

## Review

# A systematic review of prerequisites for implementing assessment for learning in classroom practice



M.C. Heitink <sup>a,\*</sup>, F.M. Van der Kleij <sup>b</sup>, B.P. Veldkamp <sup>a</sup>, K. Schildkamp <sup>a</sup>,  
W.B. Kippers <sup>a</sup>

<sup>a</sup> University of Twente, Postbus 217, 7500 AE, Enschede, The Netherlands

<sup>b</sup> Australian Catholic University, 1100 Nudgee Rd, Banyo, QLD 4014, Australia

## ARTICLE INFO

## Article history:

Received 25 June 2015

Received in revised form 26 November 2015

Accepted 1 December 2015

Available online 17 December 2015

## Keywords:

Assessment for learning

Formative assessment

Systematic review

Classroom practice

## ABSTRACT

Although many researchers acknowledge that Assessment for Learning can significantly enhance student learning, the factors facilitating or hindering its implementation in daily classroom practice are unclear. A systematic literature review was conducted to reveal prerequisites needed for Assessment for Learning implementation. Results identified prerequisites regarding the teacher, student, assessment and context. For example, teachers must be able to interpret assessment information on the spot, student engagement in the assessment process is vital, assessment should include substantial, constructive and focussed feedback, and the school should have a school-wide culture that facilitates collaboration and encourages teacher autonomy. The results of this review contribute to a better understanding of the multiple facets that need to be considered when implementing Assessment for Learning, from both a theoretical and a practical standpoint.

© 2015 Elsevier Ltd. All rights reserved.

## Contents

1. Introduction .....	2
2. Methods .....	3
2.1. Procedure .....	3
2.2. Databases and search terms .....	3
2.3. Selection process .....	4
2.4. Data extraction and data analysis .....	4
3. Results .....	5
3.1. Teacher .....	5
3.1.1. Teacher knowledge and skills .....	5
3.1.2. Teacher beliefs and attitudes .....	7
3.2. Assessment .....	7
3.2.1. Assessment content and presentation .....	7
3.2.2. Alignment and integration .....	8
3.3. Context .....	8
3.3.1. Leadership and culture .....	8

\* Corresponding author.

E-mail address: [m.c.heitink@utwente.nl](mailto:m.c.heitink@utwente.nl) (M.C. Heitink).

3.3.2.	Support and professional development .....	9
3.4.	Student .....	9
3.4.1.	Student knowledge and skills .....	9
3.4.2.	Student beliefs and attitudes .....	9
4.	Conclusions and discussion .....	9
4.1.	Prerequisites for the effective implementation of AfL in the classroom .....	9
4.2.	Implications for practice .....	11
4.3.	Limitations .....	11
4.4.	Implications for further research .....	11
	References .....	12

## 1. Introduction

Assessment plays a crucial role in education. A distinction is often made between formative and summative purposes of assessment. Where summative assessment primarily focuses on assessing learning outcomes, formative assessment aims to gain insights into learning processes that can be used to support learning through tailored instruction and targeted feedback (Stobart, 2008).

Formative assessment has been on policy agendas internationally for decades, but implementation has proven to be challenging (e.g., Birenbaum et al., 2015; Marshall & Drummond, 2006). Although many researchers acknowledge that formative assessment can have a positive effect on learning, the proof for this is based on limited sound scientific evidence (Bennett, 2011). Moreover, the differing conceptualisations of formative assessment have led to a wide variety of practices, and it is unclear which factors facilitate or hinder its implementation. The purpose of this review was to get a sense of the prerequisites needed for the implementation of ongoing formative assessment practice that has the potential to support learning in the classroom. This study focuses specifically on a formative assessment approach called 'Assessment for Learning' (AfL) (Assessment Reform Group, 1999), in order to gather information from studies that look at relatively consistent underlying principles and intentions that shape formative assessment uses. In the following paragraphs, the key terminology used throughout this paper is clarified and the research background is described.

The literature includes a wide range of definitions of formative assessment, each having different strategies for using evidence from assessment to enhance learning and with differing emphases on social dimensions (see Brookhart, 2007 for an overview of definitions). For example, a phrase often used in formative assessment literature is the use of assessment evidence to provide feedback to "close the gap" between students' current performance and the goal (Sadler, 1989). Definitions of formative assessment differ with respect to, for example, the specific roles of not only teacher but also the student in this process as receivers, users and providers of feedback.

Three distinct approaches have evolved over time, namely: 'data-based decision-making' (DBDM), 'diagnostic testing' (DT), and 'assessment for learning' (AfL) (Van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015). DBDM involves the systematic collection and analysis of data to inform decisions that focus on improvement of teaching, curricula and (school) performance (Schildkamp & Kuiper, 2010). DT concerns the 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 (Crisp, 2012). AfL is an approach to formative assessment that occurs as part of ongoing classroom practices (Klenowski, 2009), that is viewed as a social and contextual event and that focuses on the quality of the learning process (Stobart, 2008). Feedback is continually incorporated in this process to guide future learning, and is aimed at the class or individual level. Students play a vital role in AfL and are expected to engage in assessing their own and their peer's learning (Elwood & Klenowski, 2002). A major long-term goal of AfL is to foster student autonomy by helping students learn how to learn (Black, McCormick, James, & Pedder, 2006; James et al., 2007; Pedder & James, 2006).

The publications of Black and Wiliam (1998a, 1998b) on formative classroom assessment were followed by a boost in research on formative assessment, especially work on AfL, with researchers reporting effects of AfL implementation in many countries. Much of this research has been centred around the five key strategies for implementing AfL identified in Black and Wiliam's and William and Thompson's (2007) work:

1. Clarifying and sharing learning intentions or goals and success criteria;
2. Generating opportunities to effectively gather evidence of student learning through informal and formal assessment, e.g., through classroom discussions, questioning or learning tasks;
3. Providing formative feedback to students to support their learning;
4. Supporting students in acting as instructional partners through discussion and peer assessment; and
5. Activating students as agents in their own learning through self-assessment and self-regulation.

These five key strategies are based on the central notion of using assessment evidence to inform learning. They have been interpreted in various ways, and numerous researchers have emphasised the need for deep engagement with these principles

in order to achieve the ultimate goals of AfL; promote deep learning and learner autonomy (Hayward, 2014; Marshall & Drummond, 2006; Pedder & James, 2006).

AfL can be approached from a measurement perspective or an inquiry perspective (Hargreaves, 2005). When approached from a measurement perspective, AfL is characterized by the use of formally gathered (quantitative) evidence about student learning to formulate feedback and to inform decisions, based on assessment activities that aim to determine whether, or to what extent, a pre-set level of performance has been achieved. Approaching AfL from an inquiry perspective results in the use of primarily qualitative information (e.g., observations, demonstrations and conversations) to generate feedback, in a process of discovery, reflection, understanding and review. This perspective acknowledges the power of social interaction and student autonomy in enhancing student learning, and is more congruent with current understandings of AfL (e.g., Klenowski, 2009; Ruiz-Primo & Furtak, 2006). Researchers have emphasised that quality implementation of AfL requires adopting an inquiry approach to AfL (Hargreaves, 2005; Wyatt-Smith, Klenowski, & Colbert, 2014) and an in-depth engagement with the five key strategies (Black & Wiliam, 1998a; 1998b; William & Thompson, 2007) by both teachers and students, as an integral aspect of daily classroom practice (Pedder & James, 2006).

Although there is evidence that AfL can help students learn, a number of studies show no to little effects on student learning. For example, in a meta-analysis that studied the effects of formative assessment on student achievement, Hendriks, Scheerens, and Slegers (2014) concluded that most studies found little to no effects. This is likely due to the ineffective implementation of formative assessment approaches, such as AfL (Bennett, 2011). Engaging deeply with the underlying ideas of AfL has proven to be challenging for many teachers, for example as a consequence of constraints imposed by the particular policy context (Marshall & Drummond, 2006) through, for instance, accountability pressure.

No systematic analysis has been conducted on evidence gathered from studies of AfL and identifying factors that contribute to or hinder implementation has not been a primary focus of any review study published so far. Therefore, this review study focused on identifying relevant prerequisites for effective AfL implementation. In order to systematically gather these data from selected studies, four categories often distinguished in school evaluation literature (e.g., Mandinach & Jackson, 2012; Schildkamp & Kuiper, 2010) were used: the teacher, the student, assessment and the context. AfL literature emphasises the crucial roles of both the teacher and student in teaching, learning and assessment (Elwood & Klenowski, 2002; Pedder & James, 2006). The category assessment includes the means by which evidence is gathered about student learning, this covers both assessment instruments (e.g., a test or learning task) and processes (e.g., questioning and classroom discussion). The category context includes both factors internal to the school (e.g., leadership) and factors external to the school (e.g., educational policy). In practice, a sound implementation of AfL would require a balance among factors in these interrelated categories. This review was guided by the following research question: *Which prerequisites regarding the teacher, student, assessment and context need to be considered when implementing AfL in the classroom?*

## 2. Methods

The following paragraphs describe the methods used to conduct this review. In the literature, the terms AfL and formative assessment are often used interchangeably (Van der Kleij et al., 2015). Moreover, the specific term 'AfL' is not always used and operationalisations of AfL might differ throughout the research. Therefore, the literature review initially focused more broadly on studies of implementation of formative assessment in the classroom. However, the selection of studies specific to AfL suitable for identifying the prerequisites for effective implementation was the ultimate focus of this review.

### 2.1. Procedure

This review took an approach used in systematic review studies in the social sciences (Petticrew & Roberts, 2006). This stepwise process encompassed formulating research questions, defining search terms, selecting databases, conducting the literature search, formulating inclusion criteria and applying these to selected relevant literature, and the extraction of data. A library professional was consulted to ensure effective strategies for the literature search. A data extraction form was used to collect similar data from each publication. Additionally, the scientific quality of each publication was assessed; only studies that met the minimally satisfactory quality requirements were selected. Initial literature searches were conducted in March, 2014.

### 2.2. Databases and search terms

Five scientific databases were used to retrieve relevant literature: Education Resources Information Center (ERIC), Web of Science, Scopus, PsychINFO and Picarta. These databases were chosen as containing most publications regarding educational research. The same search strategy was used for every database. The search process started with broad search terms such as "formative assessment", "assessment for learning", and related terms as found in the thesaurus. Due to the increased popularity of formative assessment following the 1998 publications by Black and Wiliam (1998a; 1998b), the search was limited to publications after 1998. Also, words related to "feedback" and "classroom" were added to the search terms, as feedback is a crucial part of formative assessment and this review specifically focuses on the classroom level of implementation. Finally, the search results were narrowed down by selecting only publications in the English language.

### 2.3. Selection process

All publications were exported to Thomson Reuters Endnote X7 (2013). After removing duplicates, title and abstract scans were conducted using the following inclusion criteria:

- 1) The study was published in a scientific, peer-reviewed journal or was a dissertation.
- 2) The study involved empirical research.
- 3) The study was conducted in the context of primary, secondary or vocational education.
- 4) The study focused on the use of formative assessment in classroom practice.

The first criterion aims to select studies of adequate scientific quality. Studies that were not published in a peer-reviewed scientific journal (e.g., in books, book chapters and conference proceedings) were not selected. The quality of the remaining studies was investigated more comprehensively during data extraction. Through the second criterion, theoretical articles, reviews and opinion pieces were excluded, as this review focuses on evidence-based factors that have proven to be influential for the implementation of formative assessment in practice. Although it is necessary to be careful in generalizing from the results of case studies, case studies were included in the selection for their practical examples of 'real life' contexts, an essential element in formative assessment, particularly in AfL. Furthermore, in a recent large-scale review study the majority of AfL studies were categorized as small-scale case studies (Baird, Hopfenbeck, Newton, Stobart, & Steen-Utheim, 2014). The third criterion specifies the educational context of studies that were included in the review; primary, secondary and vocational education. We included these contexts based on practitioners' requests for more guidance on how to implement AfL. The fourth criterion restricted the selection to studies that focused on the use of formative assessment in classroom practice in general, which was essential in order to identify prerequisites for effective implementation of AfL in the classroom.

If it was unclear whether or not the study fully satisfied the inclusion criteria, the publication remained in the selection. After the title and abstract scan, full-text versions of the remaining publications were obtained. Only full text versions that were available through the library facilities at hand were included. When it became clear during data extraction that a study did not match the inclusion criteria after all, the publication was removed retroactively.

### 2.4. Data extraction and data analysis

Each of the selected publications was read completely, and relevant results were recorded using a data extraction form. The data extraction form was generated over multiple trials to ensure usability and consistency in the data extraction procedures across researchers. Additionally, a coding instruction document for using the data extraction form was created in order to enhance reliability across different members of the research team. The data extraction form consisted of the following sections:

- General information: author, publication year, title, context.
- Research design: research question, methods, instruments, primary formative assessment focus.
- Research population: number of respondents, sampling method.
- Results and conclusions: answer to the research question, identified prerequisites regarding teacher, student, assessment and context.
- Quality check: clear research goal, appropriate methods used, reliability, sample quality, quality of data analysis.

Over 50% of the studies were blindly double-coded to confirm the reliability of the data extraction process. All differences were discussed between researchers and subsequently adjusted. This process resulted in an agreement rate of 80% (Cohen's Kappa = 0.620), which is substantial (Landis & Koch, 1977).

**Table 1**  
Quality questions.

Category	Quality question
General	1. Is the research objective clear?
	2. Is the research approach in combination with the chosen method capable of finding a clear answer to the research question?
Selection sample	3. Does the study include enough data to ensure the validity of the conclusions?
	4. Is the context of the research clear (country, sampling of the schools/teachers/students)?
Method	5. Do the researchers describe the research methods used?
	6. Do the authors give an argument for the methods chosen?
	7. Do the researchers take into account other variables that might be influential?
Data analyses	8. Are the data analysed in an adequate and precise way?
	9. Are the results presented clearly?
	10. Do the researchers report on reliability and validity of the results?
Conclusion	11. Is the research question answered using empirical evidence gained from the research?

During data extraction, the selection of studies was further narrowed. Studies were classified as having a primary focus on AfL, DBDM or DT in the section on research design. Only studies classified as focusing on AfL were selected for this review.

The quality check consisted of 11 questions that were scored with 0, 0.5 or 1 point (cf. [Petticrew & Roberts, 2006](#)). The questions can be found in [Table 1](#). Based on these questions, a quality score was assigned to every study. To be considered for inclusion in this review, the overall quality score had to be at or above the cut-off score of  $0.5 \times 11 = 5.5$ . When multiple researchers coded the study, the quality scores were averaged. The total score for the quality check resulted in one of the following three decisions: a score below 5 meant the study was excluded from the review. Studies with a score between 5 and 7 were discussed between researchers, which led to a decision on inclusion. Studies with a score of 7 or higher were selected for the review.

Following data extraction, the specifics of the studies were analysed according the predetermined categories, 'teacher', 'student', 'assessment' and 'context' and the different educational settings. Within these categories, clusters were made corresponding to themes that emerged from the data. Analysis showed that only one publication in our selection considered the context of vocational education. Additional searches were conducted to collect more studies regarding this type of school. The databases were searched again with the strategy used in the initial search, but using search terms related to "formative assessment" and "vocational education". Additionally, a search using the names of key authors in this field of research was conducted.

### 3. Results

The initial search resulted in 2533 publications. After removing duplicates, conference proceedings, books, and so forth, the systematic search resulted in 1743 publications. Applying the remainder of the inclusion criteria using the information from the title and abstract scan left 117 studies available for data extraction. Because the full text documents of 15 studies could not be obtained, these studies were excluded. Upon further examination of the full text documents, another 42 studies were excluded due to a mismatch with the inclusion criteria. This process resulted in 60 publications suitable for analysis. Of the 60 publications, 26 were categorized as primarily focused on AfL. All but two AfL publications were above 7 on the quality check ( $M = 8$ ); those two studies each received 6.5 points. Only one study was conducted in the context of vocational education. The supplementary search for formative assessment studies in vocational education resulted in another 84 possibly relevant publications. However, only two additional publications satisfied the inclusion and quality criteria. Because of the limited availability of formative assessment literature of sufficient quality ( $n = 3$ ) in vocational contexts and hence the limited generalizability of conclusions based on these studies' results, vocational education was not taken into consideration in this paper. Thus, the literature search resulted in 25 relevant studies for this review.

Of these 25 studies related to AfL, nine were conducted in the context of PE (ages 4–11), ten in SE (ages 12–18) and six where SE and PE overlapped. Results included data from eleven different countries, but most of the studies were conducted in the US ( $n = 9$ ). Twelve studies used a qualitative approach, four a quantitative approach and nine studies applied a mixed methods design. An overview of the studies selected for this review can be found in [Table 2](#).

The results are structured according the predetermined categories (teacher (T), student (S), assessment (A) and context (C)), beginning with the category with the most evidence for corresponding prerequisites and ending with the category with the least evidence for corresponding prerequisites. This does not imply that prerequisites related to the last category are not important, but the results suggest there is less evidence available in the literature for the importance of these prerequisites (so far). Similar sub-categories within these predetermined categories were clustered, and eight overarching aspects important in the implementation of AfL emerged. Only aspects for which there was evidence from at least two studies were included in the results. [Table 3](#) shows the numbers of studies per prerequisite category and aspect. The study ID numbers correspond with the numbers used in [Table 2](#).

As expected, based on the relatively broad inclusion criteria, the studies in this review focused on many aspects of AfL. The studies encompassed a wide variety of research contexts, dependent variables, research methods and underlying AfL philosophies. Also various depths of AfL implementation were represented by the studies, ranging from a measurement orientation to an inquiry orientation ([Hargreaves, 2005](#)). It is important to note that this review does not aim to provide a recipe for success for any and all AfL implementation. Rather, by examining which aspects have been identified as important when implementing AfL, we hope to generate a better understanding of the multiple facets that need to be considered, from both a theoretical and a practical standpoint. Some aspects identified in this review are very general in nature and may seem crucial to any educational implementation process. Wherever such universally applicable aspects are identified, we have made efforts to specify the particular relevance for AfL. In the subsequent paragraphs, the results will be reported separately for each of the categories. As no clear differences were found between the PE and SE contexts, results for both settings are presented simultaneously.

#### 3.1. Teacher

##### 3.1.1. Teacher knowledge and skills

Seventeen studies reported results regarding teachers' knowledge and skills related to AfL. Although only four studies explicitly used the term 'assessment literacy', many studies referred to assessment literacy in general; that is, the knowledge and skills teachers need to collect, analyse and interpret evidence from assessment and adapt instruction accordingly

**Table 2**  
Overview of the selected studies.

Nr	First author and publication year	Context			Research design <sup>c</sup>	Sample size			Prerequisite category <sup>d</sup>
		Country <sup>a</sup>	Education setting <sup>b</sup>	Subject		Schools	Educators	Students	
1	Aschbacher and Alonzo (2006)	US	PE	Science	MM	n/a	25	245	T, S, A
2	Birenbaum et al. (2011)	IL	PE, SE	Humanities, Math	MM	n/a	128	22	T, A, C
3	Bryant and Carless (2010)	HK	PE	Language	QL	1	2	34	T, S, A
4	Feldman and Capobianco (2008)	US	SE	Science	QT	n/a	8	n/a	T, S, C
5	Fletcher and Shaw (2012)	AU	PE	Writing	MM	1	16	256	T, S, A
6	Fox-Turnbull (2006)	NZ	PE	Humanities	QL	6	n/a	53	T, A
7	Furtak (2012)	US	SE	Science	QL	1	6	n/a	T
8	Gamlem and Smith (2013)	NO	SE	n/a	QL	4	6	150	T, A
9	Gottheiner and Siegel (2012)	US	SE	Science	QL	n/a	5	n/a	T, C
10	Hargreaves (2013)	UK	PE	Numeracy, Literacy	QL	1	n/a	9	T, A
11	Harris and Brown (2013)	NZ	PE, SE	Language, Math	QL	1	3	99	S, A, C
12	Harris et al. (2014)	NZ	PE, SE	n/a	MM	11	13	193	T
13	Havnes et al. (2012)	NO	SE	Language, Math	MM	5	192	391	T, S, A, C
14	Kay and Knaack (2009)	CA	SE	Science	MM	6	7	213	T, A, C
15	Lee (2011)	HK	SE	Language	MM	1	4	138	T, S, A, C
16	Lee et al. (2012)	US	SE	Math, Science	MM	2	18	n/a	T, S, C
17	Newby and Winterbottom (2011)	UK	SE	Science	QL	1	n/a	157	S, A, C
18	Ní Chróinín and Cosgrave (2013)	IE	PE	Physical education	QL	n/a	5	n/a	T, A, C
19	O'Loughlin et al. (2013)	IE	PE	Physical education	QL	1	1	22	T, A
20	Penuel et al. (2007)	US	PE, SE	Science	QT	n/a	498	n/a	T, A, C
21	Phelan et al. (2012)	US	PE, SE	Math	QT	7	36	1332	C
22	Rakoczy et al. (2008)	DE, CH	SE	Math	MM	n/a	n/a	1255	T, C
23	Riggan and Olah (2011)	US	PE	Math	QL	n/a	39	n/a	A
24	Sach (2013)	UK	PE, SE	n/a	QL	3	3	n/a	T, C
25	Yin et al. (2014)	US	PE	Science	QT	1	1	52	T

Note. 'n/a' indicates information that was unavailable in the cited publication.

<sup>a</sup> Country codes according to ISO: United states of America (US), Israel (IL), Hong Kong (HK), Australia (AU), New-Zeeland (NZ), Norway (NO), United Kingdom (UK), Canada (CA), Ireland (IE), Germany (DE), Switzerland (CH).

<sup>b</sup> Primary Education (PE) (ages 4–11), Secondary Education (SE) (ages 12–18).

<sup>c</sup> Quantitative research design (QT), Qualitative research design (QL), Mixed Methods (MM).

<sup>d</sup> Teacher (T), Student (S), Assessment (A), Context (C).

**Table 3**  
Number of studies per prerequisite category and aspect.

Prerequisite category	Studies (identified by number)	Totals
<b>Teacher</b>	<b>1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 18, 19, 20, 22, 24, 25</b>	<b>21</b>
• Teacher knowledge and skills	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 18, 20, 24, 25	19
• Teacher beliefs and attitudes	2, 3, 4, 5, 10, 13, 15, 16, 18, 19, 20, 22	12
<b>Assessment</b>	<b>1, 2, 3, 5, 6, 8, 10, 11, 13, 14, 15, 17, 18, 19, 20, 23</b>	<b>16</b>
• Assessment content and presentation	1, 6, 8, 10, 13, 14, 15, 18, 23	9
• Alignment and integration	1, 2, 3, 5, 8, 11, 13, 14, 15, 17, 18, 19, 20	13
<b>Context</b>	<b>2, 4, 9, 11, 13, 14, 15, 16, 17, 18, 20, 21, 22, 24</b>	<b>14</b>
• Leadership and culture	2, 4, 11, 13, 14, 15, 16, 17, 18, 22, 24	11
• Support and professional development	2, 4, 9, 16, 18, 20, 21	7
<b>Student</b>	<b>1, 3, 4, 5, 11, 13, 15, 16, 17</b>	<b>9</b>
• Student knowledge and skills	1, 3, 5, 11, 17,	5
• Student beliefs and attitudes	3, 4, 5, 11, 13, 15, 16, 17	8

Note. Boldface sections indicate category totals.

(Birenbaum, Kimron, & Shilton, 2011; Bryant & Carless, 2010; Gottheiner & Siegel, 2012; Lee, 2011; Lee, Feldman, & Beatty, 2012).

Results of multiple studies indicated that teachers need the ability to integrate AfL with pedagogical content knowledge (PCK) to be able to cater for their students' learning needs and provide useful feedback (Aschbacher & Alonzo, 2006; Birenbaum et al., 2011; Feldman & Capobianco, 2008; Fletcher & Shaw, 2012; Fox-Turnbull, 2006; Furtak, 2012; Gottheiner & Siegel, 2012; Harris, Brown, & Harnett, 2014; Kay & Knaack, 2009; Lee, 2011; Lee et al., 2012; Penuel, Boscardin,



Masyn, & Crawford, 2007; Yin, Tomita & Shavelson, 2014). Results further suggest that teachers' pedagogical knowledge (PK) and content knowledge (CK) impact their ability to provide students with useful feedback. Without understanding a concept or without knowing common misconceptions related to a subject, teachers were not able to provide accurate and complete feedback (Furtak, 2012; Gottheiner & Siegel, 2012; Ní Chróinín & Cosgrave, 2013; Yin et al., 2014).

Teachers need to be able to create educational situations in which AfL strategies can be employed and should know the roles of students, peers and teachers in the various AfL practices. Eliciting students' thinking to reveal their learning process and common misconceptions is an important guiding strategy that teachers should master. Aschbacher and Alonzo (2006) found that using questions or directions that provide conceptual focus were most effective for eliciting students' thinking and fostering learning. This was especially true compared to overly prescriptive guidance strategies in which students were prompted to copy answers from the blackboard. Other studies specifically found that 'discussion' is a pedagogy often used in AfL. Teachers should be able to foster the participation of students in discussions about their answers, expertise or feedback (Aschbacher & Alonzo, 2006; Bryant & Carless, 2010; Feldman & Capobianco, 2008; Fox-Turnbull, 2006; Gamlem & Smith, 2013; Gottheiner & Siegel, 2012; Hargreaves, 2013; Havnes, Smith, Dysthe, & Ludvigsen, 2012; Lee et al., 2012; Penuel et al., 2007). Discussion can give teachers valuable insight into students' thinking, difficulties and understanding. This information can be used in adjusting instruction and providing feedback. Because AfL takes place in everyday classroom practice, such as during discussions, it was also noted that teachers need the ability to interpret information about student learning on the spot (Birenbaum et al., 2011; Feldman & Capobianco, 2008; Lee, 2011; Lee et al., 2012).

Moreover, teachers need knowledge and skills to develop assessments that achieve the desired purpose. This includes the ability to construct questions (used in daily classroom practice) that elicit evidence about student learning and to critically evaluate assessment instruments (Feldman & Capobianco, 2008; Gottheiner & Siegel, 2012).

The need for teachers to have knowledge and skills regarding the use of hardware and software related to computer-based assessment has also been reported as vital (Feldman & Capobianco, 2008; Lee et al., 2012). This is not specific to AfL, but is important for formative assessment in general now that ICT and assessment have become more integrated. An example specific to AfL is when teachers use computer response systems (CRS) for immediately gathering students' responses to short questions. Students discuss a question presented by the teacher with peers or think about it individually, and then report their answer using a CRS. After a histogram of responses is displayed, whole class discussion can take place. Teachers use the responses collected with the CRS and the classroom discussion for adjusting their teaching plan on the spot or revisit them for later use (Lee et al., 2012).

Teachers' assessment experience and confidence in their professional judgment were also identified as a feature that leads to successful implementation. Experience with AfL activities helps teachers gain a deep understanding of AfL and fosters their confidence in their instructional decisions (Birenbaum et al., 2011; Feldman & Capobianco, 2008; Fletcher & Shaw, 2012; Ní Chróinín & Cosgrave, 2013).

### 3.1.2. Teacher beliefs and attitudes

Studies that referred to teacher beliefs and attitudes mostly focused on the level of commitment to the underlying ideals of AfL, instead of merely the use of a series of techniques that are routinely applied. In these studies, knowledge, skills and attitude are often conceived of as intertwined. Fourteen studies found evidence for characteristics of teachers' beliefs and attitudes as important for 'deep' implementation of AfL.

Teachers' beliefs, attitudes, perspectives and philosophy about teaching and learning influence the quality of AfL implementation (Havnes et al., 2012; Lee et al., 2012; Penuel et al., 2007; Rakoczy, Klieme, Bürgermeister, & Harks, 2008; Sach, 2013). In addition, both Aschbacher and Alonzo (2006) and Birenbaum et al. (2011) found that the quality of AfL practices is influenced by the extent to which teachers feel responsible for student attainment of goals rather than just coverage of the curriculum. Teachers with less of a feeling of responsibility also felt less inclination to evaluate student work, give effective feedback and revise teaching plans where needed.

Furthermore, several studies specifically reported that teachers should have a constructivist view of learning and pedagogical strategies when it comes to the implementation of AfL (Birenbaum et al., 2011; Penuel et al., 2007; Rakoczy et al., 2008; Sach, 2013). In addition, AfL practices that either represented features of high quality instruction or resulted in high student performance incorporated student-centred AfL activities (Birenbaum et al., 2011; Bryant & Carless, 2010; Fletcher & Shaw, 2012; Hargreaves, 2013; Havnes et al., 2012; O'Loughlin, Ní Chróinín, & O'Grady, 2013). For example, Birenbaum et al. (2011) found that student-centred assessment tasks and instruction positively relate to known characteristics of AfL practices (e.g., focus on learning processes and promotion of self-regulated learning). Fletcher and Shaw (2012) showed that student-directed assessment resulted in higher achievement results and higher levels of motivation and enjoyment compared to teacher-directed assessment. As the implementation of AfL often means practical change, results also showed that teachers should be willing to change their assessment practices (Feldman & Capobianco, 2008; Lee, 2011).

## 3.2. Assessment

### 3.2.1. Assessment content and presentation

In nine studies evidence was found regarding content-related aspects of assessment tasks and feedback. Several studies emphasized the importance of substantial, constructive and focused feedback. This kind of feedback specifies incorrect responses and suggests how to improve learning based on students' progress (Hargreaves, 2013; Havnes et al., 2012; Kay &

Knaack, 2009; Lee, 2011; Ní Chróinín & Cosgrave, 2013; Rakoczy et al., 2008). These studies showed that feedback can have a substantial impact on students' motivation for learning. Overly directive, dishonest, wordy and unfocused feedback is perceived as irrelevant and frustrating by students (Hargreaves, 2013; Kay & Knaack, 2009). The more students are provided with positive informational feedback containing cues on how to proceed, the more students feel excited, stimulated and interested in the subject, and the more deeply they tend to elaborate on the content (Rakoczy et al., 2008). Gamlem and Smith (2013) did a study on feedback and found four feedback typologies that illustrate these results. The two types that were perceived as most useful by students include active formative feedback that was given on the spot and for which opportunities (time) to work with the feedback were provided. Feedback was part of a dialogue between the teacher and the students or among students.

Assessment tasks should elicit students' pre-existing ideas in order to be able to move in the zone of proximal development. In this way prior knowledge can be addressed, which helps students to construct their own knowledge. Furthermore, assessment tasks should be meaningful and authentic, as this resulted in significantly increased performance, creativity and motivation (Fox-Turnbull, 2006). Riggan and Olah (2011) examined how teachers collect, interpret, and act on different types of assessment information. They found that assessments were used for organizational purposes, for identifying specific weaknesses of individual students, for explaining students' thinking and problem-solving processes, and for informing teachers' pacing decisions (e.g., whether the class could move on to a new unit). These assessment interactions were mostly characterized by open-ended questions.

### 3.2.2. Alignment and integration

Fifteen studies emphasized the need for alignment and integration of AfL in the curriculum and instructional tasks. Alignment of AfL with the curriculum and standards is vital; several studies show that systematic feedback procedures that include setting specific assessment goals and criteria (e.g., through rubrics) should be in place, as these help students to structure their assessments, become aware of what is expected of them and realize how their effort contributes to their achievement (Birenbaum et al., 2011; Fletcher & Shaw, 2012; Gamlem & Smith, 2013; Havnes et al., 2012; Newby & Winterbottom, 2011; Ní Chróinín & Cosgrave, 2013; O'Loughlin et al., 2013).

AfL practices (including feedback) need to be closely integrated into classroom instruction and should not be viewed as add-on activities (Havnes et al., 2012; Kay & Knaack, 2009; Lee, 2011; Penuel et al., 2007). This implies that assessment tasks should be integrated with teaching strategies and learning of content. Integration of AfL starts with small changes, such as giving students formal opportunities to act on feedback (Gamlem & Smith, 2013) and placing less emphasis on scores/grades than in traditional teaching methods (Lee, 2011). Furthermore, AfL integration can be fostered by using peer- and self-assessment, which helps teachers by sharing responsibilities for learning and assessment practices, and stimulates students' involvement in their own learning process (Bryant & Carless, 2010; Fletcher & Shaw, 2012; Harris & Brown, 2013; Havnes et al., 2012; Kay & Knaack, 2009; Lee, 2011; Ní Chróinín & Cosgrave, 2013).

## 3.3. Context

### 3.3.1. Leadership and culture

Eleven studies found prerequisites regarding leadership and culture. School leaders play an important role in the facilitation of AfL implementation. Leadership should focus on establishing a school-wide AfL culture with a vision, norms, goals and expectations for AfL use (Havnes et al., 2012; Sach, 2013). Most of these studies found collaboration to be an important focal point in this culture. For long-term benefits, teachers need to work collaboratively and engage in communities of practice (Birenbaum et al., 2011; Kay & Knaack, 2009; Lee, 2011). The school leader needs to facilitate AfL use by scheduling it into practical activities, providing professional development and making available the time needed for preparing AfL-based instruction and in-class pedagogical practices (e.g., discussion) (Lee et al., 2012; Ní Chróinín & Cosgrave, 2013).

Birenbaum et al. (2011) examined the impact of the school-based professional learning community using six attributes (organizational structure, professional learning, social climate, motivation, obligation/responsibility and internal regulation). They found that all six attributes affect the quality of AfL practice. Professional learning had the highest impact on the quality of AfL practice, including learning from mistakes, student-centred instruction and fostering inquisitiveness. Additionally, teachers working in positive climates were more internally motivated and committed to improving their practices. Other authors also reported on the influence of the classroom and school atmosphere (Harris & Brown, 2013; Havnes et al., 2012; Newby & Winterbottom, 2011; Rakoczy et al., 2008). A climate with trust, mutual respect and cooperation resulted in higher quality AfL practices than a negative climate that included mistrust, disruptive competition and stress. Positive and trustful classroom relations between teacher, students and peers are also important. The classroom atmosphere needs to signal the philosophy that mistakes are an opportunity to learn and should encourage honest reflection, so that critical feedback will be perceived as constructive instead of judgmental.

Furthermore, teachers working in schools with a decentralized organizational structure showed higher quality AfL practices than teachers who worked in a centralized organizational structure. This is supported by the finding that teachers feel pressured by the accountability system. These teachers tended to have students copy the correct information that has been written on the blackboard, which yielded lower performance than when students were allowed to do their own thinking. This pleads for respecting teachers' autonomy and professionalism and giving them ownership of assessment practices (Aschbacher & Alonzo, 2006; Birenbaum et al., 2011; Sach, 2013).



### 3.3.2. Support and professional development

Eight studies identified long-term professional development as an important feature of successful implementation of AfL (Aschbacher & Alonzo, 2006; Gottheiner & Siegel, 2012; Lee et al., 2012; Penuel et al., 2007; Phelan et al., 2012). Two of these studies showed that teacher professional development has a direct positive impact on student performance (Aschbacher & Alonzo, 2006; Phelan et al., 2012). Teachers need support regarding the use of assessment results, AfL-related teaching strategies, basic principles of good feedback, and effective questioning. This support should also include instructional resources, materials, and examples (Aschbacher & Alonzo, 2006; Gottheiner & Siegel, 2012; Lee et al., 2012; Ní Chróinín & Cosgrave, 2013; Penuel et al., 2007; Phelan et al., 2012). Again, practice-centred collaboration was found to be an important means of facilitating this kind of support. Results show teachers need to engage in conversations with colleagues about formative assessment and teaching, and collaborate on shared problems and dilemmas (Birenbaum et al., 2011; Feldman & Capobianco, 2008).

### 3.4. Student

#### 3.4.1. Student knowledge and skills

Only four studies referred to the abilities that students need to make effective use of AfL, specifically in peer- and self-assessment. Newby and Winterbottom (2011) studied the use of AfL strategies and found that providing assessment criteria was important for the success of peer- and self-assessment. However, without appropriate training in the use of such criteria, students only gave feedback or shared ideas that focused on relatively straightforward improvements. Similar to Newby and Winterbottom (2011), three additional studies found evidence that students require training in providing and receiving feedback and in the use of assessment criteria for peer- and self-assessment. Students should be able to accurately assess their peers' work, identify meaningful areas for improvement and provide high quality feedback (Bryant & Carless, 2010; Harris & Brown, 2013; Newby & Winterbottom, 2011). Students can become increasingly autonomous in using set assessment criteria to assess their own or peers' work. Moreover, results suggest that students benefit more from self-directed assessment than from teacher-directed assessment in terms of learning outcomes (Fletcher & Shaw, 2012).

#### 3.4.2. Student beliefs and attitudes

Eight studies imply that a positive attitude and taking an active role in their own learning process fosters autonomy and responsibility in learning for students (Bryant & Carless, 2010; Fletcher & Shaw, 2012; Harris & Brown, 2013; Havnes et al., 2012; Lee, 2011; Newby & Winterbottom, 2011). The study by Havnes et al. (2012) shows that teachers believe that passive students cannot use the teachers' feedback effectively. These authors also found that students perceived their active involvement as meaningful and useful in future learning. Fletcher and Shaw (2012) found that higher levels of student autonomy in learning and assessment processes was positively related to learning. Increased levels of behavioural, emotional and cognitive engagement also resulted in a sense of responsibility for their own learning. Furthermore, inappropriate student behaviour negatively affected the implementation of AfL in the classroom (Lee, 2011; Lee et al., 2012).

## 4. Conclusions and discussion

### 4.1. Prerequisites for the effective implementation of AfL in the classroom

A meta-analysis and review conducted on the effects of formative assessment, such as AfL, shows that formative assessment often has limited to no effects (Hendriks et al., 2014). This seems to be due to the often ineffective implementation of formative assessment (Bennet, 2011). This study focussed on one specific type of formative assessment, called AfL. Prerequisites for effective AfL implementation were investigated by means of systematic review. AfL must be implemented properly to lead to increased student learning.

Of the 1743 publications initially found, only 25 studies pertained to AfL and satisfied the inclusion and quality criteria. Similar results were found for both PE and SE contexts. In this paper, VE was not considered due to the low number of studies of sufficient quality (only three) found in this educational setting. This suggests that the availability of high quality research in this field is limited, especially in VE. Results were reported from all over the world, but the US was represented most frequently. Most of the selected studies used a qualitative research design.

The identified prerequisites were grouped into the categories 'teacher', 'student', 'assessments' and 'context', where context is limited to the contextual factors internal to the school because this review identified limited evidence with respect to contextual factors external to the school. Both sets of results not only focus on the separate categories but also suggest relationships between these categories. Based on the results, Fig. 1 visualizes a conceptual model in which the different prerequisites important in AfL implementation are brought together. The Figure shows that these prerequisites influence the establishment of an AfL-based learning environment ('AfL in classroom practice') and therefore, eventually, student learning. The context greatly determines how successfully the implementation of AfL is facilitated. Teachers and students are related through their interaction in practice, for example, in discussions to elicit students' thinking or feedback as an interactive dialogue. Both students and teachers are related to assessments through 'enactment'. The term enactment was chosen to interpret the translation of beliefs, attitudes, knowledge or skills into actions in practice. Teachers, for example, interpret assessment information on the spot and adapt instruction. Students use the assessment criteria in peer- and self-assessment.

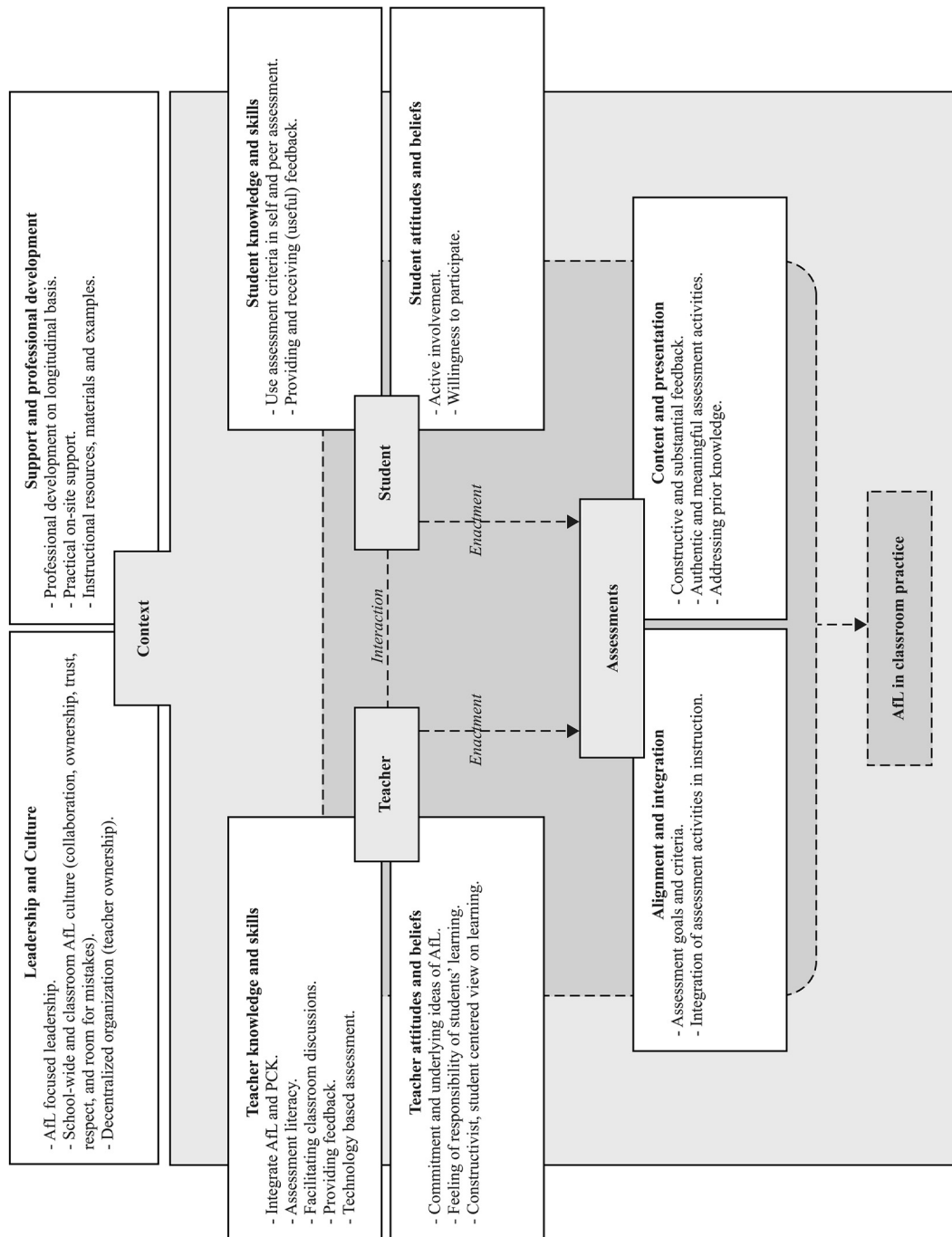


Fig. 1. Conceptual model for AfL implementation.

#### 4.2. Implications for practice

First, the results of our study show that it is crucial to invest in professional development. Sustainable implementation of any educational change requires change beyond surface structures or procedures, focused on altering the knowledge, skills, attitudes and beliefs of all involved agents (cf. Coburn, 2003). Furthermore, the results of the literature review support the notion that attitudes and beliefs that underlie teachers' pedagogical choices have an important influence on the extent to which the teacher can implement AfL to its full potential (cf. Hargreaves, 2005). Namely, the literature on AfL suggests that beliefs and attitudes of teachers who are committed to the underlying ideas of AfL are fundamentally different from those of teachers that only use routine/procedural expertise (Lysaght & O'Leary, 2013; Marshall & Drummond, 2006; Pedder & James, 2012). Professional development should explicitly address how all five AfL strategies can be integrated in classroom practice in order to maximise its potential impact; Lee, 2011

Second, although previous research shows that educational policies have major implications for the implementation of AfL (Birenbaum et al., 2015; Marshall & Drummond, 2006), only one study in our review referred to a policy-related factor external to the school, namely the pressure of the accountability system on teachers' work (Sach, 2013). This can lead to a narrow focus on the use of standardised test results, and result in practices unhelpful to student learning, such as teaching to the test, and even manipulating test scores and test samples (e.g., Booher-Jennings, 2005). The influence of factors from both school based policy and educational policy should be taken into account when implementing AfL, which will be unique for each particular context.

Third, the role of the student should be taken into account when implementing AfL. It should be noted that only a relatively small number of studies referred to prerequisites regarding students. This is remarkable, as modern assessment theory emphasizes a critical role for students in assessment (Black, 2015), particularly in AfL as its definition focuses on everyday classroom life in which teachers and students are key agents (Pedder & James, 2012). Studies that did consider the role of students all emphasized on the need for training in the use of assessment criteria for providing meaningful feedback. However, students also need to be motivated to invest effort in seeking and processing feedback (Ruiz-Primo & Furtak, 2006; Timmers & Veldkamp, 2011), which was not considered in these studies.

Furthermore, although we found no explicit differences in the prerequisites for primary and secondary education, the specific implications for practice may differ. For example, teachers in both contexts would be required to hold pedagogical content knowledge, but in primary education this knowledge would need to cover a considerable body of the curriculum, whereas in secondary education this would be specific to a particular subject area that a teacher teaches.

Finally, these results highlighted that these researchers aligned AfL with a constructivist approach to learning. The constructivist, student-centred approach to learning suggests students and teachers should work together interactively, embrace the social interaction between students, peers and teachers, and share responsibility for learning (Black & William, 1998a, 1998b; Heritage, 2010; Tunstall & Gipps, 1996). This is not to say that other learning theories are not important in AfL, for example, AfL is also underpinned by ideas from metacognitivism and social cultural theory. However, in much research on formative assessment the focus is the teacher–student interaction, and underlying theoretical assumptions about the nature of learning remain implicit (Van der Kleij et al., 2015). In practice, teachers usually do not base their teaching on just one educational approach or philosophy but use a mixture in which one approach is more dominant than the others (Niederhauser & Stoddart, 2001).

#### 4.3. Limitations

As with every systematic review, one limitation of this study is that it was impossible to include all relevant studies. The search strategies and selection criteria used in this review determined the selected set of studies. Furthermore, time pressure, publication bias and researcher bias were potential threats to the quality of this review. However, these threats were constantly monitored and addressed (Petticrew & Roberts, 2006). The research team frequently discussed methods and results, data extraction forms were supported by clear instructions, data were double-coded and low quality studies were excluded. Additionally, external experts were involved to verify the chosen strategies and the identified results.

#### 4.4. Implications for further research

The results of this review show that most studies were based on a small-scale, qualitative research design, as was also noted by Baird et al. (2014). Few AfL studies are based on large-scale or quantitative research. When studies do focus on large-scale implementation of AfL, the results are often limited to student and teacher perceptions as the main dependent variable (cf. Hopfenbeck & Stobart, 2015). Therefore, the time has come to invest as well in large-scale quantitative studies investigating the factors that enable or hinder the implementation of AfL in the classroom. This can provide insight into which factors matter the most when using AfL in the classroom to improve student learning.

Furthermore, the results of this review can inform additional future research, as they contribute to a better understanding of the multiple facets that need to be considered when implementing AfL, both theoretically and practically. Future studies should preferably focus on large-scale research in local contexts that take a more comprehensive view and consider the prerequisites identified as important for effective AfL implementation in this study. The conceptual framework can be used to inform practical initiatives such as professional development in which the entire school is involved. It should be remembered

that educational practice is greatly influenced by the dynamics of classroom life and depends on the classroom context. This means that an exact prescription for success cannot be provided, and local practitioners need to translate prerequisites important for AfL implementation to their local context for such implementation to lead to increased student learning.

## References<sup>\*</sup>

- \*Aschbacher, P., & Alonzo, A. (2006). Examining the utility of elementary science notebooks for formative assessment purposes. *Educational Assessment*, 11, 179–203. [http://dx.doi.org/10.1207/s15326977ea1103&4\\_3](http://dx.doi.org/10.1207/s15326977ea1103&4_3).
- Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box*. Retrieved from [http://www.nuffieldfoundation.org/sites/default/files/files/beyond\\_blackbox.pdf](http://www.nuffieldfoundation.org/sites/default/files/files/beyond_blackbox.pdf).
- Baird, J., Hopfenbeck, T. N., Newton, P., Stobart, G., & Steen-Utheim, A. T. (2014). *State of the field review: Assessment and learning (Case No. 13/4697)*. Retrieved from the University of Oxford, Norwegian Knowledge Centre for Education website <http://goo.gl/r8zTcG>.
- Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, 18, 5–25. <http://dx.doi.org/10.1080/0969594X.2010.513678>.
- Birenbaum, M., DeLuca, C., Earl, L., Heritage, M., Klenowski, V., Looney, A., et al. (2015). International trends in the implementation of assessment for learning: implications for policy and practice. *Policy Futures in Education*, 13, 117–140. <http://dx.doi.org/10.1177/1478210314566733>.
- \*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, 35–48. <http://dx.doi.org/10.1016/j.stueduc.2011.04.001>.
- Black, P. (2015). Formative assessment – an optimistic but incomplete vision. *Assessment in Education: Principles, Policy & Practice*, 22, 161–177. <http://dx.doi.org/10.1080/0969594X.2014.999643>.
- Black, P., McCormick, R., James, M., & Pedder, D. (2006). Learning how to learn and assessment for learning: a theoretical inquiry. *Research Papers in Education*, 21, 119–132. <http://dx.doi.org/10.1080/02671520600615612>.
- Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5, 7–74. <http://dx.doi.org/10.1080/0969595980050102>.
- Black, P., & Wiliam, D. (1998b). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139–148.
- Booher-Jennings, J. (2005). Below the bubble: “Educational triage” and the Texas accountability system. *American Educational Research Journal*, 42, 231–268. <http://dx.doi.org/10.3102/00028312042002231>.
- Brookhart, S. B. (2007). Expanding views about formative classroom assessment: a review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 43–62). New York, NY: Teachers College Press.
- \*Bryant, D. A., & Carless, D. R. (2010). Peer assessment in a test-dominated setting: empowering, boring or facilitating examination preparation? *Educational Research for Policy and Practice*, 9, 3–15. <http://dx.doi.org/10.1007/s10671-009-9077-2>.
- Coburn, S. E. (2003). Rethinking scale: moving beyond numbers to deep and lasting change. *Educational Researcher*, 32, 3–12. <http://dx.doi.org/10.3102/0013189X032006003>.
- Crisp, G. T. (2012). Integrative assessment: reframing assessment practice for current and future learning. *Assessment & Evaluation in Higher Education*, 37, 33–43. <http://dx.doi.org/10.1080/02602938.2010.494234>.
- Elwood, J., & Klenowski, V. (2002). Creating communities of shared practice: the challenges of assessment use in learning and teaching. *Assessment & Evaluation in Higher Education*, 27, 243–256. <http://dx.doi.org/10.1080/0260293022013860>.
- \*Feldman, A., & Capobianco, B. M. (2008). Teacher learning of technology enhanced formative assessment. *Journal of Science Education and Technology*, 17, 82–99. <http://dx.doi.org/10.1007/s10956-007-9084-0>.
- \*Fletcher, A., & Shaw, G. (2012). How does student-directed assessment affect learning? Using assessment as a learning process. *International Journal of Multiple Research Approaches*, 6, 245–263. <http://dx.doi.org/10.5172/mra.2012.6.3.24510.1037/0022-3514.45.2.357>.
- \*Fox-Turnbull, W. (2006). The influences of teacher knowledge and authentic formative assessment on student learning in technology education. *International Journal of Technology and Design Education*, 16, 53–77. <http://dx.doi.org/10.1007/s10798-005-2109-1>.
- \*Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching*, 49, 1181–1210. <http://dx.doi.org/10.1002/tea.21054>.
- \*Gamlem, S. M., & Smith, K. (2013). Student perceptions of classroom feedback. *Assessment in Education: Principles, Policy and Practice*, 20, 150–169. <http://dx.doi.org/10.1080/0969594X.2012.749212>.
- Gottheiner, D. M., & Siegel, M. A. (2012). Experienced middle school science teachers' assessment literacy: investigating knowledge of students' conceptions in genetics and ways to shape instruction. *Journal of Science Teacher Education*, 23, 531–557. <http://dx.doi.org/10.1007/s10972-012-9278-z>.
- Hargreaves, E. (2005). Assessment for learning? Thinking outside the (black) box. *Cambridge Journal of Education*, 35, 213–224. <http://dx.doi.org/10.1080/03057640500146880>.
- \*Hargreaves, E. (2013). Inquiring into children's experiences of teacher feedback: reconceptualising assessment for learning. *Oxford Review of Education*, 39, 229–246. <http://dx.doi.org/10.1080/03054985.2013.787922>.
- Harris, L. R., & Brown, G. T. L. (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, 101–111. <http://dx.doi.org/10.1016/j.tate.2013.07.008>.
- \*Harris, L. R., Brown, G. T. L., & Harnett, J. A. (2014). Understanding classroom feedback practices: a study of New Zealand student experiences, perceptions, and emotional responses. *Educational Assessment, Evaluation and Accountability*, 1–27. <http://dx.doi.org/10.1007/s11092-013-9187-5>.
- \*Havnes, A., Smith, K., Dysthe, O., & Ludvigsen, K. (2012). Formative assessment and feedback: making learning visible. *Studies in Educational Evaluation*, 38, 21–27. <http://dx.doi.org/10.1007/s10972-012-9278-z>.
- Hayward, L. (2014). Assessment for learning and the journey towards inclusion. In L. Florian (Ed.), *SAGE handbook of special education* (2nd ed., pp. 523–535). London, UK: SAGE.
- Hendriks, M. A., Scheerens, J., & Sleegers, P. (2014). Effects of evaluation and assessment on student achievement: a review and meta-analysis. In M. Hendriks (Ed.), *The influence of school size, leadership, evaluation, and time on student outcomes* (pp. 127–174). Enschede: University of Twente.
- Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Thousand Oaks, CA: Corwin Press.
- Hopfenbeck, T. N., & Stobart, G. (2015). Large-scale implementation of assessment for learning. *Assessment in Education: Principles, Policy & Practice*, 22, 1–2. <http://dx.doi.org/10.1080/0969594X.2014.1001566>.
- James, M., McCormick, R., Black, P., Carmichael, P., Drummond, M.-J., Fox, A., et al. (2007). *Improving learning how to learn – Classrooms, schools and networks*. Abingdon, UK: Routledge.
- \*Kay, R., & Knaack, L. (2009). Exploring the use of audience response systems in secondary school science classrooms. *Journal of Science Education and Technology*, 18, 382–392. <http://dx.doi.org/10.1007/s10956-009-9153-7>.
- Klenowski, V. (2009). Assessment for learning revisited: an Asia-Pacific perspective. *Assessment in Education: Principles, Policy & Practice*, 16, 263–268. <http://dx.doi.org/10.1080/09695940903319646>.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. <http://dx.doi.org/10.2307/2529310>.

\* References with an asterisk indicate studies included in the systematic review.

- Lee, I. (2011). Bringing innovation to EFL writing through a focus on assessment for learning. *Innovation in Language Learning and Teaching*, 5, 19–33. <http://dx.doi.org/10.1080/17501229.2010.502232>.
- \*Lee, H., Feldman, A., & Beatty, I. D. (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, 523–539. <http://dx.doi.org/10.1007/s10956-011-9344-x>.
- Lysaght, Z., & O'Leary, M. (2013). An instrument to audit teachers' use of assessment for learning. *Irish Educational Studies*, 32, 217–232. <http://dx.doi.org/10.1080/03323315.2013.784636>.
- Mandinach, E. B., & Jackson, S. S. (2012). *Transforming teaching and learning through data-driven decision making*. Thousand Oaks, CA: Corwin.
- Marshall, B., & Drummond, M. J. (2006). How teachers engage with assessment for learning: lessons from the classroom. *Research Papers in Education*, 21, 133–149. <http://dx.doi.org/10.1080/02671520600615638>.
- \*Newby, L., & Winterbottom, M. (2011). Can research homework provide a vehicle for assessment for learning in science lessons? *Educational Review*, 63, 275–290. <http://dx.doi.org/10.1080/00131911.2011.560247>.
- \*Ní Chróinín, D., & Cosgrave, C. (2013). Implementing formative assessment in primary physical education: teacher perspectives and experiences. *Physical Education and Sport Pedagogy*, 18, 219–233. <http://dx.doi.org/10.1080/17408989.2012.666787>.
- Niederhauser, D. S., & Stoddart, T. (2001). Teachers' instructional perspectives and use of educational software. *Teaching and Teacher Education*, 17, 15–31. [http://dx.doi.org/10.1016/S0742-051X\(00\)00036-6](http://dx.doi.org/10.1016/S0742-051X(00)00036-6).
- \*O'Loughlin, J., Ní Chróinín, D., & O'Grady, D. (2013). Digital video: the impact on children's learning experiences in primary physical education. *European Physical Education Review*, 19, 165–182. <http://dx.doi.org/10.1177/1356336x13486050>.
- Pedder, D., & James, M. (2006). Professional learning as a condition for assessment for learning. In J. Gardner (Ed.), *Assessment and learning* (pp. 27–43). London, UK: Sage.
- Pedder, D., & James, M. (2012). Professional learning as a condition for assessment for learning. In J. Gardner (Ed.), *Assessment and learning* (pp. 33–48). London: Sage.
- \*Penuel, W. R., Boscardin, C. K., Masyn, K., & Crawford, V. M. (2007). Teaching with student response systems in elementary and secondary education settings: a survey study. *Educational Technology Research and Development*, 55, 315–346. <http://dx.doi.org/10.1007/s11423-006-9023-4>.
- Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford, UK: Blackwell.
- \*Phelan, J. C., Choi, K., Niemi, D., Vendilinski, T. P., Baker, E. L., & Herman, J. (2012). The effects of POWERSOURCE<sup>®</sup> assessments on middle-school students' math performance. *Assessment in Education: Principles, Policy and Practice*, 19, 211–230. <http://dx.doi.org/10.1080/0969594X.2010.532769>.
- \*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, 111–124. <http://dx.doi.org/10.1027/0044-3409.216.2.111>.
- \*Riggan, M., & Olah, L. N. (2011). Locating interim assessments within teachers' assessment practice. *Educational Assessment*, 16, 1–14. <http://dx.doi.org/10.1080/10627197.2011.551085>.
- Ruiz-Primo, M. A., & Furtak, E. M. (2006). Informal formative assessment and scientific inquiry: exploring teachers' practices and student learning. *Educational Assessment*, 11, 237–263. <http://dx.doi.org/10.1080/10627197.2006.9652991>.
- 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, 322–335. <http://dx.doi.org/10.1080/03004279.2013.813956>.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119–144. <http://dx.doi.org/10.1007/BF00117714>.
- Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, 26, 482–496. <http://dx.doi.org/10.1016/j.tate.2009.06.007>.
- Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. Abingdon, England: Routledge.
- Timmers, C. F., & Veldkamp, B. P. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers in Education*, 56(3), 923–930. <http://dx.doi.org/10.1016/j.compedu.2010.11.007>.
- Tunstall, P., & Gipps, C. (1996). Teacher feedback to young children in formative assessment: a typology. *British Educational Research Journal*, 22, 389–404. <http://dx.doi.org/10.1080/0141192960220402>.
- Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2015). Integrating data-based decision making, assessment for learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice*, 22, 324–343. <http://dx.doi.org/10.1080/0969594X.2014.999024>.
- William, D., & Thompson, M. (2007). Integrating assessment with instruction: what will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching learning* (pp. 53–82). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wyatt-Smith, C., Klenowski, V., & Colbert, P. (2014). Assessment understood as enabling. In C. Wyatt-Smith, V. Klenowski, & P. Colbert (Eds.), *Designing assessment for quality learning* (pp. 1–20). Dordrecht, The Netherlands: Springer International.
- \*Yin, Y., Tomita, M. K., & Shavelson, R. J. (2014). Using formal embedded formative assessments aligned with a short-term learning progression to promote conceptual change and achievement in science. *International Journal of Science Education*, 36, 531–552. <http://dx.doi.org/10.1080/09500693.2013.787556>.