation resulting in the loss of 6 intervention and 11 control teachers. Some teachers had different experiences of the program as 29 teachers transferred into the CASL schools, and 53 teachers transferred into the control schools. These late-entry teachers participated in learning-team activities. Missing scores for some students who transferred out. CASL implementation fidelity below recommendations. CASL group schools were provided FAPD materials, videoconference and district-level facilitator. However, training was internal/'natural' – with no involvement/requirements from research team. It was thus variable.
(Reddy et al. 2017)	A randomised controlled trial with two conditions: (1) an immediate coaching and (2) a wait-list control group; measured by comparing the post-coaching evaluation on teachers' progress to control.	Teacher-level team reporting on activities' use and perceived intervention success may be biased as it may relate to view of success/achievement in needed employment. Unclear if FAPD impacted student learning. Unclear how specific coaching components and rating processes influenced fidelity and outcomes sustaining teacher and student learning and social behaviours and how these strategies work together to orchestrate enriched classroom environments for all students including those with disabilities.
(Roschelle et al. 2010)	Three studies (two randomised controlled experiments and one embedded quasi-experiment) evaluating replacement units targeting learning.	Various teachers dropped out of the study prior to formative assessment, after formative assessment, in subsequent years for a variety of reasons. Researchers call for expanding the approach of integrating software, curriculum, and teacher professional development to cover the key ideas in algebra, geometry and statistics.
(Shechtman et al. 2010)	Two large-scale randomised experiments with one embedded quasi-experiment using classroom data.	Researchers discuss a need in the field for richer models of how 'mathematical knowledge for teaching' works in the context of complete instructional system. Various teachers dropped out of the study.
(Smit et al. 2017)	Quasi-experimental study with quantitative longitudinal analyses using curriculum, achievement tests and questionnaires.	Some missing data for individual participants due to drop-out or item-skipping on questionnaires.

Study (researchers, year)	Reported method	Limitations noted by researchers
		Low sample size on the second level of the calculation of the standard errors in the complex SEM models mean results related to student achievement might not become significant until a larger experimental group is considered. Effects on students' outcomes could not be observed. Potential sample bias for motivated teachers.
(van den Berg et al. 2018)	A quasi-experiment to compare 2 conditions including treatment and control, with achievement tests.	FAPD hours unspecified. Analysis at student level and class level only. Some missing data (student illness, test not taken). Not controlled for teachers' skill in formative assessment application. Researchers asserted no 'business-as-usual' dynamic. Selection bias may have occurred; schools disinclined to participate in one of the conditions withdrew.
(Witmer et al. 2014)	Randomly assigned experimental PD on how to administer and interpret formative assessment in classrooms.	Low participant number. Analysis at student level and class level only. Some missing data for individual participants. Not controlled for teachers' adjustment of teaching based on collected evidence of student learning; self-regulated learning, colleague collaborators and whole-school efforts.
(Yin et al. 2015)	A teacher survey-based comparison of two FAPDs designed for networked classrooms with randomised groups.	Due to the longitudinal feature, one teacher failed to finish the questionnaires at pre-test, and four teachers failed to finish the post-test 3. Therefore, the sample size varies from 26 to 30 in the analysis. Analysis at teacher level only; self-assessment.

8.7 Part B: Teacher-level factors

8.7.1 Role-based conditions

This section provides a review of teacher-level factors affecting successful FAPD implementation. It includes an analysis of two possible knowledge deficits most frequently suggested as hindering implementation of effective formative assessment practices: lack of assessment knowledge and lack of pedagogical content knowledge (PCK).

8.7.2 Teacher knowledge and skills

As Heitink et al. (2016) recently completed a systematic review of factors affecting the implementation of assessment for learning in classrooms, this section begins by summarising their findings. Note they first divided formative assessments into three 'distinct approaches': (1) data-based decision-making (defined as 'involving systematic collection and analysis of data to inform decisions that focus on improvement of teaching, curricula and school performance'), (2) diagnostic testing ('mapping out of individual learners' task-response patterns to reveal their (possibly inadequate) solution strategies and using this as an indication of each learner's developmental stage'), and (3) assessment for learning (AfL). The authors broadly define AfL as a formative assessment approach that occurs as part of ongoing classroom practices and viewed as social and contextual, and focusing on the quality of the learning process. This broad definition leads them to include mostly qualitative and mixed-methods studies, making it narrative by nature.

8.7.3 Pedagogical content knowledge

Twenty-one of the 25 studies reviewed by Heitink et al. (2016) examined teacher knowledge and skills while 19 examined teacher beliefs and attitudes. In terms of knowledge and skills, the main conclusion was that teachers need diverse sets of knowledge and skills to successfully collect, analyse and interpret evidence from assessments, and to adapt instruction correspondingly. PCK was the most commonly noted prerequisite for successful AfL. Several reviewed papers suggest that, without full understanding of the content being learned and common misconceptions and learning gaps, it is unlikely teachers can provide accurate and complete feedback. Several papers not included in the review reached the same conclusion, particularly with regards to Science. For example, Sabel, Forbes and Flynn (2016) reported teachers with higher levels of Science content knowledge evaluated students' ideas more effectively than teachers with lower levels of content knowledge. Haug and Ødegaard (2015) studied primary school Science classes, concluding that when PCK is lacking, teachers' interpretation of students' responses and their subsequent instructional modifications are unlikely to align with the scientific idea that the key concepts represent. Therefore, it is likely that teachers' ability to design formative assessments and interpret the information accurately requires substantial knowledge of both the content being taught and the instructional approaches that can be adopted in response to assessment information (Haug & Ødegaard 2015). Teachers with strong content knowledge are more likely to identify and flexibly adapt instruction to a student's place in the knowledge-acquisition trajectory (Aschbacher & Alonzo 2004; Duschl & Gitomer 1997; Fennema et al. 1993). Teachers with strong content understanding are also more adroit at considering students' learning in direct relation to the content rather than in relation to more general developmental terms (Johnston, Afflerbach & Weiss 1993). To sum up, when teachers' pedagogical content knowledge is deep and flexible, they can break down critical concepts, find appropriate entry points for all students, and redesign instruction to match students' understandings and misconceptions, as evidenced in their formative assessments.

Some researchers also suggest that implementing formative assessments can be a powerful opportunity for teachers to use, integrate and generate PCK (Falk 2012). However, we suspect that for this process to succeed, substantial professional development on instructional strategies must be available, particularly in terms of differentiated instruction that many teachers find difficult (e.g. Fuchs & Vaughn 2012).

Finally, we should note the association between PCK and formative assessments is not uniformly found. For example, Forbes, Sabel and Biggers (2015) find primary teachers' knowledge of geoscience disciplinary content is unrelated to their formative assessment practices in the subject. Herman et al. (2015) suggest that, despite weaknesses in teachers' PCK and assessment knowledge, their formative assessment practices were positively related to student learning.

8.7.4 Assessment knowledge

In contrast to Herman et al.'s (2016) optimistic stance, several studies identify teachers' lack of assessment knowledge as problematic and sometimes difficult to improve (e.g. Sabel, Forbes & Flynn 2016). Sixteen studies in Heitink et al. (2016) examine one or more aspects of assessment knowledge, or integration of PCK and assessment knowledge. The authors concluded teachers need knowledge and skills to develop assessments that achieve the desired purpose. This includes the ability to construct questions that elicit reliable and valid evidence about student learning and to critically evaluate assessment instruments. Experience with formative assessment and confidence in their professional judgement were associated with deeper understanding of formative assessment, confidence in instructional decisions and, as a result, successful formative assessment implementation (e.g. Birenbaum, Kimron & Shilton 2011; Fletcher & Shaw 2012). Brink and Bartz' (2017) in-depth case studies similarly show that, when teachers are provided with relevant information about formative assessment through professional development, they are more positive about using formative assessments and their implementation is greatly improved. In turn, Abrams et al. (2015) note that the teachers they studied were able to use benchmark test results formatively when provided with valid and transparent test items, support to understand items and time to discuss results with other teachers. In summary, it seems while teachers' insufficient assessment knowledge may hinder successful implementation of formative assessments, this problem can be addressed with targeted professional development and school-based supports.

Finally, when formative assessment is computer-based, teachers need to have knowledge and skills regarding hardware and software use (Feldman & Capobianco 2008; Lee et al. 2012). However, Tolo, Chan and Hopfenbeck (2018) note that even when teachers have the required digital competencies, formative assessments supported by technology may not lead to more data-based instructional planning or feedback to students if a teacher's pedagogic preferences and orientations (more on this below) do not align with tool usage.

While existing literature on the relationship between teacher knowledge and assessment practices is partly mixed, it seems a vicious cycle leads from insufficient initial teacher preparation and further professional development opportunities to a lack of assessment and pedagogical content knowledge, and to difficulties in implementing formative assessments and differentiated instruction successfully. Further, teaching experience alone does not solve the problem. There is potential to break this cycle, however, with appropriate professional development and school-level supports. These findings underscore both the potential and challenge of bringing effective formative assessment practices to fruition, as well as the need for continued research.

8.7.5 Teachers' pedagogical beliefs and attitudes

Teachers' beliefs about and conceptions of teaching influence all aspects of their teaching, including formative assessment practices. Teachers may not view formative assessments as integral to their instructional practice (De Lisle 2015), and their beliefs, attitudes, perspectives and philosophy about teaching and learning can influence implementation quality (e.g. Havnes et al. 2012; Lee, Feldman & Beatty 2012; Sach 2013). Teachers' beliefs about content (Young 2006), what they regard as valid assessment approaches (McMillan & Nash 2000), and what they consider the role of assessment is in instructional planning (Torrance & Pryor 2001) can lead to ignoring assessment practices and results that are inconsistent with those beliefs. For example, teachers who feel responsible for student attainment and learning rather than just covering the curriculum are more likely to evaluate student work, give effective feedback and revise instruction as needed (Aschbacher & Alonzo 2006; Birenbaum, Kimron & Shilton 2011). In a study aimed at helping teachers design and implement classroom-based performance assessments, Borko et al. (1997) noted teachers were likely to ignore new ideas and practices incompatible with their own philosophies.

Several studies have specifically noted one particular teacher belief. They report teachers with a constructivist view of learning and teaching are more likely to use formative assessments than teachers taking a more instructivist approach (e.g. Penuel et al. 2007; Rakoczy et al. 2008). For example, Box et al. (2015) suggest high school Science teachers' view of learning and teaching plays a critical role in shaping their assessment practices and affects their ability to convert promoted theories about assessment into actual classroom practice. Similarly, Goertz, Olan & Riggan (2009) suggest

Mathematics teachers who focused on students' conceptual understanding showed better diagnostic and analytic abilities and used higher quality formative assessment practices. These teachers were also more likely to respond to assessment results with instructional rather than organisational changes: they provided alternative means of representing mathematical concepts or tried to activate students' prior knowledge, as opposed to reteaching, grouping students, or identifying specific students for additional supports. Focus on building conceptual understanding is clearly in line with the constructivist approach.

Finally, we should note that the negative connection between traditional conceptions of teaching and learning, instructivist approaches and formative assessment is not always evident. Alt's (2018) recent study of 127 Israeli primary Science teachers did not show a significant negative connection between teachers' traditional conceptions of teaching and learning and their tendency to use constructivist activities and formative assessments in their classes. Alt interprets this data to indicate that, in practice, teachers need to respond to school-level expectations to use inquiry-based activities and formative assessments – they may therefore use these even if it conflicts with their beliefs about teaching and learning. Unfortunately, Alt's study does not examine whether the quality of the practices varied as a function of teaching and learning conceptions.

To conclude, we quote Borko et al. (1997, p. 27) who notes if they were to 'embark on another staff development effort, we would build in explicit attention to beliefs as well as practices.' In light of the above studies, this conclusion seems to hold, even 22 years later.

9 Chapter 9: Conclusion and recommendations

9.1 A problem of definition

Some researchers define formative assessment as any interaction that generates data on student learning that is used to inform teaching and learning content and strategy. The academic community uses a number of definitions for formative assessment, and there is no consensus on a defined set of practices for its successful implementation in schools. This makes it difficult to compare studies of formative assessment's effectiveness.

9.2 Conditions for success are unclear

While the current academic literature assumes formative assessment leads to better student learning outcomes, this literature review demonstrates we know very little about the conditions that enable effective implementation of formative assessments, including the optimal school and educational system structures and supports.

As a body of work, the literature lacks clarity and consensus about what formative assessment is, how it is executed and the conditions for enhanced efficacy. Claims about effect sizes made in a number of the meta-analyses reviewed can be questioned on the basis of included studies and the lack of samples large enough to allow determination of effects for specific student groups. In terms of individual studies, more rigorously designed experimental research (particularly for Arts learning and Writing) need to be undertaken to establish the effectiveness of different formative assessment practices.

9.3 Findings for impact on teaching practice and student learning are mixed

This review found the impact of formative assessment on teaching practice and student learning progress/outcomes were mixed. Significant findings were reported in some studies, but benefits were isolated to specific groups. Many papers omitted details of control group activities, making it impossible to distinguish formative assessment's influence on students. To make matters more complicated, research shows formative assessment in different fields is often non-transferable, requiring field-specific evidence-based best-practice models to be established.

Self-assessment using a combination of scripts and rubrics has the potential to improve student performance and achievement in Years 7–8 Science. A combination of group discussion, feedback from peers and feedback from teachers has the potential to improve Year 9 student Science performance. Research also shows that peer and self-assessment practices can help develop self-regulation and motivation in Writing tasks involving primary students.

Greater benefits to student learning do appear when students are provided with targeted, individualised feedback immediately and frequently, suggesting embedded formative feedback may be critical to the effective implementation of formative assessment, particularly in Mathematics and Science. Elaborate feedback with prompts is generally more effective than merely recognising errors or giving correct answers. However, more investigation is required before firm conclusions can be drawn.

9.4 Technology likely to impact outcomes in tandem with other factors

Most studies included formative assessment technologies, tools and resources, and results show that the most successful have: (1) a valid task model demonstrating the sequence of activities to be done to meet learning outcomes and how to move through them, (2) a valid cognitive model outlining prerequisite cognitive and learning skills for successful progression, and (3) an evidence-based instructional intervention.

It is important to note that the impact and effectiveness of tools also differs across disciplines, with Arts outcomes being the most difficult to assess. A variety of assessment tools are required to effectively improve these skills, including observation, student–teacher collaboration and self-, peer and teacher feedback.

9.5 Professional development and technology alone won't achieve optimal effects

Many professional development studies show no effect on student learning because links with practice are not investigated. Simply providing professional development, much like introducing technology in isolation, does not automatically equate to adapted assessment and improved instructional practices in the classroom.

However, high-quality professional development research suggests general ways to improve teacher learning experiences of formative assessment. Extended, process-oriented professional development that focused on collaboration, feedback and discussion was more effective than brief pedagogical interventions in successfully changing classroom assessment practices.

Practice-centred collaboration within schools or across networks of teachers is a critical ingredient for effective practices, as is collectively sharing developmental work and actively engaging in instructional inquiry.

9.6 Conditions for effective implementation

Environmental conditions and teacher-level factors both play a role in the effective implementation of formative assessment practices. Creating optimal educational system structures, supports and conditions requires school leaders who understand formative assessment, can provide a rationale for its use, can create a supportive and non-threatening environment and model the effective use of assessment data. School leaders have a responsibility to ensure that accountability pressures on teachers do not lead to unintended impacts on instruction and assessment practices.

Strategically aligning expertise and resources to facilitate teacher learning and develop communities focused on improving formative assessment practice are necessary to achieve optimal implementation. Increased focus on student teachers' assessment and data literacy, and a classroom philosophy that views mistakes as opportunities to learn are key to achieving better results for students.

Deep, flexible pedagogical content knowledge helps teachers achieve optimal knowledge, skills and beliefs about formative assessment. This allows them to adapt instruction techniques that will better address student learning gaps and misconceptions.

Our research found optimal use of computer-based formative assessment is dependent on teachers' pedagogic preferences and orientations. Teachers also require an awareness of the valid task models, cognitive models and evidence-based interventions for addressing learning gaps. Different disciplines require different technological interventions. Particularly, it was found occasional formative assessment in Reading using digital technologies without other activities resulted in smaller impacts on student learning.

In summary, while it is difficult to find consensus in the literature about the steps towards successful implementation of formative assessment, a number of themes do present themselves. The best outcomes are seen when teachers have deep pedagogical content knowledge and knowledge of assessment practices and can use both to target individual and classroom knowledge gaps. Providing targeted, tailored feedback to students immediately and often supports improved student learning outcomes. School leaders who support collaborative and continual formative assessment modelling, learning and practice in a non-threatening pedagogical environment are best placed to foster increased formative assessment results. Finally, adjusting formative assessment teaching and technology for each discipline leads to more effective results.

9.7 Recommendations for key stakeholders

9.7.1 Initial teacher education providers

- Provide learning opportunities to improve graduate teachers' pedagogical content knowledge, assessment and data literacy.

9.7.2 Researchers

- Develop a consensus around the nature and definition of formative assessment to guide research and practice.
- Undertake rigorous experimental design studies to explore the links between specific aspects of formative assessment (for example, peer assessment and feedback) and student learning outcomes across K–12.
- Undertake studies of formative assessment professional development that focus on the fidelity of implementation of formative assessment practices.
- Undertake research to identify learning progressions for key concepts in the K–12 curriculum.

9.7.3 School systems

- Undertake a review of current formative assessment practices and supports (including professional development) across school systems.
- Allocate appropriate resources to support the implementation of formative assessment. This is essential for ensuring the fidelity of implementation.

9.7.4 Schools and school leaders

- Provide teachers with ongoing support with both hardware and software related to online formative assessment.
- Ensure that technology is integrated with the curriculum rather than an add-on.
- Promote the sharing of formative assessment practices both within and between schools.
- Ensure that school leaders and executive understand formative assessment, create a supportive and non-threatening environment and allocate appropriate resources to support the implementation of formative assessment.

- Design professional assessment opportunities that focus on building teachers' pedagogical content knowledge, assessment and data literacy.

9.7.5 Teachers/educators

- Ensure that feedback is individualised, timely and aligned with the curriculum.
- Provide elaborated feedback with prompts rather than information about errors and correct answers.
- Consider the nature of the content and/or skill domain when selecting formative assessment tools (including online tools). Tools and resources need to be fit for purpose.
- Increase awareness of the valid task models, cognitive models and evidence-based interventions for addressing learning gaps.
- Integrate formative assessment practices and technology as a regular component of the curriculum.

10 References

Abrami, PC, Venkatesh, V, Meyer, EJ & Wade, CA 2013, ‘Using electronic portfolios to foster literacy and self-regulated learning skills in elementary students’, *Journal of Educational Psychology*, vol. 105, no. 4, pp. 1188–1209.

Abrams, LM, McMillan, JH, & Wetzel, AP 2015, ‘Implementing Benchmark Testing for Formative Purposes: Teacher Voices about What Works’, *Educational Assessment, Evaluation and Accountability*, vol. 27, no. 4, pp. 347-375.

Abu-Hamour, B & Mattar, J 2013, ‘The Applicability of Curriculum-Based-Measurement in Math Computation in Jordan’, *International Journal of Special Education,* vol. 28, no. 1, pp. 111-119.

Accardo, AL and Kuder, SJ 2017, ‘Monitoring Student Learning in Algebra,’ *Mathematics Teaching in the Middle School,* vol. 22, no. 6, pp. 352-359.

Act, N.C.L.B, 2002, ‘No child left behind act of 2001’, *Publ. L*, pp.107-110.

Al Otaiba, S, Connor, CM, Folsom, JS, Greulich, L, Meadows, J & Li, Z 2011, ‘Assessment data-informed guidance to individualize Kindergarten reading instruction: Findings from a cluster-randomized control field trial’, *Elementary School Journal, 111*, pp. 535–560.

Albers, CA & Hoffman, A 2012, ‘Using Flashcard Drill Methods and Self-Graphing Procedures to Improve the Reading Performance of English Language Learners’, *Journal of Applied School Psychology*, vol. 28, no. 4, pp. 367–388.

Alt, D 2018, ‘Science Teachers’ Conceptions of Teaching, Attitudes Toward Testing, and Use of Contemporary Educational Activities and Assessment Tasks', *Journal of Science Teacher Education,* vol. 29, no. 7, pp .600-619.

Amabile, TM 1983, *The social psychology of creativity*, Springer, New York.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US), 2014. *Standards for educational and psychological testing*, American Psychological Association, Washington DC.

Anderson, J L & Barnett, M 2013 ‘Learning Physics with Digital Game Simulations in Middle School Science’, *Journal of Science Education and Technology,* vol. 22, no.6, pp. 914-926.

Andersson, C & Palm, T 2017, ‘The impact of formative assessment on student achievement: A study of the effects of changes to classroom practice after a comprehensive professional development programme’, *Learning and Instruction*, 49, pp. 92-102.

Andersson, C & Palm, T 2017a, ‘The characteristics of formative assessment that enhance student achievement in Mathematics’, *Education Inquiry*, vol. 8, no. 2, pp.104-122.

Andersson, C & Palm, T 2018, ‘Reasons for teachers’ successful development of a formative assessment practice through professional development – a motivation perspective’, *Assessment in Education: Principles, Policy & Practice*, vol. 25, no. 6, pp. 576-597.

Andersson, C, & Palm, T 2017, ‘Characteristics of improved formative assessment practice’, *Education Inquiry*, vol. 8, no. 2, pp. 104-122.

Andersson, N 2018, ‘Making space for assessment: dance teachers' experiences of learning and teaching prerequisites’, *Research in Dance Education*, vol. 19, no. 3, pp. 274-293.

Andrade, H & Cizek, G 2010, ‘*Handbook of Formative Assessment’*, Routledge, Florence.

Andrade, H, Hefferen, J & Palma, M 2014a, ‘Formative assessment in the visual arts’, *Art Education,* vol. 67, no. 1, pp. 34-40.

Andrade, H, Lui, A, Palma, M & Hefferen, J 2015, ‘Formative assessment in dance education’, *Journal of Dance Education*, vol. 15, no. 2, pp. 47-59.

Andrade, HL & Cizek, GJ 2010, *Handbook of formative assessment,* Routledge, New York, NY.

Ardoin, SP, Christ, TJ, Morena, LS, Cormier, DC & Klingbeil, DA 2013, ‘A systematic review and summarization of the recommendations and research surrounding Curriculum-Based Measurement of oral reading fluency (CBM-R) decision rules’, *Journal of School Psychology*, vol. 51, no.1, pp. 1-18.

Aschbacher, P & Alonzo, A 2004, ‘*Using Science notebooks to assess students' conceptual understanding.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA’.

Aschbacher, P & Alonzo, A 2006, ‘Examining the utility of elementary Science notebooks for formative assessment purposes’, *Educational Assessment, 11*, pp. 179-203.

Assessment Reform Group 1999, ‘*Assessment for learning: Beyond the black box*. Cambridge, UK: School of Education, Cambridge University’ Retrieved from http://www.nuffieldfoundation.org/sites/default/files/files/beyond_blackbox.pdf.

Axtell, PK, McCallum, RS, Bell, S & Poncy, B 2009, ‘Developing math automaticity using a class-wide fluency building procedure for middle school students: A preliminary study’, *Psychology in the Schools*, vol. 46, no. 6, pp. 526-538.

Aydemir, Z, Ozturk, E & Horzum, MB 2013, ‘The effect of reading from screen on the 5th Grade elementary students’ level of reading comprehension on informative and narrative type of texts’, *Educational Sciences: Theory & Practice*, vol. 13, no. 4, pp. 2272–2276.

Baker, SK, Nelson, NJ, Stoolmiller, M, Kennedy Paine, P, Turtura, J, Crone, D & Fien, H 2018, ‘Intervening with Struggling Readers in Seventh Grade: Impact Evidence from Six School Districts’, *Journal of Research on Educational Effectiveness*, vol. 11, no. 4, pp. 479-506.

Bartholomew, SR, Strimel, GJ & Yoshikawa, E 2019, ‘Using adaptive comparative judgment for student formative feedback and learning during a middle school design project’, *International Journal of Technology and Design Education,* vol. 29, no. 2, pp. 363-385.

Bartlett, E & McPhail, G 2016, ‘Teacher expertise and feedback in music composition’, *Pacific-Asian Education Journal,* vol. 28, no. 1, pp. 63-74.

Baten, E, Praet, M & Desoete, A 2017, ‘The relevance and efficacy of metacognition for instructional design in the domain of Mathematics’, *ZDM*, vol. 49, no. 4, pp. 613-623.

Bennett, RE & Gitomer, DH 2009, ‘Transforming K–12 assessment: Integrating accountability testing, formative assessment and professional support’, in *Educational assessment in the 21st century*, Springer, Dordrecht, pp. 43-61.

Bennett, RE 2011, 'Formative assessment: A critical review', *Assessment in Education: Principles, Policy & Practice*, vol. 18, no. 1, pp. 5-25.

Berry, R 2011, ‘Assessment Trends in Hong Kong: Seeking to Establish Formative Assessment in an Examination Culture’, *Assessment in Education: Principles, Policy & Practice,* vol*.* 18, no. 2, pp. 199-211.

Bhagat, KK & Spector, JM 2017, ‘Formative assessment in complex problem-solving domains: The emerging role of assessment technologies’, *Journal of Educational Technology & Society*, vol. 20, no. 4, pp. 312-317.

Birenbaum, M, Kimron, H & Shilton, H 2011, ‘Nested contexts that shape assessment ‘for’ learning: school-based professional learning community and classroom culture’, *Studies in Educational Evaluation, 37*, pp. 35-48.

Birenbaum, M, Kimron, H, Shilton, H & Shahaf-Barzilay, R 2009, ‘Cycles of inquiry: Formative assessment in service of learning in classrooms and in school-based professional communities’, *Studies in Educational Evaluation,* vol. 35*,* no .4, pp. 130-149.

Black, P, & Wiliam, D 1998, ‘Assessment and classroom learning’, *Assessment in Education: principles, policy & practice*, vol. 5, no. 1, pp. 7-74.

Black, P, & Wiliam, D 1998a, ‘Inside the black box: Raising standards through classroom assessment’, *Phi Delta Kappan*, vol. 80, no. 2, pp. 139-148.

Black, P, & Wiliam, D 2009,’ Developing the theory of formative assessment’, *Educational Assessment Evaluation and Accountability*, vol. 21, no. 1, pp. 5-31.

Bloom, BS, Hastings, JT & Madaus, GF (Eds.) 1971, *Handbook of formative and summative evaluation of student learning*, McGraw-Hill, New York.

Boakes, NJ 2009, ‘Origami instruction in the middle school Mathematics classroom: Its impact on spatial visualization and geometry knowledge of students’, *RMLE Online*, vol. 32, no. 7, pp. 1-12.

Bond, JB & Ellis, AK 2013, ‘The effects of metacognitive reflective assessment on fifth and sixth graders' Mathematics achievement’, *School Science and Mathematics,* vol. 113, no. 5, pp. 227-234.

Borko, H, Mayfield, V, Marion, S, Flexer, R & Cumbo, K 1997, ‘Teachers’ developing ideas and practices about Mathematics performance assessment: Successes, stumbling blocks, and implications for professional development,’ *CSE Technical Report 423*, National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles.

Box, C, Skoog, G & Dabbs, JM 2015, ‘A case study of teacher personal practice assessment theories and complexities of implementing formative assessment’, *American Educational Research Journal,* vol. 52, no .5, pp. 956-983.

Bramwell-Lalor, S & Rainford, M 2015, ‘Advanced Level Biology Teachers' Attitudes towards Assessment and Their Engagement in Assessment for Learning’, *European Journal of Science and Mathematics Education*, vol. 4, no. 3, pp. 380-396.

Briggs, DC, Ruiz-Primo, MA, Furtak, E, Shepard, L, & Yin, Y 2012, ‘Meta-analytic methodology and inferences about the efficacy of formative assessment’, *Educational Measurement: Issues and Practice*, vol. 31, no. 4, pp.13-17.

Brink, M & Bartz, DE 2017, ‘Effective Use of Formative Assessment by High School Teachers. Practical Assessment’, *Research & Evaluation,* vol. 22, no. 8, pp. 1-10.

Brookhart, SM 2010, *Formative assessment strategies for every classroom* (2nd edition), Alexandria, VA: ASCD.

Brookhart, SM, Moss, CM & Long, BA 2010, ‘Teacher inquiry into formative assessment practices in remedial reading classrooms’, *Assessment in Education: Principles, Policy and Practice*, vol. 17, no. 1, pp. 41-58.

Bryant, DP, Bryant, BR, Roberts, G, Vaughn, S, Pfannenstiel, KH, Porterfield & Gersten, R 2011, ‘Early numeracy intervention program for first-grade students with Mathematics difficulties’, *Exceptional children*, vol. 78, no. 1, pp. 7-23.

Burns, MK, Codding, RS, Boice, CH & Lukito, G 2010, ‘Meta-analysis of acquisition and fluency math interventions with instructional and frustration level skills: Evidence for a skill-by-treatment interaction’, *School Psychology Review*, vol. 39, no. 1, pp. 69.

Campbell, YC & Filimon, C 2018, ‘Supporting the argumentative writing of students in linguistically diverse classrooms: An action research study’, *RMLE Online*, vol. 41, no. 1, pp.1-10.

Campus, R 2014, *IRubric: Dance performance evaluation rubric* [Online forum comment]. Retrieved from: http://www.rcampus.com/rubricshowc.cfm?code.E8X3A9&sp.yes&.

Carlson, D, Borman, GD & Robinson, M 2011, ‘A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and Mathematics achievement’, *Educational Evaluation and Policy Analysis,* vol. 33, no. 3, pp. 378-398.

Cawsey, C, Hattie, J & Masters, G 2019, *Growth to Achievement: on-demand resources for teachers*, Australian Government Department of Education and Training.

Cech, SJ, 2008, ‘Test industry split over ‘formative’ assessment’, *Education Week*, vol. 28, no. 4, pp.1-15.

Chappell, S, Arnold, P, Nunnery, J & Grant, M 2015, ‘An Examination of an Online Tutoring Program's Impact on Low-Achieving Middle School Students' Mathematics Achievement’ *Online Learning*, vol. 19, no. 5, pp. 37-53.

Chase, CC & Klahr, D 2017, ‘Invention versus Direct Instruction: For Some Content, It's a Tie’, *Journal of Science Education and Technology*, vol. 26, no. 6, pp. 582-596.

Chen, F & Andrade, H 2018, ‘The impact of criteria-referenced formative assessment on fifth-grade students’ theatre arts achievement’, *The Journal of Educational Research,* vol. 111, no. 3, pp. 310-319.

Chen, F, Lui, AM, Andrade, H, Valle, C & Mir, H 2017, 'Criteria-referenced formative assessment in the Arts', *Educational Assessment, Evaluation and Accountability,* vol. 29, pp. 297-314.

Cizek, G. J 2010, An introduction to formative assessment: History, characteristics, and challenges, in HL Andrade & GJ Cizek (eds.), *Handbook of formative assessment,* Routledge, New York.

Cizek, GJ, Andrade, HL & Bennett, RE 2019, Formative assessment: history, definition, and progress, in Andrade, HL, Bennett, RE & Cizek, GJ (eds.), *Handbook of Formative Assessment in the Disciplines*, Routledge, New York.

Clark, I 2012, 'Formative assessment: Assessment is for self-regulated learning', *Educational Psychology Review,* vol. 24, no. 2, pp. 205-249.

Clarke, B, Baker, S, Smolkowski, K, Doabler, C, Strand Cary, M & Fien, H 2015, 'Investigating the efficacy of a core Kindergarten Mathematics curriculum to improve student Mathematics learning outcomes', *Journal of Research on Educational Effectiveness*, vol. 8, no. 3, pp. 303-324.

Clarke, B, Doabler, CT, Strand Cary, M, Kosty, D, Baker, S, Fien, H & Smolkowski, K 2014 'Preliminary Evaluation of a Tier 2 Mathematics Intervention for First-Grade Students: Using a Theory of Change to Guide Formative Evaluation Activities', *Grantee Submission*, vol. 43, no. 2, pp. 160-177.

Clarke, B, Smolkowski, K, Baker, SK, Fien, H, Doabler, CT & Chard, DJ 2011 'The impact of a comprehensive Tier I core Kindergarten program on the achievement of students at risk in Mathematics', *The Elementary School Journal,* vol. 111, no. 4, pp. 561-584.

Connor, CM & Morrison, FJ 2017, 'Child characteristics by instruction interactions, literacy, and implications for theory and practice, in Cain, K, Compton, D, & Parrila, R (eds.),' *Theories of reading development*, John Benjamins: Amsterdam.

Connor, CM, Morrison, FJ & Katch, EL 2004, 'Beyond the reading wars: The effect of classroom instruction by child interactions on early reading', *Scientific Studies of Reading,* no. 8, pp. 305-336.

Connor, CM, Morrison, FJ & Petrella, JN 2004, 'Effective reading comprehension instruction: Examining child by instruction interactions', *Journal of Educational Psychology,* no. 96, pp. 682-698.

Connor, CM, Morrison, FJ, Fishman, B, Crowe, EC, Al Otaiba, S & Schatschneider, C 2013, 'A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade', *Psychological Science,* no. 24, pp. 1408–1419.

Connor, CM, Morrison, FJ, Fishman, B, Giuliani, S, Luck, M, Underwood, P, & Schatschneider, C 2011, 'Classroom instruction, child X instruction interactions and the impact of differentiating student instruction on third graders' reading comprehension', *Reading Research Quarterly,* no. 46, pp. 189-221.

Connor, CM, Morrison, FJ, Fishman, BJ, Schatschneider, C & Underwood, P 2007, 'The Early Years: Algorithm-guided individualized reading instruction', *Science,* no. 315, pp. 464–465.

Connor, CM, Morrison, FJ, Schatschneider, C, Toste, J, Lundblom, EG, Crowe, E & Fishman, B 2011, 'Effective classroom instruction: Implications of child characteristic by instruction interactions on first graders' word reading achievement', *Journal for Research on Educational Effectiveness,* no. 4, pp. 173-207.

Connor, CM, Spencer, M, Day, SL, Giuliani, S, Ingebrand, SW, McLean, L & Morrison, FJ 2014, ‘Capturing the complexity: Content, type, and amount of instruction and quality of the classroom learning environment synergistically predict third graders’ vocabulary and reading comprehension outcomes’, *Journal of Educational Psychology,* no. 106, pp. 762-778.

Copland, M 2002, Leadership of inquiry: Building and sustaining capacity for school improvement in the Bay Area School Reform Collaborative, Center for Research on the Context of Teaching, Stanford University, Stanford, CA.

Corcoran, T, Shields, P & Zucker, A 1998, *SSIs and professional development for teachers,* SRI International, Menlo Park, CA.

Cordray, D, Pion, G, Brandt, C, Molefe, A & Toby, M 2012, *The impact of the Measures of Academic Progress (MAP) program on student reading achievement* (NCEE 2013-4000), National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education, Washington, DC.

Dandolo Partners 2013, DER mid-program review: Assessing progress of the DER and potential future directions. Final report, Department of Education. Canberra.

Darling-Hammond, L, Wei, RC, Andree, A, Richardson, N & Orphanos, S 2009, *Professional Learning in the Learning Profession: a status report on teacher development in the United States and abroad,* National Staff Development Council & School Redesign Network, Stanford University, Stanford, CA.

De Lisle, J 2015, ‘The promise and reality of formative assessment practice in a continuous assessment scheme: the case of Trinidad and Tobago’, *Assessment in Education: Principles, Policy & Practice,* vol. 22, no. 1, pp. 79-103.

Decristan, J, Hondrich, L, Büttner, G, Hertel, S, Klieme, E, Kunter, M, Lühken, A, Adl-Amini, K, Djakovic, Mannel, S, Naumann, A, & Hardy, I 2015, ‘Impact of Additional Guidance in Science Education on Primary Students’ Conceptual Understanding’, *The Journal of Educational Research*, vol. 108, no. 5, pp. 358-370.

Deneen, CC, Fulmer, GW, Brown, GTL, Tan, K, Leong, WS, Tay, HY 2019, ‘Value, practice and proficiency: Teachers’ complex relationship with assessment for learning’, *Teaching and Teacher Education,* no. 80, pp. 39-47.

Denis, JM 2018, ‘Assessment in music: A practitioner introduction to assessing students’, *Update: Applications of Research in Music Education,* vol. 36, no. 3, pp. 20-28.

Deno, SL, Mirkin, PK & Chiang, B 1982, ‘Identifying valid measures of reading’, *Exceptional Children,* no. 49, pp. 36-45.

Desimone, L 2009, ‘Improving impact studies of teachers' professional development: Toward better conceptualizations and measures’, *Educational Researcher*, vol.38, no. 3, pp. 181-199.

Dignath, C & Buttner, G 2008, ‘Components of fostering self-regulated learning among students. A meta- analysis on intervention studies at primary and secondary school level’, *Metacognition and Learning*, vol. 3, no. 3, pp. 231-264.

Duit, R & Treagust, D 2003, ‘Conceptual change: A powerful framework for improving Science teaching and learning’, *International Journal of Science Education*, vol. 25, no. 6, pp. 671-688.

Duschl, R & Gitomer, D 1997, ‘Strategies and challenges to changing the focus of assessment and instruction in Science classrooms’, *Educational Assessment,* vol. 4, no. 1, pp. 37–73.

Egan, TM, Cobb, B & Anastasia, M 2009, ‘Think time: Formative assessment empowers teachers to try new practices’, *Journal of Staff Development,* vol. 30, no. 4, pp. 40-42.

Espin, CA, Wayman, MM, Deno, SL, McMaster, KL & de Rooij, M 2017, ‘Databased decision making: Developing a method for capturing teachers' understanding of CBM graphs’, *Learning Disabilities Research & Practice,* no. 32, pp. 8-21.

European Commission 2011, ‘Evidence on the use of ICT for the assessment of key competences’, European Commission, Brussels, Belgium.

Faber, J, Luyten, JW & Visscher, AJ 2017, 'The effects of a digital formative assessment tool on Mathematics achievement and student motivation: Results of a randomized experiment', *Computers & education*, vol. 106, pp. 83-96.

Faber, J, and Visscher, AJ 2018, ‘The effects of a digital formative assessment tool on spelling achievement: Results of a randomized experiment’, *Computers & education*, vol. 122, pp.1-8.

Faggella-Luby, M, Griffith, RR, Silva, C & Weinburgh, MH, 2016, ‘Assessing ELLs' Reading Comprehension and Science Understandings Using Retellings’, *Electronic Journal of Science Education*, vol. 20, no. 3, pp.150-166.

Falk, A 2012, ‘Teachers Learning from Professional Development in Elementary Science: Reciprocal Relations between Formative Assessment and Pedagogical Content Knowledge’, *Science Education*, vol. 96, no. 2, pp. 265-290.

Fantuzzo, JW, Gadsden, VL & McDermott, PA 2011, ‘An integrated curriculum to improve Mathematics, language, and literacy for Head Start children’, *American Educational Research Journal,* vol. 48, no. 3, pp.763-793.

Feldman, A & Capobianco, BM 2008, ‘Teacher learning of technology enhanced formative assessment’ *Journal of Science Education and Technology,* no. 17, pp. 82-99.

Fennema, E, Franke, M, Carpenter, T & Carey, D 1993, ‘Using children’s mathematical knowledge in instruction’, *American Educational Research Journal*, vol. 30, no. 3, pp. 555-583.

Ferm Almqvist, C, Vinge, J, Vakeva, L & Zanden, O 2017, ‘Assessment ‘as’ learning in music education: The risk of ‘criteria compliance’ replacing ‘learning’ in the Scandinavian countries’, *Research Studies in Music Education,* vol. 39, no. 1, pp. 3-18.

Filsecker, M & Kerres, M 2012, ‘Positioning formative assessment from an educational assessment perspective: A response to Dunn & Mulvenon 2009’, *Practical Assessment, Research & Evaluation,* vol. 17, no.16, pp. 1-9.

Fletcher, A & Shaw, G 2012, ‘How does student-directed assessment affect learning? Using assessment as a learning process’, *International Journal of Multiple Research Approaches,*no. 6, pp. 245-263.

Forbes, CT, Sabel, JL & Biggers, M 2015, ‘Elementary Teachers' Use of Formative Assessment to Support Students' Learning about Interactions between the Hydrosphere and Geosphere’, *Journal of GeoScience Education,* vol. 63, no. 3, pp. 210-221.

Förster, N & Souvignier, E 2014, ‘Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept’, *Learning and Instruction*, no. 32, pp. 91–100.

Förster, N, Kawohl, E & Souvignier, E 2018, ‘Short- and long-term effects of assessment-based differentiated reading instruction in general education on reading fluency and reading comprehension’, *Learning and Instruction*, no. 56, pp. 98-109.

Fuchs, L, & Fuchs, D 1986, ‘Effects of systematic formative evaluation of student achievement: A meta-analysis’, *Exceptional Children*, vol. 53, no. 3, pp. 199-205.

Fuchs, LS & Vaughn, S 2012, ‘Responsiveness-to-Intervention: A decade later’, *Journal of Learning Disabilities,* no. 45, pp.195-203.

Fuchs, LS, Fuchs, D, Hamlett, CL & Ferguson, C 1992, ‘Effects of expert system consultation within curriculum-based measurement, using a reading maze task’, *Exceptional Children*, vol. 58, no. 5, pp. 436-450.

Fullan, MG 1993, Change Forces: probing the depths of educational reform, Falmer Press, Levittown, PA.

Gallagher, HA, Arshan, N & Woodworth, K, 2017, ‘Impact of the National Writing Project's College-Ready Writers Program in high-need rural districts, *Journal of Research on Educational Effectiveness*, vol. 10, no. 3, pp. 570-595.

Garet, M, Porter, A, Desimone, L, Birman, B & Yoon, K 2001, ‘What Makes Professional Development Effective? Analysis of a National Sample of Teachers’, *American Educational Research Journal*, no. 38, pp. 915-945.

Garet, MS, Birman, BF, Porter, AC, Desimore, L, Herman, R & Yoon, KS 1999, ‘*Designing effective professional development: Lessons from the Eisenhower Program’*, US Department of Education, Washington, DC.

Gersten, R, Chard, DJ, Jayanthi, M, Baker, SK, Morphy, P & Flojo, J 2009, ‘Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components’, *Review of Educational Research*, vol. 79, no.3, pp. 1202-1242.

Gersten, R, Dimino, J, Jayanthi, M, Kim, JS & Santoro, L 2010, ‘Teacher Study Group: The Impact of Reading Instruction and Student Outcomes in First Grade Classrooms’, *American Educational Research Journal, 47*, pp. 694–739.

Gewertz, C 2015, ‘Searching for clarity on formative assessment’, *Education Week*, vol. *35,* no. 12, p. S2.

Gikandi, JW, Morrow, D & Davis, NE 2011, ‘Online formative assessment in higher education: A review of the literature’, *Computers & education*, vol. 57, no. 4, pp. 2333-2351.

Goertz, M, Oláh, L & Riggan, M 2009, ‘*Can interim assessments by used for instructional change? CPRE Policy Briefs RB-51’,* Consortium for Policy Research in Education, Philadelphia.

Goertz, ME, Olah, LN & Riggan, M 2009a, ‘*From testing to teaching: The use of interim assessments in classroom instruction* (Research Report #65)’, University of Pennsylvania, Consortium for Policy Research in Education, Philadelphia.

Goh, K & Walker, R 2018, ‘Written teacher feedback: Reflections of year seven music students’, *Australian Journal of Teacher Education*, vol. 43, no. 12, pp. 30-41.

Gonski, D, Arcus, T, Boston, K, Gould, V, Johnson, W, O'Brien, L, Perry, LA & Roberts, M 2018, *Through growth to achievement: Report of the review to achieve educational excellence in Australian schools*, Commonwealth of Australia, Canberra.

Graham, S & Perin, D 2007, 'Writing Next: Effective strategies to improve writing of adolescents in middle and high schools' Carnegie Corporation of NewYork, New York.

Graham, S, Hebert, M & Harris, KR 2015, 'Formative assessment and writing: A meta-analysis', *The Elementary School Journal*, vol. 115, no. 4, pp. 523-547.

Haelermans, C, Ghysels, J & Prince, F 2015, 'A dataset of three educational technology experiments on differentiation, formative testing and feedback', *British Journal of Educational Technology*, vol. 46, no. 5, pp. 1102-1108.

Hall, TE, Cohen, N, Vue, G & Ganley, P 2015, 'Addressing learning disabilities with UDL and technology: Strategic reader', *Learning Disability Quarterly*, vol. 38, no. 2, pp. 72-83.

Hargreaves, A 2007, 'Five Flaws of Staff Development and the Future Beyond', *Journal of Staff Development*, vol. 28, no. 3, pp. 37-38.

Hargreaves, E 2013, 'Inquiring into children's experiences of teacher feedback: reconceptualising assessment for learning', *Oxford Review of Education,* 39, pp. 229-246.

Harris, LR & Brown, GTL 2013, 'Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: case studies into teachers' implementation', *Teaching and Teacher Education,* 36, pp. 101-111.

Hartmeyer, R, Stevenson, MP & Bentsen, P 2018, 'A systematic review of concept mapping-based formative assessment processes in primary and secondary Science education', *Assessment in Education: Principles, Policy & Practice,* vol. 25, no. 6, pp. 598-619.

Haug, BS & Odegaard, M 2015, 'Formative Assessment and Teachers' Sensitivity to Student Responses', *International Journal of Science Education*, vol. 37, no. 4, pp. 629-654.

Havnes, A, Smith, K, Dysthe, O & Ludvigsen, K 2012, 'Formative assessment and feedback: making learning visible', *Studies in Educational Evaluation,* 38, pp. 21-27.

Heitink, MC, Van der Kleij, FM, Veldkamp, BP, Schildkamp, K & Kippers, WB 2016, 'A systematic review of prerequisites for implementing assessment for learning in classroom practice', *Educational Research Review,* 17, pp. 50-62.

Hellrung, K & Hartig, J 2013, 'Understanding and using feedback–A review of empirical studies concerning feedback from external evaluations to teachers', *Educational Research Review,* 9, pp. 174-190.

Henderson, S, Petrosino, A, Guckenburg, S & Hamilton, S 2007, 'Measuring how benchmark assessments affect student achievement (Issues & Answers Report, REL2007 No. 039). US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands, Washington, DC.

Heritage, M 2010, Formative assessment: Making it happen in the classroom, Corwin Press: Thousand Oaks, CA.

Heritage, M 2013, Formative assessment in practice: A process of inquiry and action, Harvard Education Press, Cambridge, MA

Heritage, M, Kim, J, Vendlinski, T & Herman, J 2009, 'From evidence to action: A seamless process in formative assessment?', *Educational Measurement: Issues and Practice,* vol. 28, no. 3, pp. 24-31.

Herman, J, Osmundson, E, Dai, Y, Ringstaff, C & Timms, M 2015, 'Investigating the Dynamics of Formative Assessment: Relationships between Teacher Knowledge, Assessment Practice and Learning', *Assessment in Education,* vol. 22, no. 3, pp. 344-367.

Hopfenbeck, TN, Florez Petour, TM & Tolo, A 2015, 'Balancing Tensions in Educational Policy Reforms: Large-Scale Implementation of Assessment for Learning in Norway', *Assessment in Education: Principles, Policy & Practice,* vol. 22, no. 1, pp. 44-60.

Hsia, LH, Huang, I & Hwang, GJ 2016, 'Effects of different online peer-feedback approaches on students' performance skills, motivation and self-efficacy in a dance course', *Computers & Education*, 96, pp. 55-71.

Hsiao, HS, Lin, CY, Chen, JC & Peng, YF 2017, 'The influence of a Mathematics problem-solving training system on first-year middle school students', *Eurasia Journal of Mathematics, Science and Technology Education,* vol. 14, no. 1, pp. 77-93.

Huberman, AM & Miles, MB 1984, Innovation Up Close: 'How School Improvement Works', Plenum, New York.

Hudesman, J, Crosby, S, Ziehmke, N, Everson, H, Isaac, S, Flugman, B, Zimmerman, B & Moylan, A 2014, 'Using Formative Assessment and Self-Regulated Learning to Help Developmental Mathematics Students Achieve: A Multi-Campus Program', *Journal on Excellence in College Teaching,* vol. 25, no. 2, pp.107-130.

Impara, JC, Plake, BS & Fager, JJ 1993, 'Teachers' assessment background and attitudes towards testing', *Theory into Practice,* vol. 32, no. 2, pp. 113-117.

Irving, K.E, Pape, SJ, Owens, DT, Abrahamson, L, Silver, D, Sanalan, VA & Sanalan, V 2016, 'Classroom connectivity and algebra 1 achievement: A three-year longitudinal study', *Journal of Computers in Mathematics and Science Teaching*, vol. 35, no. 2, pp.131-151.

James, AO & Folorunso AM 2012, 'Effect of Feedback and Remediation on Students' Achievement in Junior Secondary School Mathematics', *International Education Studies,* vol. 5, no. 5, pp. 153-162.

Johnston, PH, Afferblach, P & Weiss, PB 1993, 'Teachers' assessment of the teaching and learning of literacy', *Educational Assessment,* vol. 1, no. 2, pp. 91-117.

Juntunen, ML 2017, 'National assessment meets teacher autonomy: national assessment of learning outcomes in music in Finnish basic education', *Music Education Research*, vol. 19, no. 1, pp. 1-16.

Kaware, SS & Sain, SK 2015, 'ICT application in education: an overview', *International Journal of Multidisciplinary Approach & Studies*, vol. 2, no. 1, pp. 25-32.

Kay, R & Knaack, L 2009, 'Exploring the use of audience response systems in secondary school Science classrooms', *Journal of Science Education and Technology,* 18, pp. 382-392.

Kazragytė, V & Kudinovienė, J 2018, ‘Formative assessment in arts education lessons: Episodic or integrated with effective teaching?’, *Pedagogika / Pedagogy,* vol. 131, no. 3, pp. 217-232.

Keuning, T, Van Geel, M & Visscher, A 2017, ‘Why a data-based decision-making intervention works in some schools and not in others’, *Learning Disabilities Research and Practice*, vol. 32, no. 1, pp. 32-45.

Kingsley, TL, Cassady, JC & Tancock, SM 2015, ‘Successfully Promoting 21st Century Online Research Skills: Interventions in 5th-Grade Classrooms’, *Reading Horizons*, vol. 54, no. 2, pp. 91-134.

Kingston, N & Broaddus, A 2017, ‘The use of learning map systems to support the formative assessment in Mathematics’, *Education Sciences*, vol. 7, no. 1, p. 41.

Kingston, N & Nash, B 2011, ‘Formative assessment: A meta-analysis and a call for research’, *Educational Measurement: Issues and Practice,* vol. 30, no. 4, pp. 28-37.

Kingston, N & Nash, B 2012, ‘How many formative assessment angels can dance on the head of a meta-analytic pin: 0.2.’, *Educational Measurement: Issues and Practice,* vol. 31, no. 4, pp.18-19.

Klein, PD & Rose, MA 2010, ‘Teaching argument and explanation to prepare junior students for writing to learn’, *Reading Research Quarterly*, vol. 45, no. 4, pp. 433-461.

Klinger, A, Volante, L & Deluca, C 2012, ‘Building teacher capacity within the evolving assessment culture in Canadian education’, *Policy Futures in Education*, vol. 10*,* no. 4, pp. 447-460.

Klute, M, Apthorp, H, Harlacher, J & Reale, M, 2017, *Formative Assessment and Elementary School Student Academic Achievement: A Review of the Evidence,* REL 2017-259. Regional Educational Laboratory Central.

Koedinger, KR, McLaughlin, EA, and Heffernan, NT 2010, ‘A quasi-experimental evaluation of an on-line formative assessment and tutoring system’, *Journal of Educational Computing Research*, vol. 43, no. 4, pp. 489-510.

Konstantopoulos, S, Miller, S & van der Ploeg, A 2013, ‘The impact of Indiana's system of interim assessments on Mathematics and reading achievement’, *Educational Evaluation and Policy Analysis*, 35, pp. 481–499.

Konstantopoulos, S, Miller, SR, van der Ploeg, A & Li, W 2016, ‘Effects of interim assessments on student achievement: Evidence from a large-scale experiment’, *Journal of Research on Educational Effectiveness*, vol. 9, no. 1, pp.188-208.

Lau, AMS 2016, ‘Formative good, summative bad?’A review of the dichotomy in assessment literature’, *Journal of Further and Higher Education,* vol. 40, no. 4, pp. 509-525.

‘Learning Audit Instrument to Stimulate Site-Based Professional Development, One School at a Time’, *Assessment in Education: Principles, Policy & Practice,* vol. 24, no. 2, pp. 271-289.

Lee, C & Wiliam, D 2005, ‘Studying changes in the practice of two teachers developing assessment for learning’, *Teacher Development*, vol. 9, no. 2, pp. 265-283.

Lee, H, Feldman, A & Beatty, ID 2012, ‘Factors that affect Science and Mathematics teachers' initial implementation of technology-enhanced formative assessment using a classroom response system’, *Journal of Science Education and Technology,* 21, pp. 523-539.

Lee, I 2011, ‘Bringing innovation to EFL writing through a focus on assessment for learning’, *Innovation in Language Learning and Teaching,* 5, pp. 19-33.

Lin, MC 2013, ‘The development of a performance assessment with performing arts teachers in Taiwan – from national policy to classroom practice’, *Research in Drama Education*, vol. 18, no. 3, pp. 296-312.

Lingam, GI & Lingam, N 2016, ‘Developing School Heads as Instructional Leaders in School-Based Assessment: Challenges and Next Steps’, Australian Journal of Teacher Education, vol. 41, no. 2, pp. 1-16.

Love, N & Crowell, M 2018, ‘Strong Teams, Strong Results: Formative Assessment Helps Teacher Teams Strengthen Equity’, *Learning Professional,* vol. 39, no. 5, pp. 34-39.

Luttenegger, KC 2009, ‘Formative Assessment Practices in Reading Instruction in Pre-Service Teachers' Elementary School Classrooms’, *Journal of Education for Teaching: International Research and Pedagogy*, vol. 35, no. 3, pp. 299-301.

Lyon. C & Leahy, S 2009, ‘Developing assessment for learning through teacher learning communities’, ETS RM-09-01, Education Testing Service, Princeton, NJ.

Lysaght, Z, and O’Leary, M 2017, Scaling up, writ small: using an assessment for learning audit instrument to stimulate site-based professional development, one school at a time’, *Assessment in Education: Principles, Policy & Practice*, vol. 24, no. 2, pp. 271-289.

Mackenzie, N, Scull, J & Bowles, T 2015, ‘Writing over time: An analysis of texts created by Year One students’, *AER*, vol. 42, no. 2, pp. 567-593.

Marzano, RJ & Kendall, JS 2007, ‘*The New Taxonomy of Educational Objectives’,* 2nd edn, Corwin Press, Thousand Oaks, CA.

Marzano, RJ 2010*, Formative assessment & standards-based grading,* Marzano Research, Bloomington, IN.

Mastrorilli, TM, Harnett, S & Zhu, J 2014, ‘Arts Achieve, impacting student success in the arts: Preliminary findings after one year of implementation’, *Journal for Learning through the Arts*, vol. 10, no. 1, pp. 1-24.

May, H & Robinson, MA 2007, ‘A randomized evaluation of Ohio’s Personalized Assessment Reporting System (PARS), Consortium for Policy Research in Education, Madison, WI.

McLaughlin, T & Yan, Z 2017, ‘Diverse delivery methods and strong psychological benefits: A review of online formative assessment’, *Journal of Computer Assisted Learning*, vol. 33, no. 6, pp. 562-574.

McMillan, JH & Nash, S 2000, ’*Teachers' classroom assessment and grading decision making,’* Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans, LA.

McMillan, JH, Venable, JC & Varier, D 2013, ‘Studies of the Effect of Formative Assessment on Student Achievement: So Much More is Needed’, *Practical Assessment, Research & Evaluation*, vol. 18, no. 2, pp. 1-5.

Means, B, Padilla, C, DeBarger, A & Bakia, M 2009, '*Implementing data-informed decision making in schools—teacher access, supports and use',* US Department of Education, Office of Planning, Evaluation, and Policy Development, Washington, DC.

Menesses, KF & Gresham, FM 2009, 'Relative Efficacy of Reciprocal and Nonreciprocal Peer Tutoring for Students At-Risk for Academic Failure', *School Psychology Quarterly,* vol. 24, no. 4, pp. 266-275.

Meusen-Beekman, KD, Brinke, D & Boshuizen, HP 2016, 'Effects of formative assessments to develop self-regulation among sixth grade students: Results from a randomized controlled intervention', *Studies in Educational Evaluation*, 51, pp.126-136.

Meyer, E, Abrami, PC, Wade, CA, Aslan, O & Deault, L 2010, 'Improving literacy and metacognition with electronic portfolios: Teaching and learning with ePEARL', *Computers and Education*, vol. 55, no. 1, pp. 84-91.

Miller, DM, Scott, CE & McTigue, EM 2018, 'Writing in the secondary-level disciplines: A systematic review of context, cognition, and content', *Educational Psychology Review,* vol. 30, no.1, pp. 83-120.

Mills, MM 2009, 'Capturing Student Progress via Portfolios in the Music Classroom', *Music Educators Journal,* vol. 96, no. 2, pp. 32-38.

Moher, D, Liberati, A, Tetzlaff, J, & Altman, DG 2009, 'Preferred reporting items for systematic reviews and meta-analysis: The PRISMA statement', *Annals of Internal Medicine*, vol. 151, pp. 264-270.

Morrison, FJ & Connor, CM 2002, 'Understanding schooling effects on early literacy', *Journal of School Psychology,* vol. 40, pp. 493-500.

Moss, C, Brookhart, S & Long, A 2013, 'Administrators' roles in helping teachers use formative assessment information', *Applied Measurement in Education*, vol. 26, no. 3, pp. 205-218.

Moss, CM & Brookhart, SM 2009, Advancing formative assessment in every classroom: A guide for instructional leaders, ASCD, Alexandria, VA.

Myhill, D, Jones, S & Wilson, A 2016, 'Writing conversations', *Research Papers in Education*, vol. 31*,* no. 1, pp. 23-44.

Newby, L, & Winterbottom, M 2011, 'Can research homework provide a vehicle for assessment for learning in Science lessons?', *Educational Review,* vol. 63, pp. 275-290.

Newman, R & Myhill, D 2016, 'Metatalk: Metalinguistic discussion about writing', *International Journal of Ed Research*, 80*,* pp.177-187.

Ní Chroinín, D & Cosgrave, C 2013, 'Implementing formative assessment in primary physical education: teacher perspectives and experiences', *Physical Education and Sport Pedagogy,* 18, pp. 219-233.

Noll, VH 1955, 'Requirements in educational measurement for prospective teachers', *School and Society,* 82, pp. 88–90.

Noyce, PE & Hickey, DT 2011, *New frontiers in formative assessment*, Harvard Education Press, Cambridge, MA.

Olakanmi, EE 2017, 'The effects of a flipped classroom model of instruction on students' performance and attitudes towards chemistry', *Journal of Science Education and Technology*, vol. 26, no. 1, pp. 127-137.

Organisation for Economic Co-operation and Development (OECD) 2015, *Students, computers and Learning: Making connection,* the PISA, OECD, Paris, France.

Osterman, KF & Kottkamp, RB 2004, '*Reflective Practice for Educators'*, 2nd edn, Corwin Press, Thousand Oaks, CA.

Panadero, E, Tapia, JA & Huertas, JA 2012, 'Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education', *Learning and individual differences,* vol. 22, no. 6, pp. 806-813.

Penuel, WR, Boscardin, CK, Masyn, K & Crawford, VM 2007, 'Teaching with student response systems in elementary and secondary education settings: A survey study', *Educational Technology Research and Development,* 55*,* pp. 315-346.

Penuel, WR, Fishman, BJ, Yamaguchi, R & Gallagher, LP 2007, 'What Makes Professional Development Effective? Strategies That Foster Curriculum Implementation', *American Educational Research Journal*, vol. 44, no. 4, pp. 921-958.

Phelan, JC, Choi, K, Niemi, D, Vendlinski, TP, Baker, EL & Herman, J 2012, 'The effects of POWERSOURCE© assessments on middle-school students' math performance', *Assessment in Education: Principles, Policy & Practice,* vol.19, no. 2, pp. 211-230.

Phelan, JC, Choi, K, Vendlinski, T, Baker, E & Herman, J 2011, 'Differential improvement in student understanding of mathematical principles following formative assessment intervention', *The Journal of Educational Research*, vol.104, no. 5, pp. 330-339.

Pinger, P, Rakoczy, K, Besser, M & Klieme, E 2018, 'Interplay of formative assessment and instructional quality—interactive effects on students' Mathematics achievement', *Learning Environments Research,* vol. 21, no. 1, pp. 61-79.

Polly, D, Wang, C, Martin, C, Lambert, R, Pugalee, D & Middleton, C 2018, 'The Influence of Mathematics Professional Development, School-Level, and Teacher-Level Variables on Primary Students' Mathematics Achievement', *Early Childhood Education Journal*, vol. 46, no. 1, pp.31-45.

Polly, D, Wang, C, Martin, C, Lambert, RG, Pugalee, DK & Middleton, CW 2017, 'The Influence of an Internet-Based Formative Assessment Tool on Primary Grades Students' Number Sense Achievement', *School Science and Mathematics*, vol. 117, no. 3-4, pp.127-136.

Ponce, HR, Mayer, RE, Figueroa, VA & López, MJ 2018, 'Interactive highlighting for just-in-time formative assessment during whole-class instruction: effects on vocabulary learning and reading comprehension', *Interactive Learning Environments*, vol. 26, no.1, pp. 42-60.

Popham, WJ, 2006, 'All about accountability/phony formative assessments: Buyer beware', *Educational Leadership*, vol. 64*,* no. 3, pp. 86-87.

Quinn, M, Miltenberger, R, James, T & Abreu, A 2016, 'An evaluation of auditory feedback for students of dance: Effects of giving and receiving feedback', *Behavioral Interventions,* vol. 32, no. 4, pp. 370-378.

Quint, J, Sepanik, S & Smith, J 2008, ‘Using student data to improve teaching and learning: Findings from an evaluation of the Formative Assessments of Students Thinking in Reading (FAST-R) program in Boston elementary schools’, MDRC, New York.

Rakoczy, K, Klieme, E, Bürgermeister, A & Harks, B 2008, ‘The interplay between student evaluation and instruction: grading and feedback in Mathematics classrooms’, *Journal of Psychology,* 216, pp. 111-124.

Rakoczy, K, Pinger, P, Hochweber, J, Klieme, E, Schütze, B & Besser, M 2019, ‘Formative assessment in Mathematics: Mediated by feedback's perceived usefulness and students' self-efficacy’, *Learning and Instruction*, 60, pp.154-165.

Randel, B, Apthorp, H, Beesley, A, Clark, T & Wang, X 2016, ‘Impacts of professional development in classroom assessment on teacher and student outcomes’, *The Journal of Educational Research,* vol. 109, no. 5, pp. 491-502.

Reddy, L, Dudek, C & Lekwa, A 2017, ‘Classroom Strategies Coaching Model: Integration of Formative Assessment and Instructional Coaching’, *Theory into Practice, Columbus,* vol. 56, no. 1, pp. 46-55.

Resendes, M, Scardamalia, M, Bereiter, C, Chen, B & Halewood, C 2015, ‘Group-level formative feedback and metadiscourse’, *International Journal of Computer-Supported Collaborative Learning*, vol.10, no. 3, pp. 309-336.

Robinson, J, Myran, S, Strauss, R & Reed, W 2014, ‘The impact of an alternative professional development model on teacher practices in formative assessment and student learning’, *Teacher Development*, vol. 18, no. 2, pp. 141–162.

Roschelle, J, Feng, M, Murphy, RF & Mason, CA 2016 ‘Online Mathematics homework increases student achievement’, *AERA Open,* vol. 2, no. 4, pp. 1-12.

Roschelle, J, Shechtman, N, Tatar, D, Hegedus, S, Hopkins, B, Empson, S, Gallagher, L 2010, ‘Integration of Technology, Curriculum, and Professional Development for Advancing Middle School Mathematics: Three Large-Scale Studies’, *American Educational Research Journal*, vol. 47, no. 4, pp. 833-878.

Roschelle, JM, Pea, RD, Hoadley, CM, Gordin, DN & Means, BM 2000, ‘Changing how and what children learn in school with computer-based technologies’, *The future of children*, vol. 10, no. 2, pp.76-101.

Rubie-Davies, CM & Rosenthal, R 2016, ‘Intervening in teachers' expectations: A random effects meta-analytic approach to examining the effectiveness of an intervention’, *Learning and Individual differences*, 50*,* pp. 83-92.

Sabel, JL, Forbes, CT & Flynn, L 2016, ‘Elementary teachers’ use of content knowledge to evaluate students’ thinking in the life Sciences’, *International Journal of Science Education,* vol. 38, no. 7, pp. 1077-1099.

Sach, E 2013, ‘An exploration of teachers' narratives: what are the facilitators and constraints which promote or inhibit ‘good’ formative assessment practices in schools?’ *Education, 3-13: International Journal of Primary, Elementary and Early Years Education*, 43, pp. 322-335.

Samo, DD, Darhim and Bana Kartasasmita 2017, ‘Culture-Based Contextual Learning to Increase Problem-Solving Ability of First Year University Student’, *Journal on Mathematics Education,* vol*.* 9, no.1, pp. 81-94.

Sanchez, CE, Atkinson, KM, Koenka, AC, Moshontz, H & Cooper, H 2017, ‘Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis', *Journal of Educational Psychology,* vol. 109, no. 8, pp. 1049-1066.

Sanzo, K, Myran, S & Caggiano, J 2014, *Formative Assessment Leadership*, Routledge, New York.

Schmid, D 2012, 'Data mining: A systems approach to formative assessment', *Journal of Dance Education*, vol.12, no. 3, pp. 75-81.

Schneider, MC & Meyer, JP 2012, 'Investigating the efficacy of a professional development program in formative classroom assessment in middle school English language arts and Mathematics', *Journal of MultiDisciplinary Evaluation*, vol. 8, no. 17, pp. 1-24.

Schneider, MC & Randel, B 2010, 'Research on Characteristics of Effective Professional Development Programs for Enhancing Educators' Skills in Formative Assessment,' in Andrade, HL & Cizek, GJ (eds.), *Handbook of Formative Assessment*, Routledge, New York.

Scott, SJ 2012, 'Rethinking the roles of assessment in music education', *Music Educators Journal*, vol. 98, no. 3, pp. 31-35.

Scriven, M. 1963, *The methodology of evaluation,* [Research Report #110], Purdue University, Lafayette.

Shechtman, N, Roschelle, J, Haertel, G & Knudsen, J 2010, 'Investigating Links from Teacher Knowledge, to Classroom Practice, to Student Learning in the Instructional System of the Middle-School Mathematics' Classroom', *Cognition and Instruction*, vol. 28, no. 3, pp. 317-359.

Shepard, L 2005, The future of assessment: Shaping teaching and learning, in *ETS Invitational Conference*, New York, NY.

Shuler, SC 2011, 'Music education for life: music assessment, Part 1: What and why', *Music Educators Journal,* vol. 98, no. 2, pp.10-13.

Sicherl Kafol, B, Kordeš, U & Holcar Brunauer, A 2017, 'Assessment for learning in music education in the Slovenian context–from punishment or reward to support,' *Music education research*, vol. 19, no. 1, pp.17-28.

Simmons, DC, Kim, M, Kwok, O, Coyne, MD, Simmons, LE, Oslund, E & Rawlinson, D 2015, 'Examining the Effects of Linking Student Performance and Progression in a Tier 2 Kindergarten Reading Intervention', *Journal of Learning Disabilities*, vol. 48, no. 3, pp. 255-270.

Slavin, RE, Cheung, A, Holmes, GC, Madden, NA & Chamberlain, A 2013, 'Effects of a data-driven district reform model on state assessment outcomes', *American Educational Research Journal*, 50, pp. 371-396.

Smit, R, Bachmann, P, Blum, V, Birri, T & Hess, K 2017, 'Effects of a rubric for mathematical reasoning on teaching and learning in primary school', *Instructional Science*, vol. 45, no. 1, pp. 603-622.

Soong, B, Mercer, N & Er, SS 2010, 'Revision by means of computer-mediated peer discussions', *Physics Education*, vol. 45, no. 3, p. 264.

Sparks, SD 2015, 'Types of assessments: A head-to-head comparison', *Education Week*, vol. 35, no.12, p. S2.

Speck, M & Knipe, C 2005, 'Why Can't We Get It Right? Designing High-Quality Professional Development for Standards-Based Schools,' 2nd edn, Corwin Press, Thousand Oaks, CA.

Staley, C 2015, ‘Tracking student growth through assessment, *Illinois Music Educator,* vol. 76, no. 1, p. 54.

Stecker, PM, Fuchs, LS & Fuchs, D 2005, ‘Using curriculum-based measurement to improve student achievement: Review of research’, *Psychology in the Schools*, vol. 42, no. 8, pp. 795-819.

Stiggins, R 1991, ‘Relevant classroom assessment training for teachers’, *Educational Measurement: Issues and Practice,* vol. 10, no. 1, pp. 7-12.

Sumantri, M S & R, Satriani 2016, ‘The Effect of Formative Testing and Self-Directed Learning on Mathematics Learning Outcomes’, *International Electronic Journal of Elementary Education,* vol. 8, no. 3, pp. 507-523.

Swapna, N, Prema, KS, Geetha, YV & Asha, SA 2017, ‘Development and validation of digital tutorial to facilitate pre-reading skill’, *JAIISH*, 36, pp. 36-47.

Terrazas-Arellanes, FE, Gallard M, AJ, Strycker, LA & Walden, ED 2018, ‘Impact of interactive online units on learning Science among students with learning disabilities and English learners’, *International Journal of Science Education*, vol. 40, no. 5, pp. 498-518.

Thoron, AC & Rubenstein, ED 2013, ‘The Effect of Vee Maps and Laboratory Reports on High- and Low-Order Content-Knowledge Achievement in AgriScience Education’, *Journal of Agricultural Education,* vol. 54, no. 3, pp. 198-208.

Tolo, A, Chan, J & Hopfenbeck, TN 2018, ‘A systematic review on teachers' implementation of technology-enhanced formative assessment: insights and implications’, *AERA Online Paper Repository*.

Torgerson, CJ 2007, ’The quality of systematic reviews of effectiveness in literacy learning in English: a ‘tertiary’ review’, *Journal of Research in Reading,* 30, pp. 287-315.

Torrance, H & Pryor, J 2001, ‘Developing formative assessment in the classroom: Using action research to explore and modify theory’, *British Educational Research Journal*, vol. 27, no. 5, pp. 615–631.

Torrance, H 2007, ‘Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning’, *Assessment in Education: Principles, Policy & Practice,* vol. 14, no. 3, pp. 281-294.

Tsuei, M 2017, ‘Learning behaviours of low-achieving children’s Mathematics learning in using of helping tools in a synchronous peer-tutoring system’, *Interactive Learning Environments*, vol. 25, no. 2, pp. 147-161.

Tsuei, M 2017, ‘Learning behaviours of low-achieving children’s Mathematics learning in using of helping tools in a synchronous peer-tutoring system’, *Interactive Learning Environments,* vol. 25, no.2, pp.147-161.

Tyler, EJ, Hughes, JC, Beverley, M & Hastings, RP 2015, ‘Improving early reading skills for beginning readers using an online programme as supplementary instruction’, *European Journal of Psychology of Education*, vol. 30, no. 3, pp. 281-294.

US National Mathematics Advisory Panel 2008, *Report of the task group on instructional practices,* National Academy Press, Washington.

Uzezi, G, 2017, 'The Effect of Learning Cycle Constructivist-Based Approach on Students' Academic Achievement and Attitude towards Chemistry in Secondary Schools in North-Eastern Part of Nigeria', *Educational Research and Reviews*, vol. 12, no. 7, pp. 456-466.

van den Berg, M, Bosker, R & Suhre, C 2018, 'Testing the effectiveness of classroom formative assessment in Dutch primary Mathematics education', *School Effectiveness and School Improvement,* vol. 29, no. 3, pp. 339-361.

van der Kleij, FM, Cumming, JJ & Looney, A 2018, 'Policy expectations and support for teacher formative assessment in Australian education reform', *Assessment in Education: Principles, Policy & Practice,* vol. 25, no. 6, pp. 620-637.

Van der Kleij, FM, Feskens, RC & Eggen, TJ 2015, 'Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis', *Review of educational research*, vol. 85, no. 4, pp. 475-511.

Veldkamp, BP, Matteucci, M & Eggen, TJ 2011, 'Computerized adaptive testing in computer assisted learning?' *Communications in Computer and Information Science*, 126, pp. 28-39.

Vogelzang, J & Admiraal, WF 2017, 'Classroom action research on formative assessment in a context-based chemistry course', *Educational Action Research*, vol. 25, no. 1, pp. 155-166.

Wang, AH, Firmender, JM, Power, JR & Byrnes, JP 2016, 'Understanding the program effectiveness of early Mathematics interventions for preKindergarten and Kindergarten environments: A meta-analytic review'*, Early Education and Development,* vol. 27, no.5, pp. 692-713.

Wang, TH 2010, 'Web-Based Dynamic Assessment: Taking Assessment as Teaching and Learning Strategy for Improving Students e-Learning Effectiveness', *Computers & Education,* vol. 54, no. 4, pp. 1157-1166.

Wayman, JC, Cho, V & Johnston, MT 2007, 'The data-informed district: A district-wide evaluation of data use in the Natrona County School District', University of Texas, Austin.

Wei, RC, Darling-Hammond, L, Andree, A, Richardson, N & Orphans, S 2009, 'Professional Learning in the Learning Profession:' A Status Report on Teacher 'Development in the United States and Abroad', National Staff Development Council.

Weinbaum, E 2009, 'Learning about assessment. An evaluation of a ten-state effort to build assessment capacity in high schools', Research report RR-61, Consortium for Policy Research in Education, Philadelphia.

Wheldall, K, Wheldall, R, Madelaine, A, Reynolds, M & Arakelian, S 2017, 'Further evidence for the efficacy of an evidence-based, small group, literacy intervention program for young struggling readers', *Australian Journal of Learning Difficulties*, vol. 22, no. 1, pp. 3-13.

Wiliam, D & Thompson, M 2008, 'Integrating assessment with learning: What will it take to make it work?' In Dwyer, CA (ed.) *The future of assessment*, Routledge, New York.

Wiliam, D 2007, *Five 'Key strategies' for effective formative assessment,* National Council of Teachers of Mathematics, Reston, VA.

Wiliam, D 2011, An integrative summary of the research literature and implications for a new theory of formative assessment, in Andrade, HL & Cizek, GJ (eds.), Handbook of formative assessment, Routledge, New York, NY.

Witmer, S, Duke, N, Billman, A & Betts, J 2014, ‘Using Assessment to Improve Early Elementary Students’ Knowledge and Skills for Comprehending Informational Text’, *Journal of Applied School Psychology,* vol. 30, no. 3, pp. 223-253.

Wong, H M 2017, ‘Implementing Self-Assessment in Singapore Primary Schools: Effects on Students' Perceptions of Self-Assessment’, *Pedagogies: An International Journal,* vol. 12, no. 4, pp. 391-409.

Wongwatkit, C, Srisawasdi, N, Hwang, GJ & Panjaburee, P 2017, ‘Influence of an integrated learning diagnosis and formative assessment-based personalized web learning approach on students learning performances and perceptions’, *Interactive Learning Environments*, vol. 25, no. 7, pp. 889-903.

Wu, HM, Kuo, BC & Wang, SC 2017, ‘Computerized Dynamic Adaptive Tests with Immediately Individualized Feedback for Primary School Mathematics Learning’, *Educational Technology & Society*, vol. 20, no.1, pp. 61-72.

Wylie, EC & Lyon, CJ 2009, ‘*How Much Is Enough: what is needed for a district to take on the formative assessment challenge?’* Paper presented at American Educational Research Association, San Diego, 13-17 April.

Yang, EF, Chang, B, Cheng, HN & Chan, TW 2015, ‘Improving pupils' mathematical communication abilities through computer-supported reciprocal peer tutoring’, *Journal of Educational Technology & Society*, vol. 19, no. 3, p. 157.

Yin, Y, Olson, J, Olson, M, Solvin, H & Brandon, PR 2015, ‘Comparing two versions of professional development for teachers using formative assessment in networked Mathematics classrooms’, *Journal of Research on Technology in Education,* vol. 47, no. 1, pp. 41-70.

Young, VM & Kim, DH 2010, ‘Using assessments for instructional improvement: A literature review’, *Education Policy Analysis Archives, 18,* pp.19. Retrieved June 30, 2019, from http://epaa.asu.edu/ojs/article/view/809

Young, VM 2006, ‘Teachers’ use of data: Loose coupling, agenda setting, and team norms’, *American Journal of Education,* vol. 112, no. 4, pp. 521-548.

Zeuch, N, Förster, N & Souvignier, E 2017, ‘Assessing teachers' competencies to read and interpret graphs from learning progress assessment: Results from tests and interviews’, *Learning Disabilities Research & Practice,* 32*,* pp. 61-70.

Zhang, B & Misiak, J 2015, ‘Evaluating three grading methods in middle school Science classrooms’, *Journal of Baltic Science Education,* vol. 14, no. 2, pp. 207-215.

Zucker, A, Kay, R & Staudt, C 2014, ‘Helping students make sense of graphs: an experimental trial of SmartGraphs software’, *Journal of Science Education and Technology*, vol. 23, no. 3, pp. 441-457.

11 Appendix – Studies reviewed in the report

11.1 The Arts

Authors	Online tool (Y/N)	Domain A = Arts M = Mathematics PD = Professional Development R = Reading S = Science W = Writing	Location of study	Sample characteristic P = Primary S = Secondary Typical sample vs atypical sample. If atypical then describe characteristics	Form of formative assessment (description of the task used) What is the source of the assessment tool used? Who is the author? (e.g. classroom teacher, school-based learning community, assessment expert working with teachers, ready-made package (standardised/non-standardised).	Impact of the formative assessment on student learning outcomes (cite measures of impact here)	What is being measured? L = Learning (meets L/O) PR = Progress G = Gaps S = Specific difficulties R = Reasons for difficulties (cognitively diagnostic/task diagnostic)	Who is the feedback to? L = learner T = teacher S = software Who's behaviour is expected to change as a result of this feedback?	Type of feedback to learner: NA = Not applicable S = Score/grade provided only SF = Score/grade & feedback re: correct answer SE = Explanation of the difference: correct results & explanation of differences between their result and the correct result; SEI = Explanation and improvement suggestions: As above but now students also receive some specific suggestions for improvement; SEA = Explanation and specific activities: Students are given information about the correct results, some explanation, and specific activities to undertake.	Type of feedback to teacher: NA = Not applicable S = Overall score only SS = Separate scores provided for specific aspects of performance I = Possible explanation of the problem areas and suggestions for additional instructional focus A = Possible explanation of the problem areas and specific instructional activities to undertake.	Evaluation is based on theoretically valid TASK model? Sequence of activities that need to be successfully completed to meet learning outcomes and how learners typically progress through them (learning progression)	The intervention is based on theoretically valid COGNITIVE model? Model of prerequisite cognitive and learning skills underlying successful progression. e.g. Does the process require a significant amount of working memory, attention, motivation, persistency, cognitive ability, language skills etc.	Are the actions/interventions following the assessment task evidence-based? (i.e. is the INSTRUCTIONAL model valid?)	What tools/resources are used in the assessment process and intervention? (Could be teacher designed or commercial).

* Chen & Andrade (2018)	N	A, PD	USA, New York	P	Criteria Referenced Formative Assessment (CRFA) – developed by experts as part of Arts Achieve project by NY Department of Education. Participating teachers received professional development focusing on effective use of criteria-referenced peer and self- assessment strategies. There is no information provided about the actual tasks undertaken as part of the Arts Achieve dataset that was used in this paper. No information was provided about type of feedback or who received this feedback except that there was peer and self-assessment/feedback.	Positive effect of the treatment condition on performance tasks (but not multiple choice or analysis tasks), Cohen's d = 0.25; statistically significant at p = .01.	PR	L, but unclear.	Not specified	Not specified	Varies depending on the individual teacher.	Not specified	Not specified	Benchmark Arts Assessments–Theatre Arts (BAATA) – used as pre- and post-tests. Developed in alignment with the New York City Department of Education Blueprints for Teaching and Learning in the Arts and the Common Core Capacities in English Language Arts. Includes multiple choice questions, constructed responses, and performance tasks. Implementation logs – filled out by teachers every 2-3 weeks to document teachers' use of treatment components.
* Chen, Lui, Andrade, Valle & Mir (2017)	N	A, PD	USA, New York	P, S	Criteria Referenced Formative Assessment (CRFA) – developed by experts as part of Arts Achieve project by NY Department of Education. Participating teachers received professional development focusing on effective use of criteria-referenced peer and self- assessment strategies. Only teacher's with sufficiently high self-reported treatment fidelity (based on implementation logs) were included in the study. Control group – business-as-usual instruction.	After matching treatment students to controls with similar scores on 12 demographic variables, control and treatment post-test scores were compared. The effect is significant (t (610) = 5.10, p = .00), with Cohen's d = .26 (95% CI = [.15, .37]).	Not specified	Not specified	Not specified	Not specified	Not specified	Not specified	Not specified	Benchmark Arts Assessments-Theatre Arts (BAATA) – used as pre- and post-tests. Developed in alignment with the New York City Department of Education Blueprints for Teaching and Learning in the Arts and the Common Core Capacities in English Language Arts. Includes multiple choice questions, constructed responses, and performance tasks. Implementation logs – filled out by teachers every 2–3 weeks to document teachers' use of treatment components.

* Mastrorilli, Harnett & Zhu (2014)	N	A	USA, New York	P, S	Criteria Referenced Formative Assessment (CRFA) – developed by experts as part of Arts Achieve project by NY Department of Education. Participating teachers received professional development focusing on effective use of criteria-referenced peer and self- assessment strategies. Control group – business-as-usual instruction.	The authors report a small effect of the intervention on student achievement (Glass's Delta = .18), after controlling for demographic and previous achievement differences. There was no significant difference in teachers' knowledge and instructional practice between treatment and control groups (p >.05), when years of experience and arts certification were controlled for.	Not specified	Not specified	Not specified	Not specified	Not specified	Not specified	Not specified	Benchmark Arts Assessments-Theatre Arts (BAATA) – used as pre- and post-tests. Developed in alignment with the New York City Department of Education Blueprints for Teaching and Learning in the Arts and the Common Core Capacities in English Language Arts. Includes multiple choice questions, constructed responses, and performance tasks. Teacher surveys – recorded demographic characteristics of the teachers.

11.2 Reading

Authors	Online tool (Y/N)	Domain A = Arts M = Mathematics PD = Professional Development R = Reading S = Science W = Writing	Location of study	Sample characteristic P = Primary S = Secondary Typical sample vs atypical sample. If atypical then describe characteristics	Form of formative assessment (description of the task used) What is the source of the assessment tool used? Who is the author? (e.g. classroom teacher, school-based learning community, assessment expert working with teachers, ready-made package (standardised/non-standardised).	Impact of the formative assessment on student learning outcomes (cite measures of impact here)	What is being measured? L = Learning (meets L/O) PR = Progress G = Gaps S = Specific difficulties R = Reasons for difficulties (cognitively diagnostic/task diagnostic)	Who is the feedback to? L = learner T = teacher S = software Who's behaviour is expected to change as a result of this feedback?	Type of feedback to learner: NA = Not applicable S = Score/grade provided only SF = Score/grade & feedback re: correct answer SE = Explanation of the difference: correct results & explanation of differences between their result and the correct result; SEI = Explanation and improvement suggestions: As above but now students also receive some specific suggestions for improvement; SEA = Explanation and specific activities: Students are given information about the correct results, some explanation, and specific activities to undertake.	Type of feedback to teacher: NA = Not applicable S = Overall score only SS = Separate scores provided for specific aspects of performance I = Possible explanation of the problem areas and suggestions for additional instructional focus A - Possible explanation of the problem areas and specific instructional activities to undertake.	Evaluation is based on theoretically valid TASK Model? Sequence of activities that need to be successfully completed to meet learning outcomes and how learners typically progress through them (learning progression)	The intervention is based on theoretically valid COGNITIVE Model? Model of prerequisite cognitive and learning skills underlying successful progression. e.g. Does the process require a significant amount of working memory, attention, motivation, persistency, cognitive ability, language skills etc.	Are the actions/interventions following the assessment task evidence-based? (i.e. is the INSTRUCTIONAL model valid?)	What tools/resources are used in the assessment process and intervention? (Could be teacher designed or commercial).

* Al Otaiba et al., 2011	Y	A	USA, Florida	P	ISI-K intervention – researcher-designed, a variant of ISI specific to Kindergarten. Includes multidimensional conceptualisation of reading instruction, student assessment and progress monitoring, A2i software, professional development for the teachers, and implementation in the classroom. Assessment to instruction (A2i) software – online tool for individualised reading instruction. The software allows to index existing literacy activities along three dimensions of instruction (code vs meaning-focused, teacher-vs child-managed, change over time), and provides recommendations on the amount and type of instruction based on student's scores.	Students in the treatment group had higher scores on word-reading measures than the controls, β = .33, p < .002, d = .52.	S	T	NA	I	Teachers in treatment and control groups both received a baseline professional development (1 day) on response to intervention and individualised instruction, however the content of the treatment group was specific to ISI-K and A2i implementation. Teachers in the treatment condition continued to receive professional development throughout the intervention, in the form of support resources, monthly school-level meetings, and fortnightly classroom-based support during literacy instruction. Fidelity of treatment was monitored by analysing videotapes of lessons.	Not specified	A2i provides recommendations about recommended changes in instruction.	Woodcock Johnson Tests of Achievement-III – was used to measure language and literacy skills (Picture Vocabulary, Letter Word Identification, and Word Attack subtests). AIMSWeb Letter Sound Fluency (Shinn & Shinn, 2004) – used to assess students' letter-sound correspondence. DIBELS Nonsense Word Fluency (NWF) and Phoneme Segmenting Fluency (PSF) tasks – district-administered test, used to assess students' ability to read letter sounds and blend them into words.
* Brookhart, Moss & Long 2010	N	R, PD	USA, mid-Atlantic state	P (K and Y1 at-risk readers)	Professional development program – taught different formative assessment practices, including letter cards, customised letter-naming drills, keeping records of feedback given to students, etc. Control condition – business as usual.	No effect on DIBELS Letter Naming Fluency in K students whose teachers underwent PD (partial η2 = .001); large effect on Y1 Phoneme Segmentation Fluency (partial η2 = .036)	L	T	NA	Not clear as different teachers used different assessments, their frequency of use not specified.	Specific sequence of activities not specified. Teachers underwent a PD program, but were not monitored for using specific formative assessments.	Yes	Teachers reported changes to instruction and differentiation, but specific actions following the assessment task are not specified in the paper.	Dynamic Indicators of Basic Early Literacy Skills (DIBELS) (Good and Kaminski 2002) – six individually administered standardised measures of early literacy development, this study administered phoneme segmentation fluency measure and letter naming fluency measure.
* Carlson et al., 2011	N	R, M	USA, multiple states	P, S	4Sight – quarterly benchmark assessments in reading and mathematics, aligned with state standards and supplemented by advice from consultants (John Hopkins Centre for Data-Driven Reform in Education intervention).	The treatment group had significantly higher post-test scores in mathematics as compared to control (p <.05), with the estimated increase of 0.06 student-level SDs in the treatment group. For reading, the difference was not significant (p >.05).	L	T	S	S	Not specified	Not specified	Not specified	State-administered achievement tests – school-level performance on mathematics and reading tests was analysed as the outcome variable.

*	Connor et al., 2007	Y	R	USA, Florida	P	Assessment to instruction (A2i) software – online tool for individualised reading instruction. The software allows to index existing literacy activities along three dimensions of instruction (code vs meaning-focused, teacher-vs child-managed, change over time), and provides recommendations on the amount and type of instruction based on student's scores.	Students in the treatment group showed more reading growth as compared to the control group, controlling for pre-test scores, as well as child and school demographics, with a residual mean difference of 2.63 points (95% CI = 0.37, 4.90). There was also an interaction with usage time, such that there was a 1 point increase in post-test score for every 50 additional minutes of A2i usage, t(20) = 2.97, p = .008.	S	T	NA	I	Teachers were given professional development on how to use the software. They were asked to teach reading for at least 90 minutes/day, to provide instruction to children with similar reading skills in small groups, and to follow the recommendations of A2i in regard to amounts and specific types of instruction. Control teachers were also expected to have a dedicated daily reading block (of 90 minutes) and to use small groups according to school policies.	Not specified	A2i provides recommendations about recommended changes in instruction. Based on classroom observation, only ~40% of teachers implemented the intervention with moderate to high fidelity.	Woodcock Johnson Tests of Achievement-III – was used to measure language and literacy skills.
*	Connor et al., 2011	Y	R	USA, Florida	P	ISI intervention – researcher-designed, includes multidimensional conceptualisation of reading instruction, student assessment and progress monitoring, A2i software, professional development for the teachers, and implementation in the classroom. Assessment to instruction (A2i) software – online tool for individualised reading instruction. The software allows to index existing literacy activities along three dimensions of instruction (code vs meaning-focused, teacher-vs child-managed, change over time), and provides recommendations on the amount and type of instruction based on student's scores.	Students in the treatment group showed significantly greater gains in word reading scores (β = 7.84, p < .021, d = .50).	S	T	NA	I	Teachers received professional development throughout the intervention, in the form of support resources, monthly school-level meetings, and fortnightly classroom-based support during literacy instruction. A2i usage was monitored throughout the year and teachers were encouraged to use it if they were found not to do so.	Not specified	A2i provides recommendations about recommended changes in instruction.	Woodcock Johnson Tests of Achievement-III – was used to measure language and literacy skills.

* Connor et al., 2013	Y	R	USA, Florida	P, 45% of students from families living in poverty	ISI intervention – researcher-designed, includes multidimensional conceptualisation of reading instruction, student assessment and progress monitoring, A2i software, professional development for the teachers, and implementation in the classroom. Assessment to instruction (A2i) software – online tool for individualised reading instruction. The software allows to index existing literacy activities along three dimensions of instruction (code vs meaning-focused, teacher- vs child-managed, change over time), and provides recommendations on the amount and type of instruction based on student's scores.	Students in the treatment condition had significantly higher reading scores than control in the first grade (d = .32), second grade (d = .44), and third grade (d = .25). Students who stayed in the treatment condition for more years showed greater gains in reading (d = .20 per year), however it was more beneficial to be in the treatment class in first grade as opposed to second or third.	S	T	NA	I	Teachers received professional development throughout the intervention, in the form of support resources, monthly school-level meetings, fortnightly classroom-based support during literacy instruction, and individual support as needed. Teachers in the control condition were given the same amount of professional development, but in mathematics (learned to apply Math Pals, Fuchs et al., 1997).	Not specified	A2i provides recommendations about recommended changes in instruction.	Woodcock Johnson Tests of Achievement-III – Letter-Word Identification and Passage Comprehension tests were used to measure language and literacy skills.
* Cordray et al., 2012	Y	R	Finland	P	Measures of Academic Progress (MAP; https://www.nwea.org/the-map-suite/) – a collection of dynamic tests in reading, language usage, mathematics, and science that place individual students on a continuum of learning from grade 3 to grade 10 in each discipline. Schools and teachers can use MAP to monitor student progress towards state proficiency standards. In this study, reading and language usage tests for Grades 4 and 5 were evaluated.	In Grade 4, there were no significant differences between treatment and control groups on either the ISAT reading score (p = .412), or the MAP composite score (p = .280). In Grade 5, there were also no significant differences on either ISAT (p = .280) or MAP (p = .889).	L	T	NA	SS	All teachers received standard professional development on how to administer and interpret MAP, as well as use MAP data to set student growth goals and evaluate instruction practices.	Not specified	Not specified	Illinois Standards Achievement Test (ISAT) – the reading scale scores were used as the outcome measure. MAP assessments in reading and language usage – administered to both treatment and control students as a composite measure to assess students' reading and literacy achievement.

* Förster & Souvignier 2014	Y	R	Germany	P	Progress monitoring assessment – 8 internet-based reading tests, one every 3 weeks. Includes a maze task and a comprehension task. Developed by the authors. LPA group – only received feedback on their progress every three weeks. LPA-G group – set goals for each assessment and then compare their performance to the goals Control group – no progress monitoring assessment, classroom activities not specified.	LPA group did significantly better than the control (z = 2.43, p = .015, d = 0.24), and also better than LPA-G group (z = -4.23, p < .001, d = -0.27).	L	T, L	S (LPA group) S/ element of SEI (LPA-G group)	S	Activities outside the formative assessments not specified.	Yes	Yes, but teachers carried out their own sequences of activities.	HAMLET 3-4 (Lehmann, Peek, & Poerschke, 2006) – standardised measure of reading comprehension.3 (out of 10) texts were used as pre-test, and another 3 as post-test. Questionnaires of reading motivation and reading self concept (Möller & Bonerad, 2007, Schöne, Dickhäuser, Spinath, & Stiensmeier-Pelster, 2012).
* Förster, Kawohl & Souvignier 2018	Y	R	Germany	P	Progress monitoring assessment – 8 internet-based reading tests, one every 3 weeks. Includes a maze task and a comprehension task. Developed by the authors. Teachers were also given support materials to aid differentiated instruction based on assessment results. The methods included repeated reading and peer-assisted learning.	Students in the treatment condition exhibited significantly higher rate of learning growth in reading fluency than the controls, γ01 = 2.28; z = 2.95; p < .01, d = .30 (1 year), d = .31 (2 years). There was no effect on reading comprehension, for both short (γ01 = −0.30; z = −0.68; p = .50) and long-term (γ01 = −0.35; z = −0.45; p = .65).	L	T, L	Score, what activity to work on	Score/progress - suggestion whether a student should focus on fluency or comprehension	Yes	Yes	Yes - repeated reading and peer-assisted learning.	Salzburger Reading Screening (SLS 1-4, Mayringer & Wimmer, 2003) – standardized screening measure for basic reading, mainly assessing reading speed. Students rate within 3 minutes whether 70 short sentences are correct. HAMLET 3-4 (Lehmann, Peek, & Poerschke, 2006) – standardised measure of reading comprehension. 3 (out of 10) texts were used as pre-test, and another 3 as post-test. Teacher questionnaire – researcher-developed, collected some self-report data on treatment fidelity.

* Hall et al., 2015	Y	R	USA, North-east	P, S – typical and with learning disabilities.	Strategic Reader – researcher-developed online tool with multiple features to support reading instruction, including progress monitoring. Curriculum-based Measurement (CBM) – progress monitoring tool build into Strategic Reader. Students are assessed on their oral reading fluency, reading comprehension, and reading comprehension strategies.	There were significant differences between pre-test and post-test scores for both online and offline conditions. The difference was not significant for students with learning disabilities in the offline treatment condition.	S	L, T.	S	SS	Both groups worked with Strategic Reader with the same scaffolding and support. The study compared the effects of using an online CBM integrated into the Strategic Reader, to the effects of doing CBM offline (pen and paper). Over 12 weeks, all students read at least 2 out of 4 available novels, responding to embedded reciprocal teaching prompts. Oral fluency and reading comprehension CBM measures were administered every two weeks, and the reciprocal teaching measure was administered before and after reading each novel. All teachers received 2 days of professional development.	Not specified	Not specified	Gates-MacGinitie Reading Test – used as pre-test and post-test of reading comprehension.
* Konstantopoulos et al., 2016	Y	R	USA, Indiana	P, S	Wireless Generation's mCLASS – commercial product for K-2; has literacy and numeracy components, both providing teachers with detailed feedback on students' error patterns and reading/problem-solving strategies. Acuity (CTB/McGraw-Hill) – commercial product for Grade 3-8; designed to forecast performance on the Indiana state test (ISTEP) through short assessments (30-35 multiple choice questions). Both assessment types provide performance reports against Indiana standards with individual and group summaries.	The authors conducted intention to treat, treatment on treated, and instrumental variables approaches to analysing the data. In all cases, there were significant ($p < .05$) negative differences between experimental and control groups on both Mathematics and Reading measures, but only in Grades K-2 (i.e. students using mCLASS performed worse than the controls). Treatment effect estimates ranged from -0.194 to -0.221. The differences between groups for K-8 and for 3-8 were not significant.	L	T By adding meaningful detail to teachers' awareness of students' current performance relative to prior performance, teachers' instruction would closely match student needs and current and intended knowledge gaps would be reduced.	NA	SS	Not specified	Not specified	The authors note that they did not provide any professional development to teachers on using the interim assessments or with specific suggestions for differentiated instruction.	Indiana state test (ISTEPC) – used as the outcome measure for mathematics and reading in Grades 3-8. Terra Nova – standardised achievement test (CTB McGraw Hill) , used as the outcome measure for mathematics and reading in Grades K-2.

*	Konstantopoulos, Miller, & van der Ploeg (2013)	Y	M, R	USA	P	Wireless Generation's mCLASS – commercial product for K-2; has literacy and numeracy components, both providing teachers with detailed feedback on students' error patterns and reading/problem-solving strategies. Acuity (CTB/McGraw-Hill) – commercial product for Grade 3-8; designed to forecast performance on the Indiana state test (ISTEP) through short assessments (30-35 multiple choice questions). Both assessment types provide performance reports against Indiana standards with individual and group summaries.	The treatment effect was significant for K-8 Mathematics (Estimate = .187, SE = .70, p < .05), but not Reading (Estimate = .098, SE = .055) in the treatment on treated analysis, but not significant for either Mathematics or Reading in the intention to treat analysis. The significance varied within more narrow age brackets, as well as between urban and rural schools. The authors conclude that the overall treatment effect was positive.	L	T. By adding meaningful detail to teachers' awareness of students' current performance relative to prior performance, teachers' instruction would closely match student needs and current and intended knowledge gaps would be reduced.	NA	SS	Not specified	Not specified	Not specified. The study only compares differences in achievement between schools in the different conditions, without controlling for actual usage of mCLASS and Acuity tools or specific instruction methods.	ISTEP+ – Indiana state test for reading and mathematics (Grades 3 - 8). Terra Nova – standardised test for mathematics and reading (Grades K - 2).

*	Simmons et al. 2015	N	R	USA, Connecticut, Florida, Texas	At-risk K students	Early Reading Intervention (ERI) – small-group intervention program that explicitly teaches phonologic, alphabetic, decoding, spelling, and sentence-reading skills to kindergarten students at early reading risk. Corrective feedback is provided to students as part of the process. Designed by researchers. Treatment condition received an adjusted form of ERI (ERI-A) with additional assessment points that allowed to adjust instruction in the second half of each unit based on students' achievement. Control condition received standard ERI.	ERI-A students in general outperformed their matched ERI counterparts with substantively important effects on all measures (δw 0.36-1.25). There was little to no effect of assessment and differentiation for slow and middle-scoring learners, with a larger impact of acceleration on faster learners.	L	L (embedded in the curriculum), T (ERI-A).	Corrective feedback embedded in the instruction process, not further specified.	SS If unit content was considered not learned, students repeat it, if learned very well, they can skip review.	The program includes 126 lessons structured in four units that progress from early phonemic and alphabetic skills to more complex regular and irregular word reading, spelling, and multiple sentence-reading skills. A typical 30-min lesson consists of seven activities, each designed to last 3 to 5 min and to actively engage students. Lessons provide explicit scripting for introducing, reviewing, and providing corrective feedback. New content is taught for 3 days and systematically reviewed. Students practice the new skill with the teacher and then apply it to discrimination or generalization tasks. The program includes four assessments, one at the end of each unit, and an instructional pacing chart.	Yes	In ERI-A condition, students who performed well on the additional assessment skipped the review stage and proceeded with typical lesson progression, meanwhile other students received additional instruction for the second half of the unit.	Peabody Picture Vocabulary Test–III (PPVT-III; Dunn & Dunn, 1997) – used as a measure of receptive vocabulary at pre-test Letter ID subtest from the WRMT-R/NU and LNF from DIBELS (Good & Kaminski, 2002) – measures of alphabetic knowledge at pre-test Blending Words (BW) and SM subtests from the CTOPP and Phoneme Segmentation Fluency (PSF) from DIBELS – post-test measures of phonological awareness. Supplementary Letter Checklist (SLC) subtest from the WRMT-R/NU – post-test measure of letter knowledge. Nonsense Word Fluency (NWF) subtest from DIBELS and the Word Attack (WA) subtest from the WRMT-R/NU – post-test measures of decoding. Word Identification (WI) subtest from the WRMT-R/NU – post-test measure of word reading. ‘Mac Gets Well’ (Makar, 1995) – post-test measure of reading fluency. Test of Written Spelling–4 (Larsen, Hammill, & Moats, 2005) – post-test measure of spelling.

* Slavin et al., 2013	N	R, M	USA, multiple states	P, S	4Sight – quarterly benchmark assessments in reading and mathematics, aligned with state standards and supplemented by advice from consultants to the school districts (John Hopkins Centre for Data-Driven Reform in Education intervention).	Being in the treatment group became a significant predictor of Grade 5 reading outcomes in the 3rd and 4th year after schools were assigned to the treatment condition (effect sizes 0.24 and 0.49 respectively). Mathematics followed the same pattern, with effect sizes 0.24 in 3rd year and 0.33 in 4th year. For Grade 8 reading outcomes, being in the treatment group was a significant predictor in the first two years (effect sizes 0.26 and 0.23 respectively), but not after that. For mathematics, Grade 8 outcomes were predicted by being in the treatment group in the 1st year after implementation (effect size = 0.17) and 4th year (effect size = 0.31).	L	T	S	S	Not specified	Not specified	Not specified	State-administered achievement tests – school-level performance on mathematics and reading tests was analysed as the outcome variable.
* Witmer et al. 2014	N	R, PD	USA	P	Concepts of Comprehension Assessment (COCA, Billman et al., 2008) – individually administered test designed to measure first- and second-grade students' specific fundamental knowledge and skills for comprehending informational text and is intended to help inform instruction.	The treatment group had significantly higher COCA scores at half-point (F(1, 120) = 17.14, p < .025, partial η2 = 0.13), and at the end of the year (F(1, 120) = 16.68, p < .025, partial η2 = 0.12). The treatment group also had significantly higher scores on the transfer writing measure at the end of the year, F(1, 108) = 9.25, p < .01, partial η2 = .08.	PR	T	NA	SS	Yes	Not specified	Not specified, as individual teachers could change instruction as desired. Self-report measures indicate that many of the participating teachers changed their writing activities and instruction as a result of their experiences with COCA.	PD for teachers on how to use and interpret COCA results (14.5 hours), experimenter designed. Prompted writing samples – students independently write for 30 minutes about the topic covered by their form of COCA assessment. Scored with a rubric. The main outcome measure is change in COCA assessment scores throughout the year.

11.3 Writing

Authors	Online tool (Y/N)	Domain A = Arts M = Mathematics PD = Professional Development R = Reading S = Science W = Writing	Location of study	Sample characteristic P = Primary S = Secondary Typical sample vs atypical sample. If atypical then describe characteristics	Form of formative assessment (description of the task used) What is the source of the assessment tool used? Who is the author? (e.g. classroom teacher, school-based learning community, assessment expert working with teachers, ready-made package (standardised/non-standardised).	Impact of the formative assessment on student learning outcomes (cite measures of impact here)	What is being measured? L = Learning (meets L/O) PR = Progress G = Gaps S = Specific difficulties R = Reasons for difficulties (cognitively diagnostic/task diagnostic)	Who is the feedback to? L = learner T = teacher S = software Who's behaviour is expected to change as a result of this feedback?	Type of feedback to learner: NA = Not applicable S = Score/grade provided only SF = Score/grade & feedback re: correct answer SE = Explanation of the difference: correct results & explanation of differences between their result and the correct result; SEI = Explanation and improvement suggestions: As above but now students also receive some specific suggestions for improvement; SEA = Explanation and specific activities: Students are given information about the correct results, some explanation, and specific activities to undertake.	Type of feedback to teacher: NA = Not applicable S = Overall score only SS = Separate scores provided for specific aspects of performance I = Possible explanation of the problem areas and suggestions for additional instructional focus A - Possible explanation of the problem areas and specific instructional activities to undertake.	Evaluation is based on theoretically valid TASK Model? Sequence of activities that need to be successfully completed to meet learning outcomes and how learners typically progress through them (learning progression)	The intervention is based on theoretically valid COGNITIVE Model? Model of prerequisite cognitive and learning skills underlying successful progression. e.g. Does the process require a significant amount of working memory, attention, motivation, persistency, cognitive ability, language skills etc.	Are the actions/interventions following the assessment task evidence-based? (i.e. is the INSTRUCTIONAL model valid?)	What tools/resources are used in the assessment process and intervention? (Could be teacher designed or commercial).
* Campbell & Filimon (2018)	N	A	USA, South Florida	S, middle school students with linguistically diverse backgrounds	Strategy-focused writing instruction – each student was provided with a writing folder with resources to aid learning (examples of past work, metacognitive strategy tool, etc.). As the students worked through the activities in the folder, teachers provided instruction, guided feedback and peer feedback.	There was a statistically significant increase in students' evidence and elaboration scores from pre-test to post-test, t(46) = -3.14, p = .003. There was also a significant increase in students' conventions of Standard English scores from pre-test to post-test, t(46) = -3.38, p = .002.	S	L	SEA	NA	Yes – paper presents a sequence of activities requiring students to engage in peer review processes and the improvement of metacognitive skills. Students received the strategy-focused writing instruction five days per week for 40 minutes during a 16-week period.	Not specified	Feedback from teacher and peers is embedded in the instruction process.	ELA-TBWRA (Florida Department of Education, 2014) – a test aligned to the FSA Standards, assesses various aspects of writing skills (organisation, evidence and elaboration, etc). Used a pre-test and post-test.

* Faber, J.M., & Visscher A.J. (2018)	Y	W	Netherlands	P	Snappet – a digital formative assessment tool. Students completed spelling tasks and receive immediate feedback on each item. Students can also view their performance grouped by specific learning goals. Assignment difficulty can be matched to performance level. Teachers can monitor overall progress of each student, whether they provide the correct response immediately or on second attempt, as well as normative feedback for each student or the entire group.	At the end of the intervention period (6 months) the treatment group did not have a significantly different spelling achievement than the control (β = .09, SE = .08, p > .05).	L, PR	L, T.	SF	SS	The study does not specify a specific sequence of activities. Authors report that most teachers prioritised completing curriculum assignments first, followed by adaptive and learning goal assignments.	Not specified	Snappet was embedded into the usual instruction methods used by the teachers.	Cito standardized spelling and mathematics tests – developed by the Dutch national institute for test development. Used as pre- and post-test to measure spelling achievement. Student survey – measured student motivation, developed by researchers. Snappet log files – used to measure Snappet usage.
* Fletcher, & Shaw (2012)	N	W	Australia, Darwin	P	Student-directed assessment planning template (SDA) – a learning process that draws on formative assessment principles. Students monitor their own progress and both teachers and students are involved in setting goals and criteria during the assessment process. Comparison group (TDA) completed the same activities but was entirely directed by the teacher and no learning goals were set by the students.	Among Year 4 students, the only significant difference between treatment and control groups were in the 'ideas' marking category, χ2 (5) = 11.01, p < 0.05, with Cramer's V = 0.51 (moderate effect size). Among Year 6 students, there were significant differences between groups on 6 out of 10 NAPLAN criteria (audience, text structure, vocabulary, paragraph, sentence structure, and spelling). The effect sizes (Cramer's V) ranged from 0.27 to 0.56 for these criteria.	L, R (maybe)	L	SEI (unclear)	S	Yes – the translated curriculum outcomes were a key part of the SDA planning template, designed to guide students in the writing project.	Yes – the focus of the study is on students' engagement by applying a constructive and self-regulated learning approach.	Formative assessment elements embedded into a larger intervention.	NAPLAN Writing section – student's NAPLAN scores from the previous year (except Year 2 students who did not complete any and so were excluded from quantitative analyses) were used as pre-test. Writing samples produced by students during the study were used as post-test, and were marked by a trained NAPLAN marking panellist, using the NAPLAN rubric.

* Gallagher, H.A, Arshan, N., &Woodworth, K.	Y	W, PD	USA, 10 states – rural districts	S	Using Sources Tool – an online portal that allows teachers to look at and assess student work, guiding teachers with a series of prompts to analyse student writing. The tool is also designed to further educate teachers by helping them identify qualities of effective arguments in their students' writing. Teachers also underwent an extensive professional development intervention over two years of this study (College-Ready Writers Program, CRWP).	CRWP group had significantly higher scores on 3 out of 4 AWC-SBA attributes (content, structure, stance). The impact estimate on the content and structure attributes is reported as 0.2 with p < .05 and for the stance attribute the impact estimate is reported as 0.15, p < 0.05.	L, G.	T	NA	I	Yes – the stated goal was to increase student writing proficiency in alignment with the new college- and career-ready standards in English language arts, and mathematics with use of supporting curricular resources.	Not specified	Formative assessment for the teachers embedded in the larger professional development intervention.	Analytic Writing Continuum for Source-Based Argument (AWC-SBA) – a measure to evaluate student writing, developed by the National Writing Project. Professional development monitoring form – a log to document professional development events and teacher participation. Teacher log and survey – administered in spring and autumn every day for 2 weeks. Teachers recorded time spent writing, length of writing assigned, and the purposes for writing that day. The survey measured broader practices and constructs more appropriately measured over a year than in a single day.
* Klein & Rose, 2010	N	W	Canada, London	P	The focus of the study was an intervention designed to teach writing as means of learning. Both experimental and control group received similar instruction with elements of formative assessment (peer and teacher feedback, opportunities to revise drafts). However, the experimental group more frequently wrote in content area subjects and were instructed in explanation writing. The control group also used rubrics to assess their writing.	Overall there was a significant difference on post-test scores between experimental and control groups, Pillai's trace = 0.99, F(6, 26) = 278.46, p < .001, η2 = 0.99. Students in the experimental group scored significantly higher than controls on four out of six post-test measures: argument genre knowledge (partial η2 = .21), explanation genre knowledge (partial η2 = .12), explanation test quality (partial η2 = .14), and post-test science knowledge (partial η2 = .21). The difference between groups were not significant for argument text quality and approach to writing.	L	L	Not specified	NA	The classroom teacher participant in 3 days of professional development to learn about writing to learn, analytic text genres, cognitive strategy instruction, and other elements of the experimental framework. The intervention was conducted over the course of half a year (October-March) and the post-test took place in June.	Not specified	Elements of formative assessment embedded int a larger intervention.	Approach to writing survey – adapted for primary school students from Inventory of Processes in College Composition (Lavelle, 2007). Used as pre-test to assess depth of students' approach to writing. Genre knowledge survey – researcher-developed, used as pre-test to assess declarative knowledge of analytic genres. Pre-test of analytic writing – a writing task developed by researchers, holistically assessed on a scale of 1-10 by 2 independent raters and analysed for rhetorical moves. Post-test – researcher-developed writing tasks designed to assess students' ability to learn through writing. Students were given a test of relevant science knowledge (human organ systems and nutrition, 11 items, Cronbach's α = .67), two writing activities (explanation, argument), and another science post-test (Cronbach's α = .74). The gain in science test scores were taken as the outcome measure.

* Meusen-Beekmana, Brinke, & Boshuizen (2016)	N	W	Netherlands	P	Participants actively contributed to setting criteria for assessment and a checklist was used to monitor the progress of peer assessment and self-assessment) Self-assessment condition: Self-assessment against criteria for improvement and teacher feedback on the process. Peer assessment condition: The students used the checklists and rubrics to peer-assess writing drafts.	There was a statistically significant effect of the intervention on self-regulation, F(2692) = 58.09; $p < 0.001$; partial ɳ2 = 0.14. Hedges's g was respectively 1.38 in the self-assessment condition, and 1.48 in the peer assessment condition. Participants also scored significantly higher on intrinsic motivation, F(16.94) = 6.49, $p < .05$; Hedges's g on intrinsic motivation are respectively 0.34 in the self-assessment condition, and 0.43 in the peer assessment condition. There was no significant differences on self-efficacy measure.	L, PR.	L	SEI	I	The intervention integrated Black and Wiliam's five key strategies (2003) in writing assignments in a peer-assessment condition and a self-assessment condition to develop self- regulated learning skills. Students completed 3 writing assignments. A series of planning and goal-setting activities preceded each assignment, designed by the researchers. Students in the control condition received the usual form of instruction.	Yes, increasing and encouraging self-reflection in writing: Students were given the opportunity to adjust their planning and complete the assignment and had self-set moments when they worked on their assignment with monitoring and reflection about where they were during the task, or what was needed for improvement.	After teachers marked the assignments, students did not have an opportunity to revise their assignments and did not receive any process-oriented instruction or cognitive-strategy instruction.	The Inventory Learning Style Questionnaire (ILS, Slaats, 1997) – a standardised measure self-regulation and motivation. The Self-Efficacy for Task Performance Questionnaire (STPQ, Van Meeuwen, Brand-Gruwel, Kirschner, De Bock, & Van Mer- riënboer, 2012) – a standardised measure of self-efficacy.

11.4 Science

Authors	Online tool (Y/N)	Domain A = Arts M = Mathematics PD = Professional Development R = Reading S = Science W = Writing	Location of study	Sample characteristic P = Primary S = Secondary Typical sample vs atypical sample. If atypical then describe characteristics	Form of formative assessment (description of the task used) What is the source of the assessment tool used? Who is the author? (e.g. classroom teacher, school-based learning community, assessment expert working with teachers, ready-made package (standardised/non-standardised).	Impact of the formative assessment on student learning outcomes (cite measures of impact here)	What is being measured? L = Learning (meets L/O) PR = Progress G = Gaps S = Specific difficulties R = Reasons for difficulties (cognitively diagnostic/task diagnostic)	Who is the feedback to? L = learner T = teacher S = software Who's behaviour is expected to change as a result of this feedback?	Type of feedback to learner: NA = Not applicable S = Score/grade provided only SF = Score/grade & feedback re: correct answer SE = Explanation of the difference: correct results & explanation of differences between their result and the correct result; SEI = Explanation and improvement suggestions: As above but now students also receive some specific suggestions for improvement; SEA = Explanation and specific activities: Students are given information about the correct results, some explanation, and specific activities to undertake.	Type of feedback to teacher: NA = Not applicable S = Overall score only SS = Separate scores provided for specific aspects of performance I = Possible explanation of the problem areas and suggestions for additional instructional focus A - Possible explanation of the problem areas and specific instructional activities to undertake.	Evaluation is based on theoretically valid TASK Model? Sequence of activities that need to be successfully completed to meet learning outcomes and how learners typically progress through them (learning progression)	The intervention is based on theoretically valid COGNITIVE Model? Model of prerequisite cognitive and learning skills underlying successful progression. e.g. Does the process require a significant amount of working memory, attention, motivation, persistency, cognitive ability, language skills etc.	Are the actions/interventions following the assessment task evidence-based? (i.e. is the INSTRUCTIONAL model valid?)	What tools/resources are used in the assessment process and intervention? (Could be teacher designed or commercial).

*	Decristan, Hondrich et al., 2015	N	S	Germany, central (urban and rural areas)	P	Inquiry-based science education (IBSE) baseline unit (5 lessons total) on floating and sinking adapted from Hardy et al., 2006; Möller, Jonen, Hardy, & Stern, 2002; it was implemented by classroom teachers after attending a workshop, standardised by providing teachers with all the materials. 4 groups: guidance by formative assessment (FA), peer-assisted learning (PAL), guidance by teacher scaffolding of instructional discourse (SID), IBSE-only intervention group (CG, control).	Controlling for pre-test, students in FA scored significantly higher in post-test than student in CG (beta = .24, SE = .12, p <.05), non significant results for other interventions.	PR & R	L	Unclear. Students were provided with ‘task-specific feedback and the assignment of differentiated tasks’.	NA	All groups worked through an IBSE unit which has been studied previously (Minner, Levy, & Century, 2009, meta-analysis). All teachers went through the same training and had the same materials. Teachers in the FA condition received information on designing and using diagnostic tasks to evaluate students' current conceptual understanding and how to provide informative and motivating feedback.	Authors note that analysing, summarising, and presenting science information requires sufficient language proficiency. Language proficiency is also required to participate in collaborative and communicative processes.	Formative assessment was embedded into a larger intervention.	Pre- and post-test – researcher designed, 8 multiple choice and 2 free response questions on floating and sinking. 7 of the questions were the same at pre-test and post-test. Science Competency Test – adapted from Trends in International Mathematics and Science Study 2007 (Martin, Mullis, & Foy, 2008), 5 additional items developed by researchers. CFT-20R – standardised measure to assess logical thinking (Weiß, 2006). Language Proficiency Test – adapted from German diagnostic tests of language comprehension (Elben & Lohaus, 2001; Glück, 2011; Petermann, Metz, & Fröhlich, 2010)., includes 20 items (Cronbach’s a D .72).

* Panadero et al 2012	N	S (Geography)	Spain, Madrid	S	Self-assessment tools – rubric and script (provided in appendices online) - rubric is more condensed, the script walks you through the bits needs to report/decisions that must be made when describing particular landscape features Feedback: performance vs process – performance feedback just stated what the student missed; process feedback did not state it directly but explained why the missing feature is important and what's the correct answer in this case. There were three between-group independent variables: (1) type of instructions, oriented to process or to performance, (2) presence or absence of self-assessment tool: control vs. rubric vs. script, and (3) feedback oriented to process or to performance.	The only significant effect on learning was that of the interaction between self-assessment tool and time, F(2, 108) = 7.85, p < .001; η2 = .127 – i.e. both rubric and script conditions outperformed control on all 3 trials.	L	L	S (performance condition) SE (process condition)	NA	It is not clear from the paper how the task related to main geography curriculum. The study took place in one session (2 hours 45 minutes on average), the students were taken out of the normal classroom environment to record audio for each individual's on-line self-regulation index as they were doing the task.	The study is primarily focused on self-regulation. Process feedback condition had higher self-efficacy scores after the intervention than performance feedback condition, F(1, 106) = 7.12, p < .01; η2 = .063.	The paper does not specify whether students learned about landscape analysis before or after the intervention.	Learning index – adopted from Alonso-Tapia & Panadero (2010). Participants wrote their conclusions once they had finished the oral analysis of each of the three landscapes. The written texts were divided into propositions, and then were evaluated as correct or incorrect using a specific analysis model for each landscape provided by two expert Social Science teachers. Questionnaire of Motives, Expectancies and Values, part A: goals and goal orientations (MEVA) – adapted from Alonso-Tapia (2005), 76 items measuring goal orientation. Emotion and Motivation Self-regulation Questionnaire (EMSR-Q) – adopted from Alonso-Tapia, Panadero, & Ruiz (2014), 36 items measuring self-regulation. On-line self-regulation index – researcher developed rubric to analyse student's self-regulation process. To obtain the data, students were asked to express their thoughts and feelings aloud while analysing the landscape. Self-efficacy questionnaire - researcher-designed, 8 items about students' perception of their ability to analyse landscapes.

*	Resendes et al., 2015	Y	S	Canada, Toronto	P	Comparative word clouds – one cloud shows more/less frequent vocabulary used by students within Knowledge Building (KB) tool, one with vocabulary from a source text of an expert on the topic, and one showing shared vocabulary between students and the expert. Clouds are continuously updated as the students use the KB tool. Epistemic discourse moves – a continuously updating bar-chart that shows frequency of use of each kind of scaffold in a specific Knowledge Forum view (e.g. if ‘my theory’ bar is very high and ‘important information + source’ is very low, it suggests that students have too many untested ideas and too few data). Control class vs 2 formative assessment conditions (A: comparative word clouds tool; B: word clouds tool + epistemic discourse moves tool)	Group B > Group A > Control on most lexical measures (p < .05; number of words written, number of domain words, number of unique domain words, % domain words above grade level), except number of academic words and % of words in the 1000 most frequent which did not significantly differ. Post-hoc tests (Tukey’s HSD) indicate that both experimental groups A and B performed better than the control group on scientificness (p < .01, Cohen's d = 0.59) and epistemic complexity (p < .05, Cohen's d = 0.39), but no significant difference between A and B.	G	L	NA The feedback only provide indications of areas that need improvement, without a score (as the learning task does not have any correct answers).	Unclear Teachers help students to progress their work with the KB tool, but the paper does not specify how involved the teachers are and whether they use the feedback tool to adjust their involvement.	All participants undertook the same 2 knowledge-building units, corresponding to Ontario curriculum's ‘Understanding Life Cycles' science strand for Grade 2 ("Growth and Change in Animals" inquiry stream). Students began with a 4-month study of birds, followed by 4 months investigating salmon.	The main research question is whether young students (~7 years old) can carry out productive metadiscourse. The study was conducting in a laboratory school at UoToronto that typically engages students in knowledge- building practices from kindergarten and use of Knowledge Forum from grade 1. The teacher helps students to explain their thinking, consider problem areas they have missed, etc. The study does not measure any cognitive variables in the students.	Formative assessment was embedded into a larger intervention.	Knowledge Forum – commercial tool, an online platform where students gather and exchange notes on a given topic. Knowledge Forum Analytic Toolkit (Burtis 1998) – used to calculate the number of notes read and written by each students, and obtain lexical measures. "Ways of Contributing to Explanation-Seeking Discourse" schema -- schema for content analysis of knowledge-building discourse (see Chuy et al., 2011) .
*	Soong et al 2010	Y	S	Singapore	S	Microsoft NetMeeting – used to collaborate anonymously with another student on solving the problems in the worksheet using the text chat and the whiteboard feature. 3 sessions over 2-week period, 1-1.5h. Following the collaboration session, teacher looks through the logs to identify problem areas to be addressed in the revision lesson.	Students in the experimental condition had significantly greater gain scores than the control group (completing normal worksheets), t = −3.20, p < 0.01. There was no significant difference in gain scores between control group and tutoring condition (students were receiving private tutoring outside the school), t = 0.672, p = 0.514. Students in the experimental condition also had significantly higher gain scores than students in the tutoring group, t = −4.89, p < .05.	S	T	NA	I? Teachers can look at text chat logs and see what questions/problem areas commonly arise.	Not specified	Not specified	After the assessment task, the teacher identified common misconceptions and conducted a revision lesson to address them.	Pre-test and post-test – no further information provided.

*	Terrazas-Arellanes et al., 2018	Y	S	USA, Oregon, Georgia	P, S – Grade 6 - 8; compare typical students, students with learning disabilities, and student learning English	Online textbook – escolar.uoregon.edu, includes interactive quizzes and summative assessments with corrective and explanatory feedback.	There was an overall significant treatment effect, F(1,29) = 16.8, p < .001, d = .65. The interactions between disability status and condition, and English language status and condition were nonsignificant, suggesting that both subgroups improved relative to controls. There was a main effect of learning disability, as they had lower scores than other students, F(2,041) = 5.6, p = .018.	L	L	SF	NA	Teachers were provided with tables showing how each unit's content aligns with NGSS and CCSS, detailed lesson plans, ideas for scaffolding activities onto background knowledge, and student assessment reports. Teachers also underwent one day professional development workshop on using interactive online resources. The intervention was based on a project-based learning (PBL) instructional method, delivered with a supportive multimedia learning environment and culturally relevant activities to address literacy, cognitive load, and access problems faced by students with learning disabilities and English learners (U.S. Department of Education, 2010).	The study specifically included English language learners because science requires greater language proficiency, placing additional demands on working memory. The intervention is designed to reduce cognitive load and thus aid learning. The same principle is behind including students with learning disabilities. In addition, the tool is designed to help them use their background knowledge via warm up activities and a structured learning approach.	Formative assessment was embedded into a larger intervention.	Pre-implementation readiness inventory – researcher designed measure for teachers to self-report their readiness to begin the project. Implementation log and post-implementation checklist – teachers in the treatment group used the log, control group used the checklist. Both meant to ensure adherence to the intervention. Post-implementation teacher and student surveys – a measure to gauge teacher and student perceptions of the intervention (treatment groups only). Content specific assessments – researcher-designed, aligned with Next Generation Science Standards. Four multiple choice tests with 25-64 items, Cronbach's alpha ranging from .81 to .93.
*	Vogelzang & Admiraal, 2017	N	S	Netherlands	S	Formative assessment – students worked through a series of questions in groups of 4, and received feedback from peers and the teachers. Students in the control group answered the questions individually.	On the topic of lactic acid, formative assessment group scored significantly higher on the post-test than the control group, F(1, 56) = 36.93, p < .001, η2 = 0.397; Cohen's d = 1.62. On the topic of polymers, formative assessment group also scored better on the post-test than the control group, F (1, 56) = 12.15, p < .001, η2 = 0.178, Cohen's d = 0.93.	L	L	Unclear. Feedback was provided to groups or individual students, focusing on students' understanding of the subject matter and their learning strategies.	NA		Not specified	Not specified	Pre- and post-test – designed by researcher (teacher). 20 declarative, procedural, schematic and strategic questions. Questions for formative assessments – same design as pre- and post-tests, 15 items.

*	Wang 2010	Y	S	Taiwan	P	GPAM-WATA (Wang et al., 2004; Wang, 2007, 2008) – an online dynamic assessment tool that uses a graduated prompt approach (increasingly specific hints). The treatment group received increasing specific hints but not the correct answer, the control group only received the correct answer.	There was a significant main effect of treatment, F(1, 109) = 235.974, p < 0.01, a significant effect of prior knowledge, F2, 109 = 8.020, p < 0.01, and a significant interaction between treatment and prior knowledge, F(2,109) = 4.185, p <.05. Effect sizes were not reported.	S (treatment) PR (control)	L, T.	S (control group) SF (treatment group received increasingly specific hints)	SS	Both groups completed a 2 week e-Learning instruction on plant photosynthesis. Students could access the materials both in class or after school. Teachers did not perform direct instruction of learning contents, but only guided the students in their learning. The only difference between conditions was whether students received only scores as feedback (control) or hints and opportunities to correct the answer (treatment).	The study separately evaluated students who had low, medium, and high scores on the pre-test as an effect of prior knowledge on learning effectiveness.	Formative assessment was embedded into the learning process.	Prior knowledge assessment (pre-test) – designed by the researchers, 25 items with Cronbach's alpha = .81. Summative assessment (post-test) – designed by the researchers, 48 items, Cronbach's alpha = .92.
*	Yin et al., 2014	N	S	USA	P	Interim formative assessments (3) – researcher-designed, part of an inquiry-based learning progression on floating and sinking. Each assessment included a predict-observe-explain (POE) activity and a challenge question. Answers were recorded individually, shared in a small group to generate group ideas, then group ideas were shared with the class. POE activity – the teacher gives an experimental demonstration, students record their predictions and then record what they observed to happen in the demonstrations, as well as their explanation of the event. Challenge question – students were given a novel scenario and asked to predict what would occur regarding floating and sinking phenomena.	MANCOVA, overall experimental group scores higher than control: Wilks' lambda = .87, F (2, 45) = 3.38, p = .043, partial etasq = .13. The experimental group scored higher on the performance assessment than the control, F(1,47) = 6.48, p = .014, partial etasq = .12, but not on multiple choice test. The overall effect of the treatment on conceptual change was significant after pre-test misconception score was controlled for, Wilks' l =.77, F (3, 43) = 4.30, p = .01, partial etasq = .23.	PR, S.	L, T.	SEI	SS	The content covered in the intervention was based on a unit of the Foundational Approaches in Science Teaching curriculum (Pottenger & Young, 1992) which focuses on explanations of 'why things sink and float'. There were 12 lessons on floating and sinking conducted, once a week. Formative assessments took place during lesson 4, 7, and 10, because students' understanding was expected to develop the most (based on the planned learning progression).	The topic of floating and sinking was chosen because it is commonly taught at this stage, but authors identified it as challenging and students often hold many misconceptions.	Formative assessment was embedded into a larger intervention.	Conceptual diagnostic items – researcher-designed, 13 multiple-choice items to assess common misconceptions about floating and sinking. Short-answer question – asks to explain why things sink and float; scored by a rubric. Developed by researchers. Achievement test – researcher-designed, 39 multiple choice questions to measure students' general achievement on the learning objectives, Cronbach's alpha = .79. Performance assessment – students were given equipment and asked to find the densities of some blocks and mystery liquids; scored by a rubric.

*	Zhang & Misiak, 2015	N	S	USA, Midwest	S	The study compared grading methods (point-based, rubric-based, a rubric with written feedback). Both the rubric, and rubric + feedback are described as formative assessment.	There was no significant difference between rubric only and point-based groups in both Year 7 (p = .08) and Year 8 (p = .43). There were significant difference in both years between the other conditions, such that rubric and feedback group performed better than rubric only (p = .01), and rubric and feedback did better than point-based group (p = .01).	L (rubric only, rubric and feedback) PR (all groups)	L	S (point-based group) SE (rubric only) SEI (rubric and feedback)	NA	All students were taught by the same teacher in an inquiry-based educational setting, using the same instructional tasks and same forms of assessments. A standard-based grading system was implemented to assess the concept development of eighth graders in the area of magnetism and electricity and seventh grade in astronomy. Each area was taught for approximately 4 weeks.	The authors point out that the rubrics help students identify specific areas to improve, in contrast to point-based grading.	Formative assessment occurred throughout each unit. Students developed content and inquiry standards through performing laboratory procedures, conducting experiments, analysing data, and exploring research questions. Tasks were levelled to guide students through concept development. After each completed task, students in the two standard-based groups responded to open-ended summary questions to communicate their progress toward standards. The summary questions and work completed on tasks were evaluated, and students were updated on their progress. After each evaluation, these students had the opportunity to make corrections and update any work which did not exceed standard level.	Pre- and post-test – researchers adapted items released from large-scale assessments in science, including the National Assessment of Educational Progress (NAEP), Trends in International Math and Science Study (TIMSS), and state standardized tests. Questions included multiple choice, short answer, and questions requiring to respond using diagrams ad tables. The number of items is not specified. Students' Motivation Towards Science Learning Questionnaire (SMTSL) – adapted from Tuan, Chin, and Shieh (2005), 35 items.

*	Zucker et al., 2013	Y	S	USA, Pensylvannia	S	Year 1: Standard curriculum vs SmartGraph exercises (+ standard curriculum). Year 2: ran 4 comparisons (more/less teacher experience, individual computers vs working in groups, more/less scaffolding, students of control teachers from Year 1 worked with SmartGraph). SmartGraphs – an online tool developed by Concord Consortium under an NSF grant specifically to help students learn to understand graphs. The tool uses a motion sensor to draw motion graphs from real data. The study included 5 activities on motion and acceleration due to gravity. The tool includes scaffolding that provides targeted hints adjusted in response to specific student answers. The scaffolding can be in the form of written hints, equations, or visual markers on the graph or table.	In the first year of the study, total gain in scores was significantly higher in the experimental group as compared to control, t = -2.669, df = 1684, p = .008, d = 0.13. Subsequent research showed that the gain scores of the experimental and control groups were not significantly different for the total test gain score, but the difference between the knowledge-integration gain scores is significant (experimental 4.59; control 4.01; t (1684) = -2.585, p = .049). Thus, the difference between the groups lies in the experimental students' greater abilities to explain their answers, indicating greater depth of understanding. In the second year of the study, all students had access to SmartGraphs. The most significant finding emerged from the cross year analysis which showed that students being taught by the same teacher in 2011 and in 2012, first without SmarthGraphs and then with, showed significantly greater learning gains in both factual and deep conceptual understanding. The effect size (d = 0.28) was correlated to moving a student from the 50th percentile on the total score to the 61st. There were some results which were both ambiguous and unclear, e.g. students using the software with less specific scaffolding for slope analysis had greater learning gains that those who utilised software with highly specific and adaptable scaffolding as related to student input.	L	S	Unclear. Students receive hints from the software.	NA	Learning goals of the unit are based on the analyses of Pennsylvania's science standards and the physical science textbooks used in the state. Activities progress from easier to more difficult.	Not specified	The teachers continued to instruct students before/after the SmartGraph activities according to their normal curriculum.	Pre- and post-test – developed by researchers, piloted for reliability and validity. Included 8 multiple choice questions and 12 open response items 'based on a knowledge-integration format'. Online weekly logs – for the teachers to document what was covered in class, technology used, special circumstances, etc.

11.5 Mathematics

Authors	Online tool (Y/N)	Domain A = Arts M = Mathematics PD = Professional Development R = Reading S = Science W = Writing	Location of study	Sample characteristic P = Primary S = Secondary Typical sample vs atypical sample. If atypical then describe characteristics.	Form of formative assessment (description of the task used) What is the source of the assessment tool used? Who is the author? (e.g. classroom teacher, school-based learning community, assessment expert working with teachers, ready-made package (standardised/non-standardised).	Impact of the formative assessment on student learning outcomes (cite measures of impact here)	What is being measured? L = Learning (meets L/O) PR = Progress G = Gaps S = Specific difficulties R = Reasons for difficulties (cognitively diagnostic/task diagnostic)	Who is the feedback to? L = learner T = teacher S = software Who's behaviour is expected to change as a result of this feedback?	Type of feedback to learner: NA = Not applicable S = Score/grade provided only SF = Score/grade & feedback re: correct answer SE = Explanation of the difference: correct results & explanation of differences between their result and the correct result; SEI = Explanation and improvement suggestions: As above but now students also receive some specific suggestions for improvement; SEA = Explanation and specific activities: Students are given information about the correct results, some explanation, and specific activities to undertake.	Type of feedback to teacher: NA = Not applicable S = Overall score only SS = Separate scores provided for specific aspects of performance I = Possible explanation of the problem areas and suggestions for additional instructional focus A - Possible explanation of the problem areas and specific instructional activities to undertake.	Evaluation is based on theoretically valid TASK Model? Sequence of activities that need to be successfully completed to meet learning outcomes and how learners typically progress through them (learning progression)	The intervention is based on theoretically valid COGNITIVE Model? Model of prerequisite cognitive and learning skills underlying successful progression. e.g. Does the process require a significant amount of working memory, attention, motivation, persistency, cognitive ability, language skills etc.	Are the actions/interventions following the assessment task evidence-based? (i.e. is the INSTRUCTIONAL model valid?)	What tools/resources are used in the assessment process and intervention? (Could be teacher designed or commercial).

* Abu-Hamour & Mattar (2013)	N	M	Jordan	P; three students with specific learning disability in mathematics in each of the two classes	Mathematics Curriculum Based Measurement (M-CBM) – progress worksheet that covers all computational skills studied that semester, the student is asked to complete as many questions as possible in 2 minutes. Partial credit is given for every correct digit. Developed by the researchers. Experimental group completed M-CBM worksheets and a summative assessment, the control group had summative assessment only.	Students who had M-CBM measure achieved higher grades in math (M = 77.43, SD = 9.11) than control group (M = 70.11, SD = 8.58). This difference was significant (t(68) = 3.45, p = 0.001) and it represented a medium-sized effect, r = .38 Students who had M-CBM measures achieved higher grades in the median M-CBM computation (M = 27.4, SD = 5.82) than control group (M = 18.2, SD = 5.23). This difference was significant (t(68) = 6.94, p < 0.001), and it represented a medium-sized effect, r = .64.	L	L, T Students are expected to improve in math achievement.	S	SS	Not specified	Not specified	Not specified	Curriculum-Based-Measurement in Math Computation (M-CBM) – multiple-skill worksheets that covered all computational skills for the second semester of third-grade math curriculum and administered them to the entire experimental class. End of academic semester test - 100 point final examination, covering: multi-digit addition without regrouping, multi-digit addition with regrouping, multi-digit subtraction without regrouping, multi-digit subtraction with regrouping, adding and subtracting, of fractions and math problem solving. Two equivalent forms of the test were used, both based on the curriculum and administered all students.
* Andersson, C., Palm, T. (2017)	N	A	Sweden	P	The study focuses on teacher professional development in which teachers learned about general principles and approaches to formative assessment. The paper notes that "all teachers implemented new activities that strengthened classroom practice based on the big idea of gathering evidence about student knowledge and skills and modifying instruction to respond better to identified student learning needs" but no more detail was provided about specific formative assessment tools used.	The result of the ANCOVA shows that, after adjusting for the pre-test scores, there was a significant difference in the scores on the post-test between the intervention group and the control group, F(1, 42) = 4.71, MSE = 7.74, p = 0.036, Cohen's d = 0.66. After controlling for the relevant type of proficiency in the pre-test, the intervention group achieved higher results on the post-test on procedural tasks and also tasks requiring other solution processes, but the difference between intervention and control groups was not significant (F(1, 42) = 3.32, MSE = 1.42, p = 0.075, d = 0.55, for the procedural tasks and F(1, 42) = 3.33, MSE = 4.28, p = 0.075, d = 0.56 for the tasks measuring other processes).	L	L Teachers who participated in the professional development program are expected to increase their knowledge/skills and change their attitudes/beliefs to formative assessment to improve their classroom instruction and foster increased student learning	Not specified	Not specified	Not specified	Not specified	The paper identifies key formative assessment strategies for teachers (clarifying learning intentions and success criteria, eliciting evidence of student understanding, providing feedback that moves learners forward, activating students as instructional resources for each other, activating students as owners of their own learning)	Pre-test – designed to provide information about attainment of Year 3 learning goals from the national curriculum. The test came in two parts, 40 minutes each, with 29 and 22 tasks in each part respectively. Calculators were not allowed. The test included number sense, use of numbers, algebra, geometry, probability and statistics, relations and change, as well as handling of procedures, use of mathematics concepts, reasoning, problem-solving, and communication. Post-test – similar to the pre-test, but developed for Year 4. It was done in two parts, with 23 and 13 tasks respectively. Calculators were allowed in part 2. The authors worked with experienced primary school teachers and national test developers to design the pre-test and the post-test. The tasks in both tests were classified as measuring either skills in applying procedures, or mastery of other mathematical processes.

*	Axtell, McCallum, & Bell 2009	N	M	USA	S – academically at-risk students enrolled in a 4-week summer program.	Detect, Practice, Repair Procedure (DPR; Poncy, Skinner & O'Mara 2006) – used to develop automaticity in solving simple numerical division problems (144 problems of two-digit number divided by a one-digit number leading to an answer 2-9). The procedure was repeated during every lesson in addition to usual instruction.	The DPR group had significantly higher mean post-test scores (M = 52.13, SD = 31.56) compared to the control group (M = 25.15,SD = 13.44), F(1, 34) = 6.49, p = .016, Cohen's d = 1.11.	L	L Students expected to develop greater fluency in solving simple numerical division problems	SF Students count the number of digits they got correct and graph their scores after each intervention trial on their grid sheets.	SS. Teacher can see which questions students got right or wrong.	The 144 division problems encompass all of the possible combinations so the assessment covers the entire topic.	Intervention focuses on automaticity (speed and accuracy) to develop fluency in solving basic division number problems.	CCC (Cover, Copy, and Compare) technique, (Hansen, 1978) – students copy the first 5 wrong or unanswered questions into the CCC matrix, add the correct answer provided by the answer sheet. Students repeat the problem to themselves 5 times, cover the matrix and copy out the problem from memory and repeat subvocally another 5 times.	Stopwatch to time 1 minute for the test Metronome to count 1.5s for each division question Pre- and post-test probes – 144 division problems leading to an answer between 2 and 9; students had 2 minutes to complete as many as possible. Intervention packet – inc. a full page of division problems for tap-a-problem, a CCC sheet with 5 rows containing 6 blank boxes, and another sheet with the same division problems as before but in random order (mad minute page). The packet also included a grid that allowed to track student's correct digits from the mad minute page.
*	Baten, Praet, & Desoete 2017	Y	M	Belgium	P – all groups included low performers (n = 55) having problems with early mathematics skills (or a Z-score on an early mathematics test of <.0.5) and at least average performing peers (n = 112).	CAI (computer assisted intervention) – the study compared a number of conditions including counting, comparison, metacognition, and active control. In all cases, the students received visual and auditory feedback after completing each item on the computer task (smiley/sad face, applause/sobbing sound); Active control condition played reading games on the computer. Authorship is unclear, but seems to be a ready-made package for use in schools.	ANCOVA with intelligence as covariate and calculation results (TEDI-MATH as post-test in kindergarten) as the dependent variable showed significant differences between groups at post-test, F(4,158) = 20.89, p < .001, η2 = 0.35, and a significant effect for intelligence, p < .001, η2 = 0.15. The Metacognitive CAI and Counting & Comparison CAI outperformed the Comparison CAI and the Active Control condition. The Comparison CAI was less effective than the Metacognitive CAI.	L	L Children are expected to improve in counting and comparing skills	SF Visual feedback was given by a happy or a sad smiley face. Auditory feedback was given by a sob when they made a mistake, or by applause when they succeeded	NA	Each CAI session went for 25 minutes, no further details about methodology is provided.	Opportunity-Propensity (O-P) model suggests that children are more likely to realize their potential for learning mathematics if they are provided Opportunities (O) to learn that content at school and in other contexts and have the motivation and capability or propensity (P) to benefit from the opportunities provided to them (Wang et al. 2013). Within this model, metacognitive skills can be considered as P-factor, whereas powerful learning designs can be seen as O-factor, both positively impacting the learning of mathematics.	Not specified	Computer program for presenting the CAI tasks – name or further details not specified.

*	Bond & Ellis 2013	N	M	USA	P	Researcher designed reflective prompt – had a verbal and a written component. First, students wrote a reflection with the help of the "I learned..." prompts, followed by a verbal "Thinking aloud" strategy (talking to another student). Students could edit their written reflections after verbal reflection.	Mean scores were significantly different between the Reflection group and Non-reflective Review group (post-test, p = .035; retention test, p = .036). Significant difference (p < .01) was found in all pairwise comparisons with the Control Group on both administrations of the mathematics test, with a large effect size (post-test, partial η2 = .269; retention test, partial η2 = .273)	L	L, T Students expected to become more reflective, leading to improved math performance	NA	NA	The post-test was researcher designed, not a standardised test. Post-test items were pilot tested resulting in 36-item multiple choice test on the maths content.	Metacognitive model of reflective assessment	Not specified	Post-test – 36 multiple choice questions drawn from "Connected Mathematics: Data About Us" (Lappan, Fey, Fitzgerald, Friel, & Phillips, 2002b). Scripted lesson plans – provided to teachers in experimental and control groups, derived from the Connected Mathematics Program. Experimental groups had lessons on probability and statistics, the control group had lessons on area and perimeter.
*	Bryant et al. (2011)	N	M	USA	P; students with mathematical difficulties	Early numeracy preventative Tier 2 intervention – teachers went through a professional development program and administered mathematics lessons that included systematic instruction, visual representations of mathematical concepts, opportunities for practice and progress monitoring. The progress monitoring measure (TEMI-PM) – 4 group-administered subtests to assess different topics in mathematics, have to complete as many items as possible in 2 minutes.	Statistically significant differences for the treatment group on the Addition and Subtraction Combinations (p ≤ .0001; g* = .55), Place Value (p ≤ .002; g* = .39), Number Sequences (p ≤ .00001; g* = .47), and the TEMI-PM Total Score (p < .01; g* = .50). No differences were found on the Magnitude Comparisons subtest (p = .16; g* = .18). g* = Hedges g, a measure of effect size very similar to Cohen's d used for sample sizes < 20.	L	T	NA	SS	None mentioned	None mentioned	The formative assessment component is embedded into a longer Early Numeracy Intervention	Texas Early Mathematics Inventories – Progress Monitoring measures (TEMI-PM) SAT-10 (Pearson, 2003) – standardized mathematics achievement test, administered at appropriate grade difficulty. TEMI-O (University of Texas System & Texas Education Agency, 2007a) – group- administered problem-solving and whole-number computation measure.
*	Carlson et al., 2011	N	R, M	USA, multiple states	P, S	4Sight – quarterly benchmark assessments in reading and mathematics, aligned with state standards and supplemented by advice from consultants (John Hopkins Centre for Data-Driven Reform in Education intervention).	The treatment group had significantly higher post-test scores in mathematics as compared to control (p <.05), with the estimated increase of 0.06 student-level SDs in the treatment group. For reading, the difference was not significant (p >.05).	L	T	S	S	Not specified	Not specified	Not specified	State-administered achievement tests – school-level performance on mathematics and reading tests was analysed as the outcome variable.

* Chappell et al (2015)	Y	M	USA, Virginia & Kansas	P, S; low achieving (Tier 3 intervention) children in grades 6-8.	Online tutoring provided by Focus EduVation (FEV) - interactive tutoring for K–12 students, done individually, using instant messaging and a virtual whiteboard. Diagnostics assessment completed before tutoring - program and school personnel developed individualised learning objectives for each student. Learning plan developed on this basis, aligned with curricular standards and scope and sequence. Tutoring was intended to provide a differentiated, engaging environment where skills were enhanced through sharing of curricular materials, practice problems and visuals, and graphic feature to aid communication, as well as collaboration between students and tutors. Focus on providing students with expert instructional explication to promote lateral transfer and scheme development.	Comparison between tutored and non-tutored students in the same school revealed no significant differences in post-test scores (ANCOVA controlling for pre-test scores, students well-matched on multiple demographic criteria - propensity score matching), although within group effect sizes for pre- to post-intervention scores were higher for tutored (d = .95) compared to non-tutored students (d = .24). Tutored students in a second school (but no within school non-tutored comparison) showed great gains (d = 1.47). Note that students in school 2 participated in an average of 23 hours of tutoring compared to 14 hours in school 1.	L	L, T	SEA Feedback to the learner which the tutors described (in logs) as guided practice using multiple explanations and representations of target concepts. Tutors described accessing prior knowledge, modelling, explaining steps in a math process, identifying process and operation errors, and scaffolding through the use of questions/prompts.	I. Because the tutor is getting direct feedback on the student performance, they are aware of problem areas and the additional instructional activities that may be required.	No single sequence of activities, as tutoring is individual to each student depending on their needs.	Student perceptions of the pacing of online tutoring were mixed, which is possibly a function of the wide variety of reasons students may be assigned to a Tier 3 mathematics intervention. For example, for some students the achievement gap may be due to processing issues, and thus may require more explanation, while for others, attention-related issues may demand a faster pace.	Formative assessment embedded in the individualised tutoring process, cannot separate.	2013/2014 administrations of the Virginia Standards of Learning (SOL) assessment – used as pre-tests and post-tests for School 1. Program-administered tests – used as pre-tests and post-test for School 2 (state-level assessments were unavailable due to moratorium on testing in 2014, no further details on program-administered tests provided).
* Clarke et al. (2011)	N	M	USA	P; students at risk for mathematical difficulties (66% of the sample for this study)	Early Learning in Mathematics (ELM) – 120 lesson program developed by the researchers, providing instruction in number operations, geometry, measurement, and vocabulary. The intervention includes formative assessment in the form of frequent reviews of key concepts.	Students not at risk for math achievement difficulties in ELM classrooms did not make gains over those in control classrooms. However, the study found statistically significant improvements for at-risk children in ELM classrooms over controls on the TEMA raw scores (t(61) = 2.39, p < .002) and EN-CBM total score (t(61) = 2.54, p = .014).	L	T, L Students (particularly at-risk students) are expected to improve their performance.	Unclear Paper says students are provided with "specific and immediate feedback by the teacher as they verbalize and explain their solutions and understanding of the underlying mathematical concepts".	NA	Not specified	Not specified	Explicit instruction model: Following teacher models, students solve similar problems and are provided specific and immediate feedback by the teacher as they verbalize and explain their solutions and understanding of the underlying mathematical concepts. Finally, instructional materials include frequent and cumulative review of key concepts.	4 researcher-developed measures: Oral Counting, Number Identification, Quantity Discrimination, Missing Numbers. Test of Early Mathematics Ability (TEMA-3) – standardised measure of formal and informal mathematics.

* Clarke et al. (2014)	N	M	USA	P; low-performing Year 1 students	Formative assessment took place during Tier 2 lessons by teachers "providing timely academic feedback to confirm correct student responses and address potential misconceptions"	A mixed effects model was used to analyse the data. The only significant difference between conditions over time was on the ProFusion measure, p = .015, Heges' g = 0.82.	L	T, L	Unclear	Unclear	Each lesson includes: introduction of new content, systematic practice and review in 4-5 brief scripted activities. Lessons include teacher modelling, scaffolded instructional examples, and opportunities for teachers to provide academic feedback based on student responses to individual and group questions.	Not specified	Formative assessment embedded in the learning process, not specified whether the activities change depending on outcome of formative assessment tasks.	Early Numeracy Curriculum-Based Measures (EN-CBM, Clarke & Shinn (2004)) – proximal measure of students' procedural fluency, all measures timed for 1 minute. SAT-10 (Harcourt Educational Measurement, 2002) – group-administered two of the mathematics subtests as distal measures of mathematics performance. ProFusion – researcher-developed measure of conceptual and procedural knowledge.
* Clarke et al. (2015)	N	M	Finland	P; includes students at risk for mathematical difficulties	Early Learning in Mathematics (ELM) – 120 lesson program developed by the researchers, providing instruction in number operations, geometry, measurement, and vocabulary. The intervention includes formative assessment in the form of frequent reviews of key concepts.	Overall, students in ELM classrooms achieved mathematic outcomes that were not significantly different than those achieved by students in control classrooms (the time/condition interaction with TEMA-3 score came closest to significance, with p = .0517; effects on other independent variables were highly non-significant).	L	T, L	Unclear Paper says students are provided with "specific and immediate feedback by the teacher as they verbalize and explain their solutions and understanding of the underlying mathematical concepts".	Unclear	A typical ELM lesson includes four to five activities, each of which focuses on one of three content strands: number and operations, measurement and data, and geometry. Across the three content strands, a fourth strand, vocabulary, is integrated to increase the amount of student mathematics discourse and use of mathematics-specific vocabulary. ELM is fully aligned with the learning objectives for kindergarten as specified in the CCSSI (2010).	Not specified	Formative assessment embedded in the learning process, not specified whether the activities change depending on outcome of formative assessment tasks.	4 researcher-developed measures: Oral Counting, Number Identification, Quantity Discrimination, Missing Numbers. Test of Early Mathematics Ability (TEMA-3) – standardised measure of formal and informal mathematics.

* Faber, J.M., Luyten, H., & Visscher, A. J. (2017)	Y	M	Netherlands	P	Snappet – digital formative assessment tool. Provides the same instructional content as regular curriculum, and assignments are comparable to those used in traditional paper-based settings. No specific statement of who the author is, but appears to be the teacher – there are curriculum assignments, and based on performance children then move on to adaptive assignments - IRT model is used to predict student ability levels in the basis of previous responses. Teachers decide which assignments students need to work on.	Students in the experimental condition showed significantly better performance at post-test (after controlling for pre-test ability) – ~0.2SD advantage over the 5 month period. Difference between control and experimental in mean achievement growth was highest for top 20% performing students ($b = 0.08$, $p < 0.01$). Note that lower performing students did show benefit from using Snappet, but comparison to control children at same ability level showed this level of benefit was not as great as for higher achieving children.	L	L, T, S Feedback to software in terms of number of curriculum and adaptive assignments completed.	S Learner is simply told whether the answer is correct/incorrect.	SS. Teacher can follow the progress of a lesson, of an individual student, or of the entire class. Individual student monitoring shows performance on specific learning goals (e.g. add numbers till 100) compared to that student's performance on other learning goals. Teacher can also see how their class performance on learning goals compared to other classes who also use Snappet.		Not specified	Snappet uses Item Response Theory to adapt item difficulty following feedback on the previous items.	Standardised assessment of math and spelling (used in most Dutch primary schools). Student motivation survey (based on two previous Dutch studies). Classroom observations of experimental group only - used to measure the degree to which teachers (based on Snappet feedback) differentiated their teaching in line with instructional need of students. The observation tool was developed by the study authors - note reliability from previous pilot testing is relatively low (alpha = .69). Snappet log files measured students intensity of use – total number of assignments completed, percentage of adapted assignments completed.
* Hsiao et al (2017)	N	M	Taiwan	S	Problem-solving Assessment, Diagnosis, and Remedial Instruction (PSADRI) system – contains two modules: for assessment, and for training and instruction. The assessment module includes portfolios that record students' scores, errors, and analysis of error types. These are used to provide personalised training and instruction. The control group learned the same content by the traditional instructional approach.	The experimental and control groups differed significantly at both pre- and post-test. Controlling for pre-test score, the experiment group scored significantly higher than the control ($F1, 150 = 8.729$, $p = .004$, $η2 = .055$) at post-test. When looking at scores on the 4 individual assessments of problem-solving ability, from problem translation and integration, students with lower scores on these items at pre-test showed greater benefit from traditional instruction, while children with higher scores showed more benefit from PSARDI. Overall problem solving was higher in the experimental group regardless of pre-test ability. No difference between experimental and control on solution planning and monitoring.	R, L	L, S	SEA Assessment module provide score and info on errors. Consequent training and instruction module is adapted based on the errors made in the assessment module.	NA	Yes – authors designed the contents of the training system in accordance with the competency indicators of math for the 7th grade. Students in the experimental and control groups learned the same content (linear equations) and carried out the same activity procedures with different teachers. The training system was used for assessment, diagnosis, and teaching guided mathematical problem-solving in the experimental group, whereas the control group participated in traditional instruction.	Yes – builds on Mayer's (1992) four stages in solving a problem: problem translation, problem integration, solution planning and monitoring, and solution execution.	The digital tool uses Item Response Theory to adjust instruction that follows the assessment module.	Pre and post-tests – researcher-developed, both contained 10 mathematical word problems. Scoring criterion of problem-solving ability – was used to evaluate students' ability to solve problem by measuring each question of the pre- and post-test. The total score for each student was assessed by two math teachers and converted into a percentile. The scoring criterion of problem-solving ability was designed by Szetela and Nicol (1992) and was modified into the Chinese version by Chin, Lin, Lin, and Tuan (2009). The Cronbach's alpha coefficients for the four items scored were between .703 and .981 and the reliability was .812

* Irving et al. (2016)	N	M	USA (28 states) & Canada (2 provinces). Note that data from Canada was not included in the analysis for this paper.	S	The longitudinal study (3-years) investigated how teachers who had received professional development on teaching with classroom connectivity technology (CCT) used it to teach Algebra. Authors claim that CCT enhances teacher opportunities for formative assessment and increases the potential for teachers to gain deeper understanding of student learning during classroom instruction. Texas Instruments Navigator – allows to connect students' graphic calculators to teacher's computer. Answers can be displayed for whole-class review, providing for formative assessment opportunities.	Analyses compared achievement within the 5 treatment groups with the control group. Analyses controlling for teaching experience and pre-test algebra scores revealed 2 (out of 5) significant treatment effects in favour of CCT (effect sizes .30 and .20). The 3 remaining comparisons were not statistically significant (effect sizes .23, .24, .13, p's > .15). There was a concern that pre-tests may have been administered late in some of the treatments groups; this may result in over-inflated pre-test scores and hence an unfair adjustment when it is controlled for in the analysis of post-test score. The same analysis including only teacher experience as a covariate revealed that 3 of 4 comparison were significant, with the treatment groups achieving higher algebra post-test scores than the control group. The authors point out that medium effect sizes are relatively rare in national randomised control studies.	L	L, T Students are expected to improve their test scores in algebra and teachers are expected to make better use of CCT in teaching algebra.	Not specified Paper describes how CCT can potentially improve the nature of formative feedback provided to teachers and students	Teachers see answers to formative assessment questions as part of the instruction process.	Not mentioned	Not mentioned	Authors refer to *representational expressivity* to describe the transformation of traditional communication forms through the use of software that broadens the representational infrastructure of the classroom. CCT technology such as that used in this study provides multiple representations of mathematical objects as well as accurate and timely collection and aggregation of data/expressions contributed by students to the discourse space.	Algebra pre-test (National Center for research on evaluation, Standards, and Student testing (CreSSt), 2004) – 32 pre-algebra and algebra multiple choice, short-answer, and extended constructed-response format items; Algebra post-test (Abrahamson et al., 2006) – 24 multiple-choice items, 5 extended-response items, and 1 three-part short answer question. Student views about Mathematics (Pape, Kaya, Owens, Irving, & Boscardin, 2006) – a survey.

*	James & Folorunso (2012)	N	M	Nigeria	S	Formative test with feedback and remediation group – after doing the formative test and receiving scores as feedback, students with the highest score in each of the 3-4 sections of the test lead the class discussion to identify the correct answer to each question. Formative test with feedback – students only receive scores as feedback. Formative test only (control) – students took the test but received no feedback.	There was a significant main effect of group, F(2, 236) = 174.976, $p < 0.05$, where feedback with remediation group had the highest mean post-test score (M = 27.45, SD = 3.38, cf. remediation only, M = 21.20, SD = 4.56; control, M = 14.43, SD = 3.34).	PR	L Students are expected to improve their performance in future topic tests.	SE	SS	Not specified	Not specified	Mastery learning and remediation. Class teachers gave feedback to students following topic tests in mathematics. This involved class discussion to encourage students to identify correct answers to the test questions, allowing students to seek clarification on areas of difficulty, and asking probing questions of the class.	Socio Economic Status Questionnaire (SESQ) – researcher-developed demographic questionnaire. Mathematics Achievement Test (MAT) – developed by the researchers, no further details about the test is provided.

* Koedinger, K.R., McLaughlin, E.A., & Heffernan, N.T. (2010)	Y	M	USA, Massachusetts	S	ASSISTments - web-based mathematics cognitive tutor, provides feedback and hints for problems. Developed by a non-profit organisation at the Worcester Polytechnic Institute. There were 3 treatment and 1 control school, whereby only students with available pre- and post-test scores were included in the study.	ASSISTment group was found to have higher post-test scores (compared to control), but this was specific to SPED children (medium effect size), not typically achieving children (not significant). The authors split the analysis by further subgroups and show significant impacts for free school lunch, non-white, and SPED. Based on reported degrees of freedom it is clear that these categories are not independent. The authors also report a significant interaction between teacher usage and student usage, F(2, 734) = 3.67, p = .03. They do not report effect size, but the reviewer's calculation is that partial η2 = .009, a very small effect size. Thus, teacher usage does not really have a practical impact either as a main effect or an interaction with student usage. Authors conclude that "we cannot be certain that the results are caused by ASSISTments due to the nature of the quasi-experimentation and potential selection bias".	L	L, T	SEA If the student gets the item incorrect they receive feedback in terms of scaffolds for that item. If the student does not know they can also request hints.	SS/I Information provided to the teacher is number of attempts made by student, number of hints requested, reaction time, and number of opportunities to practice. Online reports and automated emails provide reports on individual student's strengths and weaknesses, as well as performance of entire class. It is not really an explanation of the problem areas (I).	ASSISTments scaffolds problems into requisite skills and knowledge components. If student incorrectly answers original item or requests help, the first scaffold is automatically presented. Once in scaffold tutoring, student has to complete the series of scaffolds for that item.	Not specified	Use of ASSISTments was embedded into the classroom instruction of different teachers in different schools and was not experimentally controlled for.	MCAS (Massachusetts Comprehensive Assessment System) – designed to differentiate students from a broad range of potential ability, so many items in 7th grade test may be above or below what is studied by the students and a treatment effect is expected to be small. Scores from the previous year were used as pre-test, and the adjusted 7th grade scores as post-test.

* Konstantopoulos, Miller, & van der Ploeg (2013)	Y	M, R	USA	P	Wireless Generation's mCLASS – commercial product for K–2; has literacy and numeracy components, both providing teachers with detailed feedback on students' error patterns and reading/problem-solving strategies. Acuity (CTB/McGraw-Hill) – commercial product for Grade 3-8; designed to forecast performance on the Indiana state test (ISTEP) through short assessments (30–35 multiple choice questions). Both assessment types provide performance reports against Indiana standards with individual and group summaries.	The treatment effect was significant for K-8 Mathematics (Estimate = .187, SE = .70, p < .05), but not Reading (Estimate = .098, SE = .055) in the treatment on treated analysis, but not significant for either Mathematics or Reading in the intention to treat analysis. The significance varied within more narrow age brackets, as well as between urban and rural schools. The authors conclude that the overall treatment effect was positive.	L	T By adding meaningful detail to teachers' awareness of students' current performance relative to prior performance, teachers' instruction would closely match student needs and current and intended knowledge gaps would be reduced.	NA	SS	Not specified	Not specified	Not specified. The study only compares differences in achievement between schools in the different conditions, without controlling for actual usage of mCLASS and Acuity tools or specific instruction methods.	ISTEP+ – Indiana state test for reading and mathematics (Grades 3 - 8). Terra Nova – standardised test for mathematics and reading (Grades K - 2).
* Menesses & Gresham (2009)	N	M	USA	P; students with below-average performance in mathematics	Reciprocal Peer Tutoring (RPT) activities – students take turns tutoring each other, working together to prompt, monitor, and evaluate each other while learning a specific academic skill. Designed by researchers. Students doing RPT expected to perform better on basic maths questions that the control groups (non-RPT, business-as-usual).	A Bonferroni post-hoc test showed that the experimental condition produced higher scores (M = 34.88) than the control group (M = 19.72) and the tutees produced higher scores (M = 32.90) than the control group (M = 19.72). The main effect of time was not significant, F(1, 54) = .04, p = .842, partial n2 = .001, meaning there was no significant difference between post-intervention and follow-up scores. The interaction between group and time was also not significant, F(3, 54) = .75, p = .526.	L	L	SF	NA	Reciprocal Peer Tutoring (RPT), in which students alternate between roles of tutor and tutee so that both students have access to all of the advantages of peer tutoring (Fantuzzo, King, & Heller, 1992).	Not specified	NA	Computer-generated CBM mathematics probes (Math Computation Probes) – commercial tool (www.interventioncentral.org), used to measure procedural fluency by counting correct digits produced to 60 questions in 2 minutes.

*	Phelan et al. (2011)	N	M	USA	P	Checks for Understanding – a sequence of formative assessment activities designed by researchers. The sequence includes administering an initial check for understanding, a big idea and its application (15-20 min), and analyse results. Present instructional activities (if necessary) addressing deficiencies in conceptual understanding identified in Step 1 (one class period). Administer a second check for understanding focusing on conceptual understanding (15 min), and follow up instruction if necessary. Present instruction on applications of the big idea to problem solving and symbolic representation and computation tasks (if necessary; 15 min). Administer a third check for understanding focusing on conceptual understanding (15 min), and follow up instruction if necessary.	Hierarchical model results indicate no statistically significant main effect of transfer measure score for treatment or study design. In addition, the treatment and design interaction effect was not statistically significant. These results indicate that treatment effect did not differ by the two different designs implemented here. However, we found a main effect of the pre-test score. The estimate of the pre-test mean was 1.18 and its p value was less than .01. The students in classes with higher pre-test mean scores tended to have higher mean scores on the transfer measure as well.	L	T Students expected to possess a better understanding of the basic mathematical principles and be able to apply concepts they had learned, solve complex problems, and transfer the principles contained in the study domains.	NA	SS	Not specified	Not specified	Checks for understanding are meant to be followed up with additional instruction by the teacher if needed, focusing on the specific aspect of understanding (e.g. conceptual, application to a larger problem).	Pre-test – researcher designed, 28 items similar to past state test items for Grade 5. Transfer measure – researcher designed, includes 19 multiple choice, 9 short-answer, and 1 explanation question. Items were taken from Trends in International Mathematics and Science Study, National Assessment of Educational Progress, the Qualifications and Curriculum Authority Key Stage 3 exam, Programme for International Student Assessment, and benchmark tests used in a previous pilot study.

* Phelan et al. (2012)	N	M, PD	USA	P	Checks for Understanding – a sequence of formative assessment activities designed by researchers, same as in Phelan et al. (2011). The focus of the study was POWERSOURCE© – researcher-developed professional development program and supplementary materials designed to support the implementation of Checks for Understanding in four domains of algebra (rational number equivalence, properties of arithmetic, principles of solving linear equations, applications to other areas of mathematics).	There was a statistically significant difference between experimental and control groups across all four domains of algebra, short and long-response items, and total item scores. Broken down by domains, the largest effect size (1.25 SD) was for principles of arithmetic, followed by applications of core principles to other domains (0.89 SD), solving linear equations (0.81 SD), and for rational number equivalence (0.73 SD). These results suggest that students whose teachers completed the POWERSOURCE program outperformed control students substantially.	L	T Teachers expected to become more proficient in their subject matter knowledge, more skilled in their formative use of assessment, and better equipped to focus their instruction on key ideas.	NA	SS	Not specified	Not specified	Checks for understanding are meant to be followed up with additional instruction by the teacher if needed, focusing on the specific aspect of understanding (e.g. conceptual, application to a larger problem).	State standard test – used as pre-test; 2005-2006 data was used, from the test administered prior to pilot test year; Checks for understanding – scores were used as the outcome measure, as the main focus of the study was POWERSOURCE© (the professional development program + instructional materials)
* Pinger et al. (2018)	N	M	Germany, Hesse	S	Diagnostic tool – designed by researchers; allows to assess students' understanding and to provide feedback. Two components: (1) assessment: one or two mathematical problems and space for the student to write down the solution, (2) process-oriented feedback: three text fields to indicate strengths, weaknesses and strategies to improve.	No statistically significant effects on students' post-test achievement. The statistically significant positive coefficients for process-orientation and use of instructional time indicate that there was a positive association between these two aspects of instructional quality and students' achievement. However, the negative coefficients for the interactions in the same models indicate that this positive association was suppressed by the formative assessment intervention.	L	L, T Students expected to improve learning of Pythagoras' Theorem; teachers expected to improve their classroom instruction.	SEA	SS	The teaching unit consisted of 13 lessons (45 min each) and had four phases: (1) an introduction including a proof and technical tasks, (2) word problems, (3) modelling problems and (4) consolidation. To keep instruction as consistent as possible, all teachers received detailed guidelines which included a description of the teaching unit and a description of learning goals to be achieved in each phase. Additionally, teachers were given illustrations of obligatory teaching material to ensure that all students worked on the same tasks.	The model of instructional quality included three dimensions: cognitive activation (aspects of the instruction that promote the depth of students' cognitive engagement with the subject matter), supportive climate (a warm and caring teacher–student relationship and student-oriented individual support) and classroom management (prevention of disciplinary problems are relevant for on-task behaviour in class and thus are seen as prerequisite for high-quality motivational and cognitively-activating learning activities)	It is not clear whether the teaching sequence set any in-class time for students to act on the received feedback.	Pre-test – researcher designed, 19 items, focused on relevant prior knowledge such as identifying a right-angled triangle and solving equations but did not assess Pythagoras' Theorem directly. Post-test – researcher designed, 17 items, included technical tasks, word problems and modelling tasks. Items were previously analysed in a scaling study (Harks et al. 2014a). 3 questionnaire scales to capture process-oriented instruction, teacher-student relationship, and effective use of instructional time – adapted from previous research.

* Polly et al (2017)	Y	M, PD	USA, North Carolina	P (and Kindergarten)	AMC (Assessing Math Concepts Anywhere) – online formative assessment tool. Supports the process of conducting diagnostic interviews on students' mathematics understanding. Teachers use AMC in a one-on-one setting, in which students solve tasks either with manipulatives, such as counters, 10 frame mats, snap cubes, or using mental mathematical reasoning. The tool includes a built-in rubric, which then provides teachers with a rating on the assessment, which aligns to the instructional materials from the Developing Number Concepts (DNC) curricular resources (Richardson, 1998). The expectation in this project is that teachers used the assessment with every student in their class and then used the data to plan and implement differentiated, targeted instruction. In this study, all teachers used AMC, the comparison is between teachers who underwent professional development specific to using AMC and formative assessment.	Note the study does not compare outcomes for treatment vs control group. The authors looked at linear growth models within each group using multilevel hierarchical linear modelling. For the comparison group (AMC but no PD) they calculated the average intercept and slope against which to compare the intercept and growth of the treatment groups. Students in the treatment group showed significantly lower initial performance. It is not clear from the results description if the growth rate was higher for students in the AMC with PD groups. There were statistically significant relationships between the use of formative assessment practices and primary-grade students' achievement on number sense tasks using AMC. Further, students from impoverished settings, larger schools, and students who were assessed more frequently were associated with greater growth than their peers. Teachers who had undergone PD were in schools which on average were higher poverty and started at a lower achievement level. Children in these schools showed more improvement over time, but that is to be expected. Too many confounding variables to be able to clearly state the mechanisms of formative assessment that impact on learning.	L	T	NA	A. AMC reporting features provide teachers with information about students' instructional needs as well as links to the accompanying curricular resources which provide activities for students to explore, related to the skills in the assessments. The activities can be differentiated easily to meet students' mathematics needs.	The authors did not monitor or instruct teachers about any sequence of activities besides using the AMC. They do note that more frequent use of the tool was positively associated with the growth rate of student achievement.	Not specified	Use of AMC was monitored throughout the year, but the authors did not monitor the specific sequence of activities following assessment by the teachers, due to the large-scale nature of the study.	AMC assessments scores – the researchers gathered letter grades from various assessments conducted by the individual teachers throughout the year. They used Item Response Theory to transform letter grades into interval-level scale scores using the Rasch model, such that the scores have a mean of 500 and a standard deviation of 100.

* Polly et al (2018)	Y	M, PD	USA, North Carolina	P (K-1)	AMC (Assessing Math Concepts Anywhere) – online formative assessment tool. Supports the process of conducting diagnostic interviews on students' mathematics understanding. Teachers use AMC in a one-on-one setting, in which students solve tasks either with manipulatives, such as counters, 10 frame mats, snap cubes, or using mental mathematical reasoning. The tool includes a built-in rubric, which then provides teachers with a rating on the assessment, which aligns to the instructional materials from the Developing Number Concepts (DNC) curricular resources (Richardson, 1998). This study focuses of professional development (Assessment Practices to Support Mathematics Learning and Understanding for Students - APLUS) designed to support use of AMC and the related instructional activities (DNC). 72 hours of PD which includes a summer institute, workshops and classroom embedded learning activities during 1 school year. Focus on phases of number sense development (counting, addition & subtraction, early concepts of place value).	There were no consistent differences between treatment and control groups across 4 counting tasks across 6 school districts. Only significant effect is that better achievement was predicted by use of more assessments across the year (regardless of treatment or control group).	L	T	NA	A. AMC reporting features provide teachers with information about students' instructional needs as well as links to the accompanying curricular resources which provide activities for students to explore, related to the skills in the assessments. The activities can be differentiated easily to meet students' mathematics needs.	The authors did not monitor or instruct teachers about any sequence of activities besides using the AMC. They do note that more frequent use of the tool was positively associated with the growth rate of student achievement.	Not specified	Use of AMC was monitored throughout the year, but the authors did not monitor the specific sequence of activities following assessment by the teachers, due to the large-scale nature of the study.	Teacher practices and self-efficacy questionnaires – only administered to treatment group teachers as control group teacher IDs were not known until the end of the study. AMC assessment scores – do not specify how they transformed letter grades into scores.

* Racoksky et al (2019)	N	M	Germany, Hesse	S	Diagnostic and feedback tool – researcher-designed, employed according to a partly standardised procedure. The tool includes an assessment with 1–2 mathematical tasks assessing previously taught content. At the end of lesson 5, 8, and 11 the teachers asked students to complete the tasks on the diagnostic tool (15 min). The teachers assessed students' solutions and wrote process-oriented feedback on their strengths, weaknesses, and recommended strategies to continue. The tool included a list of cognitive processes and operations needed to solve the diagnostic tasks as an aid to teachers. At the end of the feedback part students were asked to complete an additional similar task and apply the strategies provided in the feedback.	Students in the formative assessment condition perceive the feedback as more useful, and reported greater self-efficacy. The authors report various indirect relationships between formative assessment, perceived usefulness, self-efficacy, and achievement. Formative assessment had no statistically significant total effect on achievement (β=0.212, SE=0.252, p = .401).	R, L.	L	SEI	NA	All the teachers were introduced to the subject-specific content and were provided with the mathematical tasks for the first 13 lessons of the teaching unit on Pythagoras' theorem. The teachers assessed students' performance at three predefined points in time (in the 5th, 8th, and 11th lessons) and provided students with written process-oriented feedback in the following lesson using the diagnostic and feedback tool.	Yes – based on Blum's approach of mathematical modelling as a cognitive domain model and its description of learning progression, the 13 lessons were divided into 4 phases: a) introduction, proof, technical tasks, b) dressed-up word problems, c) modelling problems, and d) consolidation.	Yes – students are provided with feedback and asked to complete a similar task that allowed them to implement the strategies provided as part of feedback.	Pre-test – researcher designed, 19 items, focused on relevant prior knowledge such as identifying a right-angled triangle and solving equations but did not assess Pythagoras' Theorem directly. Post-test – researcher designed, 17 items, included technical tasks, word problems and modelling tasks. Items were previously analysed in a scaling study (Harks et al. 2014a). Three questionnaire scales to capture process-oriented instruction, teacher-student relationship, and effective use of instructional time – adapted from previous research.

* Randel et al. (2016)	N	M, PD	USA, Colorado	P	The Classroom Assessment for Student Learning (CASL) – a professional development (TPL) program developed by the South Carolina Department of Education. Covers assessment purposes, accuracy of assessment, and using assessment results.	The CASL schools' adjusted mean on the CSAP Mathematics test was 502.49 (SE = 2.53), compared to an adjusted mean of the control schools of 501.91 (SE = 2.44) with an adjusted difference of 0.58 that was **not statistically significant** (SE = 3.47, p > .05).	PR	T Teachers' behaviour expected to change as a result of participating in the PL program; no involvement of any external facilitator, but teachers hold team meetings every 2-3 weeks to work through the Handbook, chapter by chapter	NA The professional development program includes ideas for teachers on how to provide feedback to students.	NA	Not specified (main focus is on the TPL program).	Not specified	Not specified	CASL implementation logs – teachers completed brief logs to describe their study of CASL materials throughout the year. At the end of the year, teachers described how they implemented CASL in the classroom throughout the year. Test of assessment knowledge – researcher-developed, 60 items to test teachers' knowledge and reasoning of generally accepted practices and principles of classroom assessment. Assessment Work Sample – an instrument to measure teacher assessment practice, adapted from an original instrument developed by National Center for Research on Evaluation, Standards, and Student Testing. Teachers present 4 graded student papers for 3 types of assessment, each sample is evaluated according to a rubric. CSAP – Colorado state No Child Left Behind assessment (mathematics section).

*	Rochelle, Feng, Murphy, & Mason (2016)	Y	M	USA, Maine	S	ASSISTments - web-based mathematics cognitive tutor, provides feedback and hints for problems. Developed by a non-profit organisation at the Worcester Polytechnic Institute.	Treatment group showed higher post-intervention scores (controlling for pre-intervention math and other demographic variables), t(20) = 2.992, p = .007. The effect of the ASSISTments intervention is greater for lower-performing students than for higher-performing students (the interaction is significant, t(2770) = 2.432, p = .015). Analyses appear very robust (multi-level HLM) and there was good attention to pre-intervention pairing followed by random assignment.	L	L, T, S	SEA If the student gets the item incorrect they receive feedback in the form of scaffolds for that item. If the student does not know the answer they can also request hints.	SS/I Information provided to the teacher is number of attempts made by student, number of hints requested, reaction time, and number of opportunities to practice. Online reports and automated emails provide reports on individual student's strengths and weaknesses, as well as performance of entire class. It is not really an explanation of the problem areas (I).	Existing skill builders in ASSISTments cover >300 topics in middle school math. Teachers can assign skill builders to students to provide practice problems that focus on a targeted skill until they reach a teacher-defined criterion for correctness (e.g. a streak of three correct answers on similar math problems). Students can be checked at 1- and 2-week intervals for retention of skills demonstrated on past problem sets, which links to the research-based instructional strategy of spaced practice (Pashler et al., 2007). For both types of content, teachers (rather than the system or intervention developers) decided how much and what type of homework was assigned, and they were asked to do so in accordance with their existing school homework policy	Not specified	Not specified	New England Common Assessment Program (NECAP) – state administered test of reading and mathematics. The authors used scores from the students' grade 6 test as pre-test. TerraNova Common Core – mathematics assessment aligned with the Common Core standards, used as post-test. System use data – electronic records within the ASSISTments system, e.g. correct responses by the student, hours of usage.

* Sumantri & Satriani (2016)	N	M	Indonesia, Jakarta	P	Formative essay tests – short-answer questions where students have to show their solution methods, designed by researchers. Multiple choice tests – designed by the researchers. Students in the self-directed learning condition also assessed their task performance and selected future tasks for improving performance (however authors do not specify how this was done).	Students in the formative assessment condition scored significantly higher on the post-test than students in the control group (F(1, 40) = 5.38, p < .05). Students with high self-directed learning also showed significantly higher scores on the post-test (F(1,40) = 27.66, p < .01) as compared to students with low self-directed learning. The interaction between condition and self-directed learning was significant, F(1,40) = 6.32, p < .05.	L	L	Not specified	NA	Not specified	Not specified	Self-directed learning which contains three elements: learners must (a) perform the tasks, (b) assess their task performance, and (c) select future tasks for improving their performance (though it is not clear from the paper how this is done)	Mathematics learning outcomes test – designed by researchers, included an essay and multiple choice questions, aligned with elementary school curriculum for fourth grade.
* Tsuei (2017)	Y	M	Taiwan	P, low-achieving Grade 3 students in remedial classes	i-GMath – a synchronous peer tutoring system on mobile tablet devices. The system was designed to provide virtual mathematics manipulatives representing students' problem solving process. A reward scheme was incorporated into the system to help retain low-achieving students' motivation to learn mathematics. Teachers can assign mathematics problems to children during the peer tutoring process. In addition, tutee can ask system for help ("please indicate the error," "give a hint", "please demonstrate the solution").The tutor has a number of tools which can be used to provide feedback to the tutee.	A significant interaction effect showed that the i-GMath group showed significant greater increases in achievement from the 1st to the 7th week, compared with the control group, F = 5.20, p < .05.	L	L and peer tutor.	Type of feedback depends on what the tutee requested, e.g. correct answer, feedback, indication of error, demonstration of solution.	NA	The remedial classes were conducted 2 times a week, for 7 weeks. It is not clear whether students undertook any other activities or instruction besides i-GMath exercises.	Not specified	Not specified.	iCBM – researcher developed curriculum based measurement model that can be run on mobile devices. The item bank included all question types presented in the students' mathematics textbooks. The teacher used the system to randomly select 10 maths problems (5 conceptual, 3 computational, and 2 application questions) from the item bank as a CBM probe. The CBM probes serve as weekly tests administered to both classes during the 7-week period.

*	van den Berg, Bosker, & Suhre (2018)	N	M	Netherlands	P	Classroom Formative Assessment (CFA) – developed by the researchers, embedded in two commonly used sets of mathematics assessments. CFA is done in cycles, whereby a teacher identifies a learning goal to be addressed during the lesson, observes each student complete a task related to learning goal and then providing instructional feedback to groups of students who did not understand the task sufficiently well. Learning goals are assessed again at the end of the week via a 8 multiple choice question quiz, allowing teachers to identify and address common misconceptions immediately after the quiz. Teachers received professional development to facilitate CFA implementation during the school year. The control group continued as usual, using half-yearly standardised tests to monitor student progress and adjust instruction.	Results indicate that the CFA teachers assessed their students' mastery of the learning goal and subsequently provided immediate instructional feedback more often during the lessons than the teachers in the control condition. However, adding teachers' participation in the CFA condition to the model as an explanatory variable did not significantly improve model fit (χ2 = .081, df = 1, p = .776), Thus, teachers' participation in the CFA condition did not improve student scores on outcome post-test.	L	L, T	SE	SS	Not specified	Not specified	CFA model consisted of four daily CFA cycles and a weekly CFA cycle incorporating three elements of formative assessment: goal setting for instruction, assessment, and instructional feedback	Pre-tests and post-tests – developed by researchers, covered same material as the curriculum for Grade 4 and 5 (separate tests). All 4 tests had 24–26 multiple choice and open-ended questions, with Cronbach's α between .81 and .84.

*	Wong (2017)	N	M	Singapore	P	Researchers developed a set of student self-assessment strategy worksheets, including checklists, learning logs, and rubrics. The materials required students to evaluate themselves on their deep understanding, strategies and reasoning, clarity, and written communication. Self-assessment criteria were explicitly explained to students. The study compared students who were trained in self-assessment and students who were not. The interventions went for 20 weeks (2 terms).	There was a statistically significant difference reported between intervention and comparison group on knowledge application, independent learning, communication and motivation over time, Pillai's Trace = .286, F (4, 141) = 14.141, p < .001, partial η2 = .29 (large effect size).	Attitudes towards and perceptions of the role of self-assessment	S	NA	NA	Not specified	Not specified	Intervention based on explicit instructions and training in self-assessment skills and criteria: (1) creating self-assessment criteria, (2) teaching the students how to apply the criteria, and (3) giving students feedback about self-assessment. This is followed by opportunities for practice.	Self-Assessment Questionnaire (SAQ) – adapted from Wong (2012) and further developed by the researchers. Included 10 Likert-scale questions for each domain (knowledge application, independent learning, communication, and motivation).
*	Wongwatkit et al (2017)	Y	M	Thailand	P	Formative assessment-based personalised web learning system – designed by researchers; students took an online conceptual pre-test and a learning style questionnaire to generate an individual learning path that matched the student's learning style. Once the student reached a set level of correctness within a unit, they could proceed to the next unit. The system shows hints to guide the student.	Learning achievement of the treatment group was significantly higher than that of the control group, F(1,58) = 6.227, p < .05, η2 = .097. There was a significant effect of learning styles (visual vs verbal), F(1,58) = 4.035, p < .05, η2 = .065, but the interaction with learning approaches (treatment vs control) was not significant.	L	L, S	SE/SEA. When the students failed to correctly answer three of five test items, the developed system showed some hints and supplementary material/tasks to guide their further learning, rather than provide correct answers.	NA	The intervention consisted of 150 minutes of activities with the online tool. The control group participated in a similar sequence of online activities for the same period of time, but their tool did not contain elements of formative assessment (only the individualised learning path).	Not specified	Yes – students are referred back to simpler pre-requisite skills. Skill hierarchy developed by teachers and experts in line with the curriculum.	Pre- and post-test – designed by 3 experimented mathematics teachers, included 10 multiple choice questions to evaluate students prior knowledge of circle area content (pre-test) and their achievement in the topic (post-test). Chronbach's alpha = 0.84 (unclear if for both tests). Questionnaire – adapted from Liaw and Huang (2013), measures students' perceptions of the learning approach.

* Wu, Kuo & Wang (2017)	Y	M	Taiwan	P	Dynamic assessment module – the learning path is adjusted for each student depending on their pattern of in/correct responses. The system can also provide instruction prompts and the process for solving the question to aid the student as necessary. There were two experimental groups: dynamic individualised assessment and individualised instruction group (DIA_II), and instruction group (nDIA_II). The students in the DIA_II group received prompts according to the options they chose and the students in the nDIA_II group received direct instructions, both using the dynamic assessment system. In the control group, cognitive skills were taught by the teachers in sequence, based on the group report of the pre-test.	All 3 groups showed significant improvement from pre- to post-test. Taking into account pre-test performance, the DIA_II group performed significantly better than the other 2 groups (which did not differ from each other), p < 0.5.	L	L, T	SEA Student receives prompts and hints on individual items. Student also sees their own profile and is provided with links to additional tutorials to support areas of need.	SS Teacher sees student profile so would know which areas might need additional instructional focus.	The dynamic assessment tool is formed based on one mathematics unit (addition and subtraction of fractions with different denominators). By analysing teaching materials and objectives, the important cognitive skills of this unit were defined by the experts.	Not specified	Yes - students are referred back to simpler pre-requisite skills. Skill hierarchy developed by teachers and experts in line with the curriculum.	Pre- and post-tests – developed by researchers; contained 20 items. Students were required to record the problem solving process in both their pre-test and post-test. The average difficulty index and the average discriminate index were 0.78 and 0.51, respectively, with Cronbach's alpha = .93. Items in the tests did not show up in the dynamic assessment adaptive system.

* Yang et al (2015)	N	M	Taiwan	P	Reciprocal peer-tutoring-enhanced mathematical communication (RPTMC) activity – involves creating, reciprocal peer tutoring, revising, and staging; the activity aims to improve students' mathematical communication ability. The assessment consisted of multi-step word problems, including continuous addition, mixed addition and subtraction, mixed addition and multiplication, and mixed subtraction and multiplication. The control variable was their daily learning approach. In other words, both groups had the same mathematics learning time in the same one-to-one self-learning mathematics environment. However, the control group practised mathematics by teacher-led instruction for solving various word problems, while the experimental group participated in the RPTMC activity to solve related word problems chosen by the teacher and researchers.	Experimental group showed significant improvement in total score for math communication (t(24) = –7.64, SE = .89, p < .001), whereas the control group showed no significant improvement (t(25) = –.27, SE = 1.00, p = .79). Improvement seen in all 3 sub-domains of math communication suggesting that sufficient practices on finding solutions and explaining them through writing/drawing and verbal forms may assist students in expressing their own mathematical concepts and understanding others' mathematical thought.	None - measuring math communication skills which may potentially help understand gaps in knowledge.	L	Learner gets verbal feedback from the peer tutor and is able to revise the response.	NA	RPTMC activities were conducted 13 times over the course of a semester, for 80 minutes every week.	Yes – the activity requires expressive communication skills, metacognition.	Not specified	Pre- and post-test – designed by researchers, both used parallel problems within Grade 2 Mathematics curriculum. Each question represented one mathematical communication sub-ability, and each sub-ability included two to three criteria to test different evaluative approaches. To ensure the reliability of the assessment, two raters evaluated the assessment independently. The inter-rater reliability of the pre-test was 0.912, p < .05, and that of the post-test was 0.905, p < .05.

11.6 Professional development

Authors	Online tool (Y/N)	Domain A = Arts M = Mathematics PD = Professional Development R = Reading S = Science W = Writing	Location of study	Sample characteristic P = Primary S = Secondary Typical sample vs atypical sample. If atypical then describe characteristics	Form of formative assessment (description of the task used) What is the source of the assessment tool used? Who is the author? (e.g. classroom teacher, school-based learning community, assessment expert working with teachers, ready-made package (standardised/non-standardised).	Impact of the formative assessment on student learning outcomes (cite measures of impact here)	What is being measured? L = Learning (meets L/O) PR = Progress G = Gaps S = Specific difficulties R = Reasons for difficulties (cognitively diagnostic/task diagnostic)	Who is the feedback to? L = learner T = teacher S = software Who's behaviour is expected to change as a result of this feedback?	Type of feedback to learner: NA = Not applicable S = Score/grade provided only SF = Score/grade & feedback re: correct answer SE = Explanation of the difference: correct results & explanation of differences between their result and the correct result; SEI = Explanation and improvement suggestions: As above but now students also receive some specific suggestions for improvement; SEA = Explanation and specific activities: Students are given information about the correct results, some explanation, and specific activities to undertake.	Type of feedback to teacher: NA = Not applicable S = Overall score only SS = Separate scores provided for specific aspects of performance I = Possible explanation of the problem areas and suggestions for additional instructional focus A - Possible explanation of the problem areas and specific instructional activities to undertake.	Evaluation is based on theoretically valid TASK Model? Sequence of activities that need to be successfully completed to meet learning outcomes and how learners typically progress through them (learning progression)	The intervention is based on theoretically valid COGNITIVE Model? Model of prerequisite cognitive and learning skills underlying successful progression. e.g. Does the process require a significant amount of working memory, attention, motivation, persistency, cognitive ability, language skills etc.	Are the actions/interventions following the assessment task evidence-based? (i.e. is the INSTRUCTIONAL model valid?)	What tools/resources are used in the assessment process and intervention? (Could be teacher designed or commercial).

* Andersson & Palm, 2017;	N	M, PD	Sweden	P	A researcher-designed PD program, aimed at preparing teachers to integrate a range of formative assessment strategies into their instruction, based on the framework of Wiliam and Thompson (2008). This approach highlights a mutual understanding of learning goals and criteria between teacher and student, the teacher eliciting evidence of student understanding and skills through classroom discussion, providing feedback and adjusting instructional activities. The teacher supports self-regulated learning and collaboration between students.	ANCOVA showed that, adjusting for the pre-test scores, the treatment group scored significantly higher at post-test than the control group, F(1,42) = 4.71, MSE = 7.74, p = .036, Cohen's d = 0.66.	Not specified	L, T.	Not specified	Not specified	Participants attended lectures on theory and implementation of FA, as well as engaged in discussions around their attempts to implement FA. This stage took 144h in total, with an additional 72h being available to read the literature, plan and reflect on FA activities.	The PD program operationalised formative assessment according to the framework of Wiliam and Thompson (2008).	Elements of FA are embedded in the instruction process. All teachers reported implementing between 8 and 34 FA activities (median = 20).	Pre-test – researcher-designed and aligned with national curriculum for mathematics. Designed to reflect attainment of Year 3 learning goals. The test was done in two parts, with 29 and 22 items respectively, questions included multiple choice, fill-in-the-blanks, and short answer. Post-test – designed similarly to pre-test, content covering Year 4 curriculum. There were two parts with 23 and 13 items, with fill-in-the-blank and short answer questions (no multiple choice). Teacher interviews and questionnaires – conducted to evaluate teacher experiences with the PD program.
* Brookhart, Moss & Long 2010	N	R, PD	USA, mid-Atlantic state	P (K and Y1 at-risk readers)	Professional development program – taught different formative assessment practices, including letter cards, customised letter-naming drills, keeping records of feedback given to students, etc. Control condition – business as usual.	No effect on DIBELS Letter Naming Fluency in K students whose teachers underwent PD (partial η2 = .001); large effect on Y1 Phoneme Segmentation Fluency (partial η2 = .036)	L	T	NA	Not clear as different teachers used different assessments, their frequency of use not specified.	Specific sequence of activities not specified. Teachers underwent a PD program, but were not monitored for using specific formative assessments.	Yes	Teachers reported changes to instruction and differentiation, but specific actions following the assessment task are not specified in the paper.	Dynamic Indicators of Basic Early Literacy Skills (DIBELS) (Good and Kaminski 2002) – six individually administered standardised measures of early literacy development, this study administered phoneme segmentation fluency measure and letter naming fluency measure.
* Chen, Lui, Andrade, Valle & Mir (2017)	N	A, PD	USA, New York	P, S	Criteria Referenced Formative Assessment (CRFA) – developed by experts as part of Arts Achieve project by NY Department of Education. Participating teachers received professional development focusing on effective use of criteria-referenced peer and self-assessment strategies. Only teacher's with sufficiently high self-reported treatment fidelity (based on implementation logs) were included in the study. Control group – business-as-usual instruction.	After matching treatment students to controls with similar scores on 12 demographic variables, control and treatment post-test scores were compared. The effect is significant (t (610) = 5.10, p = .00), with Cohen's d = .26 (95% CI = [.15, .37]).	Not specified	Not specified	Not specified	Not specified	Not specified	Not specified	Not specified	Benchmark Arts Assessments-Theatre Arts (BAATA) – used as pre- and post-tests. Developed in alignment with the New York City Department of Education Blueprints for Teaching and Learning in the Arts and the Common Core Capacities in English Language Arts. Includes multiple choice questions, constructed responses, and performance tasks. Implementation logs – filled out by teachers every 2-3 weeks to document teachers' use of treatment components.

* Fantuzzo, Gadsden, & McDermott (2011)	N	M, PD	USA, Philadelphia	Pre-K	Evidence-Based Program for Integrated Curricula (EPIC) intervention – includes integrated curriculum practices, curriculum-based assessment, and professional training and support. Routine experiences include interactive reading, working in small and large groups, transition activities, environmental changes, and home connections. EPIC Integrated Check-Ins (ICIs) – brief assessments of skill levels across the integrated scope and sequence of the curriculum. Skills include alphabet knowledge, phonemic awareness, vocabulary, print concepts, listening comprehension, mathematics, motor, social-emotional, and approaches-to-learning skills.	At the end of the 2-year long intervention, the treatment group showed significant improvements on listening comprehension and mathematics but not on alphabet knowledge or vocabulary, as compared to the control group. For listening comprehension, d = .17, and for mathematics, d = .22.	S	T	NA	SS	PD for the treatment group uses a learning community model based on distributed leadership principles (Spillane, 2006). The EPIC learning community meets routinely throughout the year in teaching teams, small groups, and large group. PD included reviewing children's responses to curriculum activities from the previous week and planning future activities. Teachers in the control group received PD in the form of didactic workshops and used the Preschool Child Observation Record (High/Scope Educational Research Foundation, 2003) to conduct individual assessments of children and monitor their progress.	The intervention includes 4 learning behaviour modules (attention control, frustration tolerance, group learning, and task approach) that are integrated into the activity sequence of the 8 curriculum units.	Formative assessments are embedded in units as standardized curriculum activities that are repeated three times throughout the year. They help teachers monitor children's progress and create a classroom profile of individual student differences in ability levels to inform instruction.	Learning Express (McDermott et al., 2009) – an individually administered adaptive battery referenced to Head Start's National Indicators and Prekindergarten Pennsylvania Learning Standards for Early Childhood. The test includes four subscales: alphabet knowledge, vocabulary, listening comprehension, and mathematics.
* Gallagher, H.A, Arshan, N., &Woodworth, K.	Y	W, PD	USA, 10 states – rural districts	S	Using Sources Tool – an online portal that allows teachers to look at and assess student work, guiding teachers with a series of prompts to analyse student writing. The tool is also designed to further educate teachers by helping them identify qualities of effective arguments in their students' writing. Teachers also underwent an extensive professional development intervention over two years of this study (College-Ready Writers Program, CRWP).	CRWP group had significantly higher scores on 3 out of 4 AWC-SBA attributes (content, structure, stance). The impact estimate on the content and structure attributes is reported as 0.2 with p < .05 and for the stance attribute the impact estimate is reported as 0.15, p < 0.05.	L, G.	T	NA	I	Yes – the stated goal was to increase student writing proficiency in alignment with the new college- and career-ready standards in English language arts, and mathematics with use of supporting curricular resources.	Not specified	Formative assessment for the teachers embedded in the larger professional development intervention.	Analytic Writing Continuum for Source-Based Argument (AWC-SBA) – a measure to evaluate student writing, developed by the National Writing Project. Professional development monitoring form – a log to document professional development events and teacher participation. Teacher log and survey – administered in spring and autumn every day for 2 weeks. Teachers recorded time spent writing, length of writing assigned, and the purposes for writing that day.8 The survey measured broader practices and constructs more appropriately measured over a year than in a single day.

* Phelan et al. (2012)	N	M, PD	USA	P	Checks for Understanding – a sequence of formative assessment activities designed by researchers, same as in Phelan et al. (2011). The focus of the study was POWERSOURCE© – researcher-developed professional development program and supplementary materials designed to support the implementation of Checks for Understanding in four domains of algebra (rational number equivalence, properties of arithmetic, principles of solving linear equations, applications to other areas of mathematics).	There was a statistically significant difference between experimental and control groups across all four domains of algebra, short and long-response items, and total item scores. Broken down by domains, the largest effect size (1.25 SD) was for principles of arithmetic, followed by applications of core principles to other domains (0.89 SD), solving linear equations (0.81 SD), and for rational number equivalence (0.73 SD). These results suggest that students whose teachers completed the POWERSOURCE program outperformed control students substantially.	L	T. Teachers expected to become more proficient in their subject matter knowledge, more skilled in their formative use of assessment, and better equipped to focus their instruction on key ideas.	NA	SS	Not specified	Not specified	Checks for understanding are meant to be followed up with additional instruction by the teacher if needed, focusing on the specific aspect of understanding (e.g. conceptual, application to a larger problem).	State standard test – used as pre-test; 2005-2006 data was used, from the test administered prior to pilot test year; Checks for understanding – scores were used as the outcome measure, as the main focus of the study was POWERSOURCE© (the professional development program + instructional materials)
* Randel et al. (2016)	N	M, PD	USA, Colorado	P	The Classroom Assessment for Student Learning (CASL) – a professional development (TPL) program developed by the South Carolina Department of Education. Covers assessment purposes, accuracy of assessment, and using assessment results.	The CASL schools' adjusted mean on the CSAP Mathematics test was 502.49 (SE = 2.53), compared to an adjusted mean of the control schools of 501.91 (SE = 2.44) with an adjusted difference of 0.58 that was not statistically significant (SE = 3.47, p > .05).	PR	T Teachers' behaviour expected to change as a result of participating in the PL program; no involvement of any external facilitator, but teachers hold team meetings every 2-3 weeks to work through the Handbook, chapter by chapter	NA The professional development program includes ideas for teachers on how to provide feedback to students.	NA	Not specified (main focus is on the TPL program).	Not specified	Not specified	CASL implementation logs – teachers completed brief logs to describe their study of CASL materials throughout the year. At the end of the year, teachers described how they implemented CASL in the classroom throughout the year. Test of assessment knowledge – researcher-developed, 60 items to test teachers' knowledge and reasoning of generally accepted practices and principles of classroom assessment. Assessment Work Sample – an instrument to measure teacher assessment practice, adapted from an original instrument developed by National Center for Research on Evaluation, Standards, and Student Testing. Teachers present 4 graded student papers for 3 types of assessment, each sample is evaluated according to a rubric. CSAP – Colorado state No Child Left Behind assessment (mathematics section).

* Reddy et al., 2017	N	PD	USA, New Jersey and New York	P	Data-driven Classroom Strategies Coaching (CSC) – researcher-designed professional development model aiming to improve teachers' use of specific evidence-based instructional and behavioural management practices in the classroom. Core components of the model include: integration of instruction and classroom behavioural management, FA with a classroom observation instrument, brief and structured problem-solving framework, establishment of measurable goals, provisions for modelling and practice, and visual performance feedback for the teachers.	No student learning outcomes were measured in this study. Teachers in the treatment condition significantly improved their use of targeted strategies taught in the program, as compared to teachers in the control condition (effect size = .54).	NA	T	NA	SS Teachers are formatively assessed on their teaching practices.	CSC constituted a sequence of four 30 minute meetings held over the course of four weeks, which included problem/needs and goal identification, plan development, plan implementation, and evaluation of the teacher.	Not specified	Not specified	CSAS – classroom observation instrument to aid FA, used as pre- and post-test in this study. Includes two forms: (a) Observer Form (CSAS-O), which can be used by instructional coaches for observing class- rooms, and (b) a teacher self-report form (CSAS- T) to self-evaluate their practice and progress throughout the coaching process. Designed by researchers in a previous study. The two forms produce visual feedback (graphs) about teachers' use of strategies during lessons and their change over time.
* Roschelle et al., 2010	Y	PD	USA, Texas	S	SimCalc – an approach which integrates an interactive representational technology (MathWorlds), paper curriculum, and teacher professional development. MathWorlds – commercial software that creates motion animation for mathematical functions created by students. MathWorlds is commonly used as an aid in FA, where the student is asked to tell stories that correspond to function and/or animation.	Overall, authors report significant differences in gain scores between experimental and control groups in all three experiments presented in this study. For Year 7 experiment (treatment vs control), the effect size of gain score difference was .63, p < .0001. For Year 8, the effect size was .56, p < .0001. In the Year7 delayed-treatment quasi-experiment (control teachers from Year 7 started the treatment in Year 8, while treatment teachers continued with the treatment), the reported effect size of gain score difference was .50, p < .0001. Authors do not report what effect size statistic they used.	L	Not specified	Not specified	Not specified	Treatment teachers attended a 3-day summer workshop introducing the respective SimCalc replacement units. The teachers worked through the SimCalc materials, and were taught techniques to prompt through exploration of mathematical ideas. The treatment replaced 1 mathematics unit in each grade (rate and proportionality for Year 7, linear function for Year 8), taught with the aid of MathWorlds software.	Not specified	The unit taught with the aid of MathWorlds was designed to be taught daily over 2-3 weeks. Further details are not specified.	Year 7 Mathematics measure – researcher designed test of mathematics achievement, reviewed by a panel of experts and piloted in the field. The test is aligned with the state curriculum and content standards, includes 30 items (Cronbach's α = .86). Year 8 Mathematics measure – designed similarly to the Year 7 measure, 36 items (Cronbach's α = .91).

* Shechtman et al., 2010	Y	PD	Finland	S	SimCalc – an approach which integrates an interactive representational technology (MathWorlds), paper curriculum, and teacher professional development. MathWorlds – commercial software that creates motion animation for mathematical functions created by students. MathWorlds is commonly used as an aid in FA, where the student is asked to tell stories that correspond to function and/or animation.	The only significant relationship between teacher MKT and student gains was in the Year 7 experiment, whereby teachers' MKT pre-test scores were a significant predictor of student M2 gains in the treatment group, β = 0.13, z = 2.6, p < .01. Note: M2 refers to the subscale of the student achievement test that contained items going beyond the basic state curriculum.	L	Not specified	Not specified	Not specified	Treatment teachers attended a 3-day summer workshop introducing the respective SimCalc replacement units. The teachers worked through the SimCalc materials, and were taught techniques to prompt through exploration of mathematical ideas. The treatment replaced 1 mathematics unit in each grade (rate and proportionality for Year 7, linear function for Year 8), taught with the aid of MathWorlds software.	Not specified	The unit taught with the aid of MathWorlds was designed to be taught daily over 2-3 weeks. Further details are not specified.	Year 7 Mathematics measure – researcher designed test of mathematics achievement, reviewed by a panel of experts and piloted in the field. The test is aligned with the state curriculum and content standards, includes 30 items (Cronbach's α = .86). Year 8 Mathematics measure – designed similarly to the Year 7 measure, 36 items (Cronbach's α = .91). Mathematical knowledge for teaching test (MKT) – designed similarly to the achievement tests. Version for Year 7 knowledge assessment had 24 items (Cronbach's α = .80), Year 8 version had 28 items (Cronbach's α = .80).
* Smit et al., 2017	N	PD	Switzerland	P	Rubric for mathematical reasoning – primarily as a tool to increase transparency of assessment criteria. The paper does not specify how the rubrics were used in the classroom or the content of the rubric. Teachers in the treatment group underwent professional development and followed a prescribed lesson plan, but no further details are specified.	The intervention did not have a significant effect of post-test scores (β = .04, p > .05), and the strongest predictor of post-test scores were the pre-test scores of mathematical reasoning (β = .96, p < .05).	Unclear	L, T.	Not specified	Not specified		Not specified	Formative assessment embedded in a larger intervention.	Mathematical reasoning test – items either developed or adapted by researchers from other standardised tests, aligned with Swiss national basic competencies. The test contained 18 open-ended questions, distributed over 2 testlets with 10 items each. Questionnaire for teachers – researcher-developed self-report measure reflecting the use of formative assessment in the classroom, as well as peer- and self-assessment. Questionnaire for students – the same questions as in the teacher questionnaire, intended to verify teachers' self-report.

*	van den Berg, Bosker, & Suhre (2018)	N	M, PD	Netherlands	P	Classroom Formative Assessment (CFA) – developed by the researchers, embedded in two commonly used sets of mathematics assessments. CFA is done in cycles, whereby a teacher identifies a learning goal to be addressed during the lesson, observes each student complete a task related to learning goal and then providing instructional feedback to groups of students who did not understand the task sufficiently well. Learning goals are assessed again at the end of the week via a 8 multiple choice question quiz, allowing teachers to identify and address common misconceptions immediately after the quiz. Teachers received professional development to facilitate CFA implementation during the school year. The control group continued as usual, using half-yearly standardised tests to monitor student progress and adjust instruction.	Results indicate that the CFA teachers assessed their students' mastery of the learning goal and subsequently provided immediate instructional feedback more often during the lessons than the teachers in the control condition. However, adding teachers' participation in the CFA condition to the model as an explanatory variable did not significantly improve model fit ($\chi2 = .081$, $df = 1$, $p = .776$), Thus, teachers' participation in the CFA condition did not improve student scores on outcome post-test.	L	L, T.	SE	SS	Not specified	Not specified	CFA model consisted of four daily CFA cycles and a weekly CFA cycle incorporating three elements of formative assessment: goal setting for instruction, assessment, and instructional feedback	Pre-tests and post-tests – developed by researchers, covered same material as the curriculum for Grade 4 and 5 (separate tests). All 4 tests had 24-26 multiple choice and open-ended questions, with Cronbach's α between .81 and .84.
*	Witmer et al. 2014	N	R, PD	USA	P	Concepts of Comprehension Assessment (COCA, Billman et al., 2008) – individually administered test designed to measure first- and second-grade students' specific fundamental knowledge and skills for comprehending informational text and is intended to help inform instruction.	The treatment group had significantly higher COCA scores at half-point ($F(1, 120) = 17.14$, $p < .025$, partial $\eta2 = 0.13$), and at the end of the year ($F(1, 120) = 16.68$, $p < .025$, partial $\eta2 = 0.12$). The treatment group also had significantly higher scores on the transfer writing measure at the end of the year, $F(1, 108) = 9.25$, $p < .01$, partial $\eta2 = .08$.	PR	T	NA	SS	Yes	Not specified	Not specified, as individual teachers could change instruction as desired. Self-report measures indicate that many of the participating teachers changed their writing activities and instruction as a result of their experiences with COCA	PD for teachers on how to use and interpret COCA results (14.5 hours), experimenter designed. Prompted writing samples – students independently write for 30 minutes about the topic covered by their form of COCA assessment. Scored with a rubric. The main outcome measure is change in COCA assessment scores throughout the year.

* Yin et al., 2015	N	PD	USA, Hawaii	S	FA-then-NAV group – received PD focusing on strategies for formative assessment in Year 1 and received PD on implementing formative assessment using TI-Navigator in Year 2. FA-and-NAV group – received PD on strategies for implementing formative assessment using TI-Navigator in both Years 1 and 2. TI-Navigator – developed by Texas Instruments; a wireless system to connect students' graphic calculators to teacher's computer. Allows teachers to easily distribute formative assessment questions and receive/display student answers.	The study did not assess student outcomes. In regard to teacher outcomes, FA-then-NAV group exhibited significant (p < .05) self-reported increases in knowledge about general assessment and FA, self-efficacy in FA, value of technology, and confidence in class technology. There was a significant decrease in interest in technology and no significant difference in self-efficacy with general technology. For FA-and-NAV group, there were significant (p <.05) increases in knowledge about FA, value of technology, and confidence in class technology. Differences on other subscales were not significant.	S	L, T.	Not specified	Not specified	Teachers received professional development during the summer break, aimed to highlight distinctions between formative and other types of assessment. After receiving PD specific to TI-Navigator, all teachers were provided with laptop computers, liquid crystal display (LCD) projectors, document cameras, and a classroom set of TI-73 calculators.	Not specified	Not specified	Researcher-developed surveys administered to the teachers as outcome measures, addressed the school environment (teacher collaboration and support), assessment survey (knowledge and self-efficacy in using FA), technology survey (teachers' beliefs about using technology), and a PD evaluation survey.