What's Known

February 2017

Formative assessment and elementary school student academic achievement: A review of the evidence

Mary Klute
Helen Apthorp
Jason Harlacher
Marianne Reale
Marzano Research

Key findings

Formative assessment is a process that engages teachers and students in gathering, interpreting, and using evidence about what and how students are learning. This review identifies rigorous studies of the effectiveness of formative assessment on elementary school student achievement. Results of the review indicate that:

- Overall, formative assessment had a positive effect on student academic achievement. On average across the studies, students who participated in formative assessment performed better on measures of academic achievement than those who did not.
- Formative assessment used during math instruction had larger effects, on average, than did formative assessment used during reading and writing instruction.
- For math, both student-directed formative assessment and formative assessment directed by other agents, such as an educator or a computer program, were effective.
- For reading, other-directed formative assessment was more effective than student-directed formative assessment.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S. Department of Education

REL CENTRAL
Regional Educational Laboratory
At Marzano Research

U.S. Department of Education
Betsy DeVos, *Secretary*

Institute of Education Sciences
Thomas W. Brock, *Commissioner for Education Research*
Delegated the Duties of Director

National Center for Education Evaluation and Regional Assistance
Audrey Pendleton, *Acting Commissioner*
Elizabeth Eisner, *Acting Associate Commissioner*
Amy Johnson, *Action Editor*
Sandra Garcia, *Project Officer*

REL 2017–259

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

February 2017

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-12-C-0007 by Regional Educational Laboratory Central administered by Marzano Research. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

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). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Central. Retrieved from http://ies.ed.gov/ncee/edlabs.

This report is available on the Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

Summary

Formative assessment is a process that engages teachers and students in gathering, interpreting, and using evidence about what and how students are learning in order to facilitate further student learning during a short period of time. The process offers the potential to guide educator decisions about midstream adjustments to instruction that address learner needs in a timely manner. Formative assessment can be implemented in classrooms in various ways. For example, formative assessment can be quick and informal, such as giving students "I learned..." prompts to reflect on and discuss their progress toward lesson objectives. Formative assessment can also be more formal and involve multiple components, such as curriculum-based measurement,[1] to frequently track and analyze individual student learning for the purpose of modifying instruction as warranted (Black & Wiliam, 1998a).

Members of Regional Educational Laboratory (REL) Central's Formative Assessment Research Alliance, including principals and district administrators, indicated that teachers in the region vary widely in their understanding of formative assessment and how to use it. They wished to focus professional development efforts on formative assessment practices that have evidence of effectiveness for promoting student learning. To address this need, this review identifies studies that examine the effectiveness of formative assessment and provides an overall average estimate of its effectiveness. Alliance members also expressed concern that teachers have difficulty finding time to use formative assessment. One approach to minimizing the formative assessment burden on teachers is to involve students more actively in the process (Black & Wiliam, 1998a). This review also compares the effectiveness of different types of formative assessment, including those directed by students and those directed by other agents, such as educators and computer software programs.

The review team conducted a comprehensive search to locate research on formative assessment interventions. After screening studies for relevance, researchers certified in the U.S. Department of Education's What Works Clearinghouse (WWC) standards and procedures coded and rated each of 76 relevant studies using systematic, rigorous, scientific evidence standards modeled after the WWC study review process and standards (U.S. Department of Education, 2014b).

The review team identified 23 studies that it determined had been conducted rigorously enough to have confidence that the formative assessment interventions caused the observed effects on student outcomes. Twenty-two of the studies compared academic outcomes for students participating in formative assessment with academic outcomes for students who did not participate in formative assessment. Nineteen of the 22 studies provided enough information to calculate an effect size, which describes the magnitude of the effect of the intervention. When examining the results across these 19 studies, the review team concluded that:

- Overall, formative assessment had a positive effect on student academic achievement. On average across all the studies, students who participated in formative assessment performed better on measures of academic achievement than those who did not.
- Formative assessment used during math instruction had larger effects, on average, than did formative assessment used during reading and writing instruction.

- Across all subject areas (math, reading, and writing), formative assessment had larger effects on student academic achievement when other agents, such as a teacher or a computer program, directed the formative assessment.
- For math, both student-directed formative assessment and formative assessment directed by other agents were effective.
- For reading, other-directed formative assessment was more effective than student-directed formative assessment.
- For writing, the effect of other-directed formative assessment on student academic achievement was small, and not enough evidence was available to determine the effectiveness of student-directed formative assessment.

Contents

Figure

Tables

Why this study?

In the past two decades assessment experts and education leaders have promoted formative assessment as a necessary complement to summative accountability assessments, which evaluate student learning after instruction has been completed (Andrade & Cizek, 2010; Heritage, 2010a; Popham, 2013; Shepard, 2000). Formative assessment is a process that engages teachers and students during instruction in gathering, interpreting, and using evidence about what and how students are learning in order to facilitate further student learning (Black & Wiliam, 2009; Heritage, 2010b; Moss & Brookhart, 2009).

This report focuses on formative assessment that occurs within a relatively short period of time, lasting up to four weeks. The frequency with which teachers formatively assess student learning varies. Short-cycle formative assessment occurs frequently, moment by moment, daily, or weekly, "within and between lessons" (National Council of Teachers of Mathematics, 2007, p. 7). Medium-cycle formative assessment occurs less frequently, "within and between instructional units" (National Council of Teachers of Mathematics, 2007, p. 7). Although assessment information from end-of-course, end-of-grade, or other summative testing can be used formatively at any time, the utility of the shorter cycle is in adjusting instruction, whereas the utility of the longer cycle is in adjusting curriculum (Brookhart, 2014; Perie, Marion, & Gong, 2009). By creating feedback loops during teaching and learning, formative assessment conducted in a short cycle has the potential to guide midstream, just-in-time adjustments to help students learn. This early recognition of learner needs is critical to prevent elementary school students whose academic development has slowed from falling further behind (Baumert, Nagy, & Lehmann, 2012; Carreker et al., 2007).

By creating feedback loops during teaching and learning, formative assessment conducted in a short cycle has the potential to guide midstream, just-in-time adjustments to help students learn

Members of Regional Educational Laboratory (REL) Central's Formative Assessment Research Alliance, including principals and district administrators, indicated that educators in the region vary widely in their understanding of formative assessment and how to implement it. The research alliance members requested a review of research evidence to help them make sound decisions on developing teacher knowledge and skills in formative assessment by identifying practices that have evidence of effectiveness for promoting student learning.

Prior research reviews have provided widely varying estimates of the effectiveness of formative assessment (Black & Wiliam, 1998a, 1998b; Kingston & Nash, 2011). The current review improves on prior reviews by considering whether the studies of formative assessment were conducted rigorously enough to have confidence that the formative assessment caused the observed effects on student outcomes. Confidently attributing causality to formative assessment requires a systematic approach that rates the evidence and sorts studies into those that meet and those that do not meet evidence standards for supporting causal inferences. The current review used an approach modeled after the What Works Clearinghouse (WWC) evidence standards and procedures to identify studies that support causal inferences (U.S. Department of Education, 2014b).[2]

Research alliance members also expressed concern that teachers have difficulty finding time to use formative assessment. One way to reduce the formative assessment burden on teachers is to involve students more actively in the process (Black & Wiliam, 1998a). To shed light on the effectiveness of different approaches to formative assessment, this review examined whether student-directed formative assessment is as effective as other-directed

formative assessment (formative assessment directed by other agents, such as educators or software programs).

The results identify what is known to be effective and what is not yet known to be effective about formative assessment for promoting student academic achievement in the elementary school grades. The results can inform teachers' selection of formative assessment and administrators' and other school leaders' decisions about how to support teachers' use of formative assessment. The results can also inform researchers about areas needing future inquiry.

What the study examined

This review used a procedure modeled after the U.S. Department of Education's WWC systematic review process (U.S. Department of Education, 2014b) to identify studies on the effectiveness of formative assessment published between 1988 and 2014. This review addresses the following research questions:

- What is the effect of formative assessment on elementary school student achievement?
- Does formative assessment have a greater effect on student achievement in some subject areas than in others?
- Does the effect of formative assessment on student achievement vary depending on whether it is student-directed or other-directed?
- Does one type of formative assessment have a greater effect on student achievement in particular subject areas?

The results of this review identify what is known to be effective and what is not yet known to be effective about formative assessment for promoting student academic achievement in the elementary school grades

To address these questions, the review team conducted a comprehensive search of research on a range of interventions that met the definition of formative assessment (see box 1 on features and types of formative assessment). Each study that met the inclusion criteria was evaluated against WWC standards and was assigned a rating (box 2). This report includes only the studies that the review team determined met WWC standards with or without reservations. More information about inclusion criteria and procedures used to search for and evaluate studies is in appendix A.

Details about each study that met standards with or without reservations were recorded. Specifically, the review team determined the type of formative assessment based on the primary agent gathering and using evidence to improve learning (see box 1). Second, the review team recorded the academic subject that the intervention addressed. Although the review team searched for studies in five core academic areas (math, reading, writing, science, and social studies), no studies that met standards focused on science or social studies. Therefore, this report focuses on results for math, reading, and writing. The studies examined the effectiveness of formative assessment for students primarily in grades 1–6 in both general and special education classes.[3]

The search identified 76 studies, 23 of which met standards with or without reservations and are included in this report. (The interventions examined in the 23 studies that met standards are described in appendix B.) Because the focus of the review was the effectiveness of formative assessment, this report focuses primarily on comparisons that test the difference between a group of students who participated in formative assessment and a group of students who did not participate in formative assessment. This analysis excludes one study that compared two types of formative assessment rather than comparing formative assessment with no assessment (Wesson, 1990).

Box 1. Features and types of formative assessment

Features of formative assessment

Formative assessment. Interventions with a process dedicated to both gathering and using assessment information about what and how students are learning to facilitate student learning over either a short cycle (within and between lessons) or a medium cycle (within and between instructional units). Formative assessment interventions can have three iterative phases: establishing learning targets, determining where students are now, and deciding how to help students improve. Interventions were included in this review if either all three or only the second and third of the iterative phases were evident. Studies were included in the review if they examined formative assessment interventions that took place in a cycle lasting up to four weeks. The review includes interventions that are replicable, including programs, practices, strategies, or activities implemented by teachers, students, or both.

Gathering assessment information. Seeking or eliciting evidence of student knowledge, understanding, or behavior.

Using assessment information. Having the explicit opportunity to apply the information to facilitate student learning. This could include a follow-up learning or assessment activity that addresses the same or related learning goal or performance task offered to students, as well as time and guidance provided to teachers for both interpreting assessment information and choosing instructional options.

Types of formative assessment

Student-directed. Students appraise or monitor their own or their peers' work, performance, strategies, or progress and have the opportunity to reflect on the assessment information they gathered to determine next steps (Black & Wiliam, 2009). Self-assessment, self-regulation, and peer assessment are all examples of student-directed formative assessment. For example, in one study, students set a goal for the number of elements to include in their stories, then examined the number of elements in their completed stories, and graphed the number of story elements they had included over time (Sawyer, Graham, & Harris, 1992).

Other-directed. Educators or computer software programs appraise or monitor student work, performance, strategies, or progress and have the opportunity to reflect on the assessment information they gather to determine next steps. For example, in one study the teacher administered an assessment on lesson objectives after delivering lessons in a large-group format. On the basis of the assessment results, the teacher divided students into two groups: students who demonstrated mastery on the assessment participated in enrichment activities, and the rest of the students received additional instruction from the teacher. Next, the teacher administered a second assessment (Null, 1990).

Teachers' support of implementation of students' self-assessment or peer assessment (for example, providing task instructions) is not considered other-directed formative assessment because the teachers themselves are not gathering, interpreting, and using assessment information.

Box 2. What Works Clearinghouse study ratings assigned to studies included in the review

To include only studies that support causal inferences about formative assessment, members of the review team who are trained and certified in the application of What Works Clearinghouse (WWC) procedures and evidence standards for comparative group designs reviewed 76 eligible studies and assigned each study one of three evidence ratings (U.S. Department of Education, 2014b).

- *Meets standards without reservations.* The highest rating a study could receive. These studies were conducted in a way that supports causal inferences about the intervention. Readers of these studies can infer, with a high degree of confidence, that the formative assessment caused the reported results. In this review 16 studies met standards without reservations.
- *Meets standards with reservations.* The middle rating a study could receive. These studies were conducted in a way that readers can infer, with a lower degree of confidence, that the formative assessment was the cause of the outcomes observed. In this review 7 studies met standards with reservations.
- *Does not meet standards.* The lowest rating a study could receive. These studies were conducted in a way that was not rigorous enough to support the interpretation that the formative assessment caused the reported results. In this review 53 studies did not meet standards.

This report focuses on 36 comparisons from the remaining 22 studies. In some cases a single study examined multiple formative assessment interventions and compared the effects of different interventions to each other as well as to the outcomes for a group of students not receiving the intervention. For example, one study compared three groups: one in which students were assessed by a computer, one in which a tutor assessed students, and one that did not receive any intervention. That study included three comparisons: computer versus human, computer versus no intervention, and human versus no intervention (Mostow et al., 2003). In addition, two studies examined the effect of formative assessment separately for different grade levels (Martens, Eckert, & Begeny, 2007; Ysseldyke & Tardrew, 2007). This situation also created multiple comparisons in the same study. In such cases a single study could have several comparisons that met criteria for the review. Each comparison was evaluated separately, and each was assigned an evidence rating.

The question on whether the effectiveness of formative assessment differs by whether it is student-directed or other-directed was examined by looking at studies that compared student-directed formative assessment with no formative assessment and studies that compared other-directed formative assessment with no formative assessment

Although examining studies that compare student-directed formative assessment to other-directed formative assessment would be useful for addressing the third research question, only one study included this type of comparison (McCurdy & Shapiro, 1992). Therefore, the third question on whether the effectiveness of formative assessment differs by whether it is student-directed or other-directed was examined by looking at studies that compared student-directed formative assessment with no formative assessment and studies that compared other-directed formative assessment with no formative assessment. Results of all studies that compared two types of formative assessment (including Wesson, 1990) are presented in appendix C.

To summarize the effectiveness of formative assessment for improving student academic outcomes, effect sizes were calculated separately for each comparison that met standards. Effect sizes are an estimate of the magnitude of the effect of an intervention (see box 3

Box 3. Interpreting effect sizes

Effect sizes describe the size of an intervention effect, in this case the difference between the scores of students who participated in formative assessment and the scores of students who did not. To allow comparisons across studies, effect sizes characterize the effect against a common point of reference. In this review, effect sizes use the standard deviation of the outcome to characterize the size of the effect (Dynarski & Kisker, 2014). The standard deviation can be interpreted as the average distance in either direction between students' scores and the average score. A small standard deviation means that students' scores tightly cluster around the average score. A large standard deviation means that students' scores spread more widely around the average score.

A useful way to understand the meaning of effect sizes for an intervention is to compare them with effect sizes for other more commonly understood differences, such as the amount of change one might expect to see in a year of schooling. In one year of schooling for students in grade 4 an effect size for academic growth is, on average, 0.36 in reading and 0.52 in math (Hill, Bloom, Black, & Lipsey, 2008). If an effect size for formative assessment in reading is 0.30, it can be interpreted as meaningful, as the gain associated with participating in the intervention is nearly as large as what one might expect, on average, from a year of schooling.

It may also be meaningful to compare effect sizes for formative assessment to estimates of the effect sizes for achievement gaps. For example, for grade 4 the effect size for the difference between students who are eligible for the federal school lunch program and those who are not is estimated to be –0.74 in reading and –0.85 in math (Hill et al., 2008). The effect sizes are negative because the target group (in this case, students who are eligible for the federal school lunch program) tend to score lower than the group to which they are compared. An effect size of 0.40 for formative assessment in math for students in grade 4 would be considered meaningful, because it is about half the size of the achievement gap associated with eligibility for the federal school lunch program at this grade.

This study uses a criterion established by the What Works Clearinghouse (WWC) for determining when an effect size is large enough to be noteworthy: an effect size greater than 0.25 or less than –0.25 is considered substantively important (U.S. Department of Education, 2014b).

Statistical significance, a way to judge the noteworthiness of the results of a research study, is influenced by both the size of the effect and the sample size. When the sample size is large, a smaller effect size will be significant. With smaller sample sizes, effects have to be larger to reach statistical significance. As a result, there can be some cases where a statistically significant finding has an effect size between –0.25 and 0.25. Effects that are statistically significant are noted in the "characterization of findings" columns in tables B2 and B3 in appendix B and table C1 in appendix C.

for how to interpret effect sizes). For three studies that involved six comparisons with a comparison group that did not participate in formative assessment, there was not enough information to calculate effect sizes. As a result, effect sizes are summarized in this report for 30 comparisons from 19 studies.

What the study found

This section describes the results for each research question.

On average across all the studies, formative assessment had a positive effect on student academic achievement

The 19 studies that met standards included 30 separate effect sizes. The average of these effect sizes was 0.26 standard deviation, which is just over the benchmark set by the WWC for a substantively important effect size (greater than 0.25 or less than –0.25). However, the effect sizes ranged from –0.46 to 1.22 (table 1).

Formative assessment in math had larger effects, on average, on student academic achievement than did formative assessment in reading and writing

The average effect size for formative assessment in math was 0.36 standard deviation, which exceeds the WWC threshold for a substantively important effect size. The average effect size was smaller for reading (0.22) and writing (0.21), approaching the threshold for a substantively important effect

The average effect size for formative assessment in math was 0.36 standard deviation, which exceeds the WWC threshold for a substantively important effect size. The average effect size was smaller for reading (0.22) and writing (0.21), approaching the threshold for a substantively important effect.

Formative assessment in writing comprised two distinct types. Two studies investigated formative assessment in spelling with special education students. Four studies examined formative assessment in composition with older elementary school students in grades 4–6. The average effect size for the studies investigating formative assessment in spelling (0.19) was slightly lower than the effect size for the studies investigating formative assessment in composition (0.22).

Table 1. Mean effect sizes for formative assessment, by subject area

Subject area	Number of studies[a]	Number of effect sizes[b]	Mean effect size	Standard deviation	Minimum effect size	Maximum effect size
Math	6	10	**0.36**	0.33	–0.18	**1.01**
Reading	7	12	0.22	0.45	**–0.46**	**1.22**
Writing	6	8	0.21	0.24	–0.20	**0.63**
Spelling	2	4	0.19	0.09	0.09	**0.30**
Composition	4	4	0.22	0.35	–0.20	**0.63**

Note: The table presents descriptive statistics for effect sizes across studies. Bolded values indicate effect sizes greater than the What Works Clearinghouse benchmark for a substantively important effect size (greater than 0.25 or less than –0.25; U.S. Department of Education, 2014b). See tables B2 and B3 in appendix B for the statistical significance of the effects in individual studies.

a. The column sum (19) does not equal the total number of studies reviewed (22) because 3 studies (Craven, Marsh & Debus, 1991; Fuchs, Butterworth & Fuchs, 1989; and Mostow et al., 2003) did not provide enough information to calculate effect sizes.

b. The number of effect sizes is greater than the number of studies because two studies of math (Fuchs, Fuchs, Hamlett, & Stecker, 1991, and Ysseldyke & Tardrew, 2007), four studies of reading (Fuchs, Fuchs, Hamlett, & Ferguson, 1992; Johnson, Graham, & Harris, 1997; Martens, Eckert, & Begeny, 2007; and McCurdy & Shapiro, 1992), and two studies of writing (Fuchs, Fuchs, Hamlett, & Allinder, 1991a, 1991b) included more than one comparison for which effect sizes could be calculated.

Source: Authors' analysis of studies published between 1988 and 2014; see appendix A for details.

Across all subject areas formative assessment had larger effects on academic outcomes when other agents directed the formative assessment

The average effect size for student-directed formative assessment was 0.20 standard deviation, which does not meet the WWC threshold for a substantively important effect (table 2). The average effect size for other-directed formative assessment was larger, at 0.29 standard deviation, which exceeds the WWC threshold for a substantively important effect.

Both student-directed and other-directed formative assessment in math were effective

Seven studies examined formative assessment in math. These studies tested 10 comparisons for which effect sizes could be calculated. The average effect size for both student-directed formative assessment (0.45) and other-directed formative assessment (0.30) was substantively important (table 3). Examples of student- and other-directed formative assessment with the largest effect sizes are presented in box 4.

The average effect size for student-directed formative assessment was 0.20 standard deviation, which does not meet the WWC threshold for a substantively important effect. The average effect size for other-directed formative assessment was larger, at 0.29 standard deviation, which exceeds the WWC threshold for a substantively important effect

Table 2. Mean effect sizes for formative assessment, by type

Type of formative assessment	Number of studies[a]	Number of effect sizes[b]	Mean effect size	Standard deviation	Minimum effect size	Maximum effect size
Student-directed	7	9	0.20	0.48	**–0.46**	**1.01**
Other-directed	13	21	**0.29**	0.30	–0.20	**1.22**

Note: The table presents descriptive statistics for effect sizes across studies. Bolded values indicate effect sizes greater than the What Works Clearinghouse benchmark for a substantively important effect size (greater than 0.25 or less than –0.25; U.S. Department of Education, 2014b). See tables B2 and B3 in appendix B for the statistical significance of the effects in individual studies.

a. The column sum (20) does not equal the total number of studies reviewed (22) because 3 studies (Craven, Marsh & Debus, 1991; Fuchs, Butterworth & Fuchs, 1989; and Mostow et al., 2003) did not provide enough information to calculate effect sizes and 1 study (McCurdy & Shapiro, 1992) is included in the total in both rows because it examined both student-directed and other-directed formative assessment.

b. The number of effect sizes is greater than the number of studies because eight studies included more than one comparison for which effect sizes could be calculated (Fuchs, Fuchs, Hamlett, & Allinder, 1991a, 1991b; Fuchs, Fuchs, Hamlett, & Ferguson, 1992; Fuchs, Fuchs, Hamlett, & Stecker, 1991; Johnson, Graham, & Harris, 1997; Martens, Eckert, & Begeny, 2007; McCurdy & Shapiro, 1992; and Ysseldyke & Tardrew, 2007).

Source: Authors' analysis of studies published between 1988 and 2014; see appendix A for details.

Table 3. Mean effect sizes for formative assessment in math, by type

Type of formative assessment	Number of studies	Number of effect sizes[a]	Mean effect size	Standard deviation	Minimum effect size	Maximum effect size
Student-directed	4	4	**0.45**	0.49	–0.18	**1.01**
Other-directed	3	6	**0.30**	0.21	0.07	**0.66**

Note: The table presents descriptive statistics for effect sizes across studies. Bolded values indicate effect sizes greater than the What Works Clearinghouse benchmark for a substantively important effect size (greater than 0.25 or less than –0.25; U.S. Department of Education, 2014b). See tables B2 and B3 in appendix B for the statistical significance of the effects in individual studies.

a. The number of effect sizes is greater than the number of studies for other-directed formative assessment because two studies (Fuchs, Fuchs, Hamlett, & Stecker, 1991, and Ysseldyke & Tardrew, 2007) included more than one comparison for which effect sizes could be calculated.

Source: Authors' analysis of studies published between 1988 and 2014; see appendix A for details.

Box 4. Examples of student-directed and other-directed formative assessment in math for which substantively meaningful positive effects were found

Flash card assessment and progress monitoring with peers (Menesses & Gresham, 2009). This study used a student-directed formative assessment for grade 2–4 students in general education that focused on building student fluency and accuracy with basic math facts. Student pairs implemented the two-part sessions. The first part of each session was three minutes of practice; the second part of each session was formative assessment. The formative assessment required peers to assess the facts just practiced by following a protocol of presenting flash cards, counting the number of correct facts, and recording the number of correct facts on a chart. To interpret and use the assessment information formatively, a decision rule was applied: when a student provided 10 correct facts during the assessment in two consecutive sessions, that student was introduced to a new set of flash card facts in the next session. The sessions occurred three times weekly for a total of 15 sessions.

Accelerated Math (Ysseldyke & Tardrew, 2007). This study used an other-directed formative assessment in math. Students took the computer-adaptive Star Math test, and on the basis of the results of the test, teachers set instructional objectives for each student and assigned them to appropriate "libraries" of material. The computer program generated worksheets containing random sets of problems from each student's library. Students completed the worksheets and recorded their answers on a scan sheet. When students' worksheets were scanned, information about their performance was added to the instructional management system. Teachers received daily reports about how students were progressing toward their customized instructional goals. The daily report flagged students who appeared to be experiencing the most difficulty so that teachers could design other interventions for these students.

In reading, other-directed formative assessment was more effective than student-directed formative assessment

The average effect size for student-directed formative assessment in reading was small and negative (–0.15). The average effect size for other-directed formative assessment in reading was positive and substantively important, at 0.41 standard deviation

Nine studies examined formative assessment in reading. These studies tested 12 comparisons for which effect sizes could be calculated. Four comparisons examined the impact of student-directed formative assessment, and eight examined the impact of other-directed formative assessment. The average effect size for student-directed formative assessment was small and negative (–0.15; table 4). The average effect size for other-directed formative assessment was positive and substantively important, at 0.41 standard deviation. See box 5 for descriptions of other-directed formative assessment in reading.

In writing, other-directed formative assessment did not have substantively important effects, and not enough evidence was available to determine the effectiveness of student-directed formative assessment

Seven studies examined formative assessment in writing, but only one examined student-directed formative assessment (Sawyer et al., 1992; described in box 6), making it impossible to examine the differential effectiveness of student-directed and other-directed formative assessment. The one study investigating student-directed formative assessment had the largest effect observed for a writing formative assessment (0.63), suggesting a need for more research on student-directed formative assessment in writing to determine whether this finding can be replicated. The average effect size for the other-directed formative assessment was low (0.15), with individual effect sizes ranging from –0.20 to 0.34.

Table 4. Mean effect sizes for formative assessment in reading, by type

Type of formative assessment	Number of studies[a]	Number of effect sizes[b]	Mean effect size	Standard deviation	Minimum effect size	Maximum effect size
Student-directed	2	4	–0.15	0.26	**–0.46**	0.17
Other-directed	8	8	**0.41**	0.41	0.02	**1.22**

Note: The table presents descriptive statistics for effect sizes across studies. Bolded values indicate effect sizes greater than the What Works Clearinghouse benchmark for a substantively important effect size (greater than 0.25 or less than –0.25; U.S. Department of Education, 2014b). See tables B2 and B3 in appendix B for the statistical significance of the effects in individual studies.

a. The column sum (10) does not equal the total number of studies reviewed that focused on reading (9) because one study (McCurdy & Shapiro, 1992) examined both student-directed and other-directed formative assessment.

b. The number of effect sizes is greater than the number of studies because two studies of student-directed formative assessment in reading (Johnson, Graham, & Harris, 1997, and McCurdy & Shapiro, 1992) and two studies of other-directed formative assessment in reading (Fuchs, Fuchs, Hamlett, & Ferguson, 1992, and Martens, Eckert, & Begeny, 2007) included more than one comparison for which effect sizes could be calculated.

Source: Authors' analysis of studies published between 1988 and 2014; see appendix A for details.

Box 5. Examples of other-directed formative assessment in reading for which substantively meaningful positive effects were found

Curriculum-based measurement (Fuchs, Fuchs, & Hamlet, 1989). Teachers selected an end-of-year goal for each of their students. Teachers then administered a reading comprehension probe at least twice weekly and recorded scores in a data management program. After 7–10 scores were entered, the data management program presented the teacher with a line graph depicting the student's current and projected progress (if the student continued making gains at the same pace as was demonstrated thus far) and an aim line (a straight line from the starting point to the goal). If the student's progress line was less steep than the aim line, teachers were prompted to change their instructional strategy, collect 7–10 more scores, and then re-evaluate. If it was steeper, they were prompted to raise the goal, collect 7–10 more scores, and then re-evaluate.

Mastery learning (Null, 1990). Teachers delivered lessons on decoding skills to students in a large-group format. After delivering the lessons, teachers administered a formative assessment. Students who scored 80 percent or higher on the assessment participated in enrichment activities, and students who scored less than 80 percent participated in reteaching activities. Next, teachers administered a parallel formative assessment to the students participating in the reteaching.

Box 6. Example of a self-directed formative assessment in writing for which a substantively meaningful positive effect was found

Self-regulation (Sawyer et al., 1992). Students first wrote a story that served as a pretest. Next, they received instruction in a five-step writing strategy. After this instruction, students were told how many story elements they included in their pretest story. Students met individually with their instructors and discussed the goal of including all of the story elements in their stories. Students practiced writing stories. After writing each story, students counted the number of story elements they included, recorded the result on a progress chart, and evaluated their progress toward the goal of including all the story elements.

Implications of the study findings

The results of this study confirm the overall positive effect of formative assessment reported in earlier reviews (Black & Wiliam, 1998a, 1998b; Kingston & Nash, 2011, 2015). This consistency with previous reviews, along with the requirement that studies meet evidence standards to be included, lends continuing support to the claim that formative assessment has a positive impact on student academic achievement. Findings from this study and from previous research (Kingston & Nash, 2011, 2015) indicate that the effectiveness of formative assessment varies by subject area, with larger effects when formative assessment was used during instruction in math than when it was used during instruction in reading or writing.

The results of this study can help teachers and administrators identify approaches to formative assessment that are appropriate and effective for particular subject areas in the elementary school grades. Student-directed formative assessment, including self- and peer assessment, were effective, on average, for math instruction. Educator- or computer program–directed approaches were effective, on average, for math and reading. Studies of formative assessment in writing focused almost exclusively on other-directed formative assessment and, on average, did not have substantively important effects on student outcomes. More research on formative assessment in writing is needed.

The results of this study confirm the overall positive effect of formative assessment reported in earlier reviews and lends continuing support to the claim that formative assessment has a positive impact on student academic achievement

Limitations of the study

The findings presented here are restricted to studies published from 1988 to 2014. Despite attempts to locate unpublished reports (see appendix A), it is possible that some relevant studies were missed. Further, reports did not always include all the information needed to rate studies and calculate effect sizes. In these cases the review team tried to contact all the study authors to request the missing information, but some authors could not be located, some no longer had access to the required information, and some did not reply.

The findings are limited to general education students in the elementary school grades and to students receiving special education instruction on elementary academic content. The review is also limited by inclusion of only studies that have comparative group designs. A review of impact studies using different designs, such as single-case designs or regression discontinuity, may yield different findings.

Moreover, formative assessment has different characteristics and types that might account for differences in its effectiveness. This review may not have examined all the influential characteristics. For example, not enough studies were located to compare the effectiveness of different features within the student-directed and other-directed categories (for example, self-directed versus peer-directed formative assessment or teacher-directed versus computer-directed formative assessment). Similarly, this review did not examine the frequency of formative assessment, which could influence its effectiveness, because the studies reviewed did not consistently describe frequency.

This review is further limited in that the estimates of the average effect size for different types of formative assessment are drawn, for the most part, from different studies. The differences observed between self- and other-directed formative assessment may be due to other differences between the studies examined, such as differences in populations studied

or other characteristics of the interventions or differences in the fidelity with which the interventions were implemented. A more rigorous test of the difference between self- and other-directed formative assessment would come from comparisons of these two types within the same study; however, only one study (McCurdy & Shapiro, 1992) compared both student- and other-directed formative assessment.

Although this review adds to the evidence base on the impact of formative assessment on student academic achievement, the findings are limited by the number of studies that were determined to have met the criteria of topic relevance, standards, and subject areas. Future research is needed to extend the evidence base, especially in grades K–3, for which the smallest number of studies on formative assessment that met standards were identified. More rigorous research is needed on the effects of formative assessment in science and social studies, because no studies investigating these subject areas that met standards were located. Across elementary school grades, future research is needed in the core academic subjects—math, reading, and writing—especially on the effectiveness of formative assessment on foundational skill development that prepares students to successfully progress on college- and career-readiness trajectories.

More rigorous research is needed on the effects of formative assessment in science and social studies, because no studies investigating these subject areas that met standards were located

Appendix A. Methodology

This appendix describes the literature search, screening process, evidence review and study rating, characterization of study findings, and analysis approach.

Literature search

The review team used the following six search strategies to identify study reports.

1. Fourteen electronic academic literature databases were searched between January and May 2014 using search strings combining intervention and study design keywords (table A1):
   - Academic Search Premier.
   - Business Source Corporate.
   - Campbell Collaboration.
   - Dissertation Abstracts International
   - EconLit.
   - Education Research Complete.
   - Education Resources Information Center.
   - EJS E-Journals.
   - Google Scholar.
   - PsycINFO.
   - ScienceDirect.
   - SocINDEX with Full Text.
   - What Works Clearinghouse (WWC) database of studies.
   - WorldCat.

2. Tables of contents of 13 journals published between 2004 and 2014 were searched to identify relevant literature.
   - *American Educational Research Journal.*
   - *Applied Measurement in Education.*
   - *Assessment for Effective Intervention.*
   - *Assessment in Education: Principles, Policy & Practice.*
   - *British Journal of Educational Psychology.*
   - *Educational Assessment.*
   - *Educational Assessment, Evaluation and Accountability.*
   - *Educational Measurement: Issues and Practice.*
   - *Educational Studies in Mathematics.*
   - *Educational Psychology in Practice.*
   - *Journal of Educational Computing Research.*
   - *Journal of Research in Reading.*
   - *Learning and Individual Differences.*

3. Numerous websites were searched to identify relevant white papers or reports:
   - Abt Associates.
   - Alliance for Excellent Education.
   - American Educational Research Association.
   - American Enterprise Institute.
   - American Institutes for Research (Regional Educational Laboratory Midwest).

Table A1. Keywords used to search academic literature databases

Concept	Keywords
Formative assessment as process	Formative* AND assessment*
	"Assessment* for learning"
	"Diagnostic assessment*"
	"Formative evaluation*"
	"Mastery learning"
	"Classroom questioning"
	"Curriculum based assessment*" OR "curriculum-based assessment*"
Goal setting	"Learning goal*"
	Learning AND (intention* OR trajectory OR progression)
	"Learning objective*"
	"Learning target*"
	Student* AND "goal setting"
Student involvement	"Student* involvement"
	Student* AND "self-monitoring"
	Student* AND "self-assessment*"
	Student* AND "self-direct*"
	Student* AND "self-regulat*"
	Student* AND reflect*
	"Peer assessment*"
Feedback	"Feedback to student*" OR "feedback for student*"
	"Progress chart*"
	"Progress graph*"
	"Progress monitoring"
	Rubric*
	"Proficiency scale*"
Using assessment	"Instructional adjustment*"
	"Instructional scaffolding"
	"Adapt* instruction"
Design	Random* OR RCT
	Quasi-experiment* OR QED
	Experiment* OR impact OR effectiveness OR causal
	Posttest OR post-test OR pretest OR pre-test
	"Efficacy trial"
	"Multisite" OR "multi*site"
Comparison	Treatment
	"Control group*" OR "comparison group*" OR "matched group*"
	Equivalence
	Baseline
	"Propensity score"

* Used in the search term to indicate that any number of characters can be substituted in place of the asterisk (wildcard).

Source: Authors' compilation.

- Best Evidence Encyclopedia.
- Brookings Institution.
- Carnegie Corporation.
- Carnegie Foundation for the Advancement of Teaching.

- Center for Research and Reform in Education at Johns Hopkins University.
- CNA Analysis and Solutions (Regional Educational Laboratory Appalachia).
- Cochrane Central Register of Controlled Trials.
- Cochrane Database of Systematic Reviews.
- Congressional Research Service.
- CRESST (National Center for Research on Evaluation, Standards and Student Testing).
- Database of Abstracts of Reviews of Effects.
- Education Development Center, Inc. (Regional Educational Laboratory Northeast and Islands).
- Education Northwest (Regional Educational Laboratory Northwest).
- Education Resources Institute.
- Florida State University (Regional Educational Laboratory Southeast).
- Government Accountability Office.
- Grants and contracts awarded by Institute of Education Sciences (not Regional Educational Laboratories).
- Heritage Foundation.
- Hoover Institution.
- ICF International (Regional Educational Laboratory Mid-Atlantic).
- Marzano Research (Regional Educational Laboratory Central).
- Mathematica Policy Research.
- MDRC.
- Mid-Continent Research for Education and Learning (Regional Educational Laboratory Pacific).
- National Association of State Boards of Education.
- National Governors Association.
- PolicyArchive.
- Policy Studies Associates.
- Promising Practices Network.
- RAND Corporation.
- SEDL (Regional Educational Laboratory Southwest).
- SRI International.
- Thomas B. Fordham Institute.
- WestEd (Regional Educational Laboratory West).
- Urban Institute.

4. Using the names of 29 interventions as keywords, two electronic bibliographic databases (ERIC and PsycINFO) and the intervention reports on the WWC publications and products website were searched. Intervention names included:
   - 6 + 1 Trait Writing Model.
   - Aligned Developmental Feedback.
   - Assessment and Learning in Knowledge Spaces.
   - ASSISTment System.
   - Calibrated Peer Review.
   - Convince Me.
   - Copy-Cover-Compare.
   - Cumulative Writing Folder Program.
   - Empowered Curriculum.
   - Feedback Dialogue.

- Functional Assessment, Collaboration, and Evidence-Based Treatment.
- GenScope.
- Interactive Strategies Approach.
- KWL.
- Mathetics.
- Mindtools.
- Pathfinder Networks.
- Peer-Assisted Learning Strategies.
- Reading Edge.
- Reflective Assessment.
- Romance Project/Foundational Approaches in Science Teaching (FAST).
- Self-Directed Learning.
- Self-Regulated Strategy Development.
- Six-Trait Rubric.
- Social Cognitive Model of Sequential Skill Acquisition.
- Thinking Actively in an Academic Context.
- Webquests.
- Write Score.
- WriteToLearn.

5. The contents and reference lists of 28 literature reviews and practitioner-oriented books on formative assessment were examined to identify relevant studies (Andrade & Cizek, 2010; Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Bennett, 2011; Black & Wiliam, 1998a, 2003, 2009; Boekaerts, Pintrich, & Zeidner, 2000; Briggs, Ruiz-Primo, Furtak, Shepard, & Yin, 2012; Brookhart, 2008, 2010; Butler & Winne, 1995; Clark, 2012; Filsecker & Kerres, 2012; Hattie & Timperley, 2007; Heritage, 2010b, 2013; Kingston & Nash, 2011, 2012; Kluger & DeNisi, 1996; Marzano, 2010; McMillan, Venable, & Varier, 2013; Moss & Brookhart, 2009; National Research Council, 2001; Noyce & Hickey, 2011; Sadler, 1989; Shepard, 2005; Shute, 2008; Wiliam, 2011).

6. Researchers and experts in formative assessment were consulted to identify possible missing studies.

The principal investigator emailed five formative assessment researchers and experts with the list of studies being considered and asked for suggestions of missing studies. The suggestions provided three additional studies to be screened for this review.

Screening process

To eliminate studies not relevant to the review, studies were screened against a set of seven relevance criteria (table A2). Members of the review team trained in applying the relevance criteria carried out a three-phase screening process. This section describes each phase and provides information about the reliability of the process.

In phase I, the reviewers focused on report titles and abstracts and screened for topic, sample, timeframe, and outcome relevance. The reviewers were instructed to keep any article that appeared to be relevant and to drop any duplicates. In addition, potentially relevant literature reviews were identified and screened-in, resulting in 2,622 reports, including literature reviews, that were retained for phase II screening.

Table A2. Relevance criteria used for screening formative assessment studies

Criterion	Description
Topic	The study focused on the effects of a formative assessment intervention. Formative assessment interventions were defined as those with a process dedicated to both gathering and using assessment information to facilitate student learning within a cycle lasting four or fewer weeks. The review includes interventions that are replicable, including programs, practices, strategies, or activities implemented by teachers, students, or both. Included interventions were dedicated to both gathering and using assessment information to facilitate student academic learning. Gathering assessment information is defined as seeking or eliciting evidence of student knowledge, understanding, or behavior. Using assessment information is defined as having the explicit opportunity to apply the information to facilitate student learning. Having the explicit opportunity to use assessment information includes, for example, a follow-up learning or assessment activity provided to students that addresses the same or related learning goal or performance task and time and guidance provided to teachers for both interpreting assessment information and choosing instructional options. Interventions that combined formative assessment with another intervention, such as providing direct instruction in strategies for writing narratives or modeling of fluent reading (bundled interventions), were included only if the instruction was contingent on the evidence gathered through formative assessment (that is, part of the formative assessment cycle) or the effects of the formative assessment were isolated and presented separately.
Timeframe	The study was published no earlier than 1988.
Student sample	The sample included students enrolled in kindergarten–grade 6. The sample can include students in grades higher than grade 6 if the majority of students in the sample are in the K–6 grade band. In addition to students, the participants involved in the formative assessment intervention may include teachers, instructional coaches, staff developers, parents, counselors, afterschool program staff, tutors, other students, or others.
Language	The study report was written in English.
Location	The study location was limited to the United States or any country that is similar enough to the United States to permit replication of the study in the United States. Countries were considered similar to the United States if English is the societal language. This is the same location criterion as the What Works Clearinghouse used for its Beginning Reading Interventions protocol (U.S Department of Education, 2014a).
Subject area context	The intervention had to take place in an authentic education setting during the regular school day. The intervention must have addressed one or more academic content or practice standards, including standards in math, reading, writing, science, and social studies.
Study design	Studies were empirical, using quantitative methods and inferential statistical analysis, and used a group comparison research design, either a randomized controlled trial or quasi-experimental design.

Source: Authors' compilation.

In phase II the reviewers screened full study reports, including potentially relevant literature reviews, for relevance of topic, timeframe, student sample, language, location, study design, and outcome. As a result of phase II screening, 716 study reports were screened-in.

In phase III the reviewers narrowed the scope of the review in three ways to match limited resources to the time available. First, the principal investigator determined that the 251 difficult-to-obtain, "maybe screened-in" studies would be excluded from further consideration because of the extensive resources that would be required to locate and screen studies not likely to meet relevancy criteria. Second, although the review originally intended to focus on grades K–12 and a wide range of subject areas, the literature search and screening process yielded too many studies to complete the relevance screen, evidence rating, and analysis of effects with the available time and resources. Therefore, the focus was narrowed

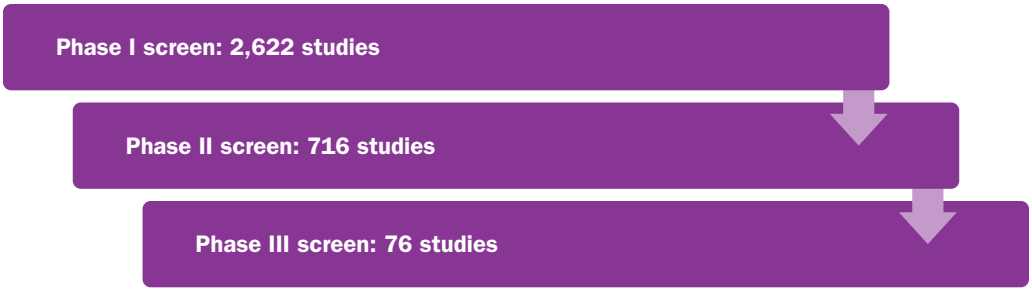
to the elementary school grades, eliminating studies involving students in grades 7–12 for this review and focusing on interventions in the core subject areas of math, reading, writing, science, and social studies. Third, studies that measured only nonacademic outcomes were eliminated (figure A1).

Three reliability checks were conducted during the 17-month (January 2014–May 2015) study identification and screening process in months 1, 6, and 15. The reliability checks showed a high percentage of agreement in screening decisions (table A3). At month 1, when six reviewers each independently screened 10 studies, 85 percent of the decisions were in agreement. At month 6, when the members of six pairs of reviewers independently screened 11 studies, 90 percent of decisions were in agreement. At month 15, when every 10th screened-out study had been identified for a second screening, the two independent reviewers agreed on 91 percent of the decisions (for 32 of the 35 studies).

Separating effects of instruction from formative assessment in selecting and reviewing studies

During the keyword search, initial screening, and WWC-certified researcher training, a decision rule was developed to screen out studies of interventions that bundled formative assessment with instruction. To distinguish effects of instruction from the effects of formative assessment, only studies of interventions in which instruction was contingent in design or content on the evidence gathered, and therefore part of the formative assessment cycle, were included. Studies were excluded if they evaluated an intervention that involved instruction that did not meet this criterion (for example, lesson scripts, instructional activities, and instructional materials that are included, not as formative assessment contingencies, but rather as part of a predefined scope and sequence for teaching particular content). This decision rule reflects the definition of formative assessment used in this

Figure A1. Study yields from each phase of the screening for formative assessment studies



Source: Authors' compilation.

Table A3. Agreement between independent screening decisions

Month	Number of studies independently screened by two researchers	Percentage of screening decisions with agreement
1	10	85
6	11	90
15	35	91

Source: Authors' analysis of studies published between 1988 and 2014.

study: it is a process dedicated to both gathering and using assessment information about student learning to facilitate further learning within a cycle lasting four or fewer weeks (see box 1 in the main report).

Studies of multicomponent interventions that were included because the instructional components were contingent on the formative assessment information gathered included three examples:

- In a fluency-building program the instructional components, such as following along while the instructor modeled fluent reading, were selected for implementation contingent on the assessment information gathered and thus were part of the formative assessment (Martens et al., 2007). The passage selected for each individual or small-group session was a passage on which "students read below 100 [correct words per minute] based on the pre-training assessment" (Martens et al., 2007, p. 45).
- A study comparing two types of formative assessment, human and computer-assisted tutoring, to each other and to a business-as-usual control group met the topic relevance criterion because the tutor gathered and used assessment information, including providing feedback and support to students contingent on the assessment information gathered (Mostow et al., 2003).
- The Accelerated Math software package, which provides ongoing feedback to students and teachers based on student performance, met the topic relevance criterion because the software "instantly scores, records student performance, updates teacher record books, creates teacher reports, creates immediate student feedback reports, and generates the next assignment for the student" (Ysseldyke & Tardrew, 2007, p. 6). The system's daily report shows each student's academic standing and "flags students who are experiencing difficulty and indicates the teacher should intervene" (p. 6). The program helps teachers group students contingent on the information provided; teachers have the opportunity to group "students who are at similar skill levels and/or experience similar learning difficulties" (p. 6). The software provides instructional resources, guidance, and practice as part of the formative assessment cycle.

In some studies the experimental design isolated the effect of formative assessment in an intervention that bundled formative assessment with instruction that was not contingent on the formative assessment information gathered. For example, in a reading study an intervention condition in which students received instruction on goal-setting and self-monitoring as a formative assessment that was bundled with reading strategy instruction was compared with a comparison condition in which students received the reading strategy instruction alone (Johnson, Graham, & Harris, 1997). Because formative assessment is added to instruction and compared to that instruction provided alone, the design of the study isolates the effects of formative assessment.

Evidence review and study rating

Nine reviewers who were certified by the WWC assessed the quality of the screened-in studies using the study review protocol and the *What Works Clearinghouse Procedures and Standards Handbook, Version 3.0* for group designs (U.S. Department of Education, 2014b). The reviewers documented information about the intervention, study design, participant sample, context, and outcome measures; confirmed topic relevance; confirmed that the

study met the protocol's study design and outcome relevance; and determined whether the study met WWC standards with or without reservations. The WWC standards include two options for assessing the extent of attrition in a randomized controlled trial.[4] As specified in the study protocol, the WWC liberal criterion for assessing attrition was used to evaluate whether a randomized controlled trial had low attrition. For each study, the reviewers also completed the data tabs in the WWC Study Review Guide to complete the calculations necessary to determine a study's rating (for example, attrition and baseline equivalence) and to calculate effect sizes (U.S. Department of Education, 2014c).[5] When a study had more than one comparison with outcomes in the same outcome domain, the WWC Study Review Guide was used to calculate a simple unweighted average effect size across the comparisons. The reviewers conducted author queries when information was missing but necessary for the rating or to compute effect sizes. For 3 of the 23 studies that met WWC standards, the reviewers were unable to obtain enough information to calculate effect sizes. These three studies included 8 of the 46 total comparisons.

When studies included students or classrooms in multiple grades but assigned them to a treatment condition and analyzed data separately for each grade level, each comparison was examined separately against the WWC standards and effect sizes were not averaged across the comparisons. This is consistent with the procedures used by the WWC, which defines a study based on which participants were randomized at the same time for randomized controlled trials and by whether samples are independent for quasi-experimental designs (U.S. Department of Education, 2014b).

A large proportion of the studies that were located (nine studies, or 39 percent of the studies that met standards) compared more than one type of formative assessment. These studies often compared different types of formative assessment with each other and with a no-intervention comparison group. These multiple studies were considered as one study following the guidance in the *What Works Clearinghouse Procedures and Standards Handbook,* version 3.0 (U.S. Department of Education, 2014b).[6] However, each relevant comparison was evaluated separately for quality of evidence, and each comparison was assigned a separate evidence rating. This report focuses primarily on comparisons between formative assessment group and no-intervention comparison groups.

All studies with comparisons that were determined to have met WWC standards with or without reservations and all studies with comparisons for which the study rating depended on the answer to an author query were examined by a second reviewer. The two reviewers met to discuss and resolve any disagreements. In all cases, reviewers were able to reach consensus. Ten percent of studies that the first reviewer determined did not meet standards were subject to a second review. Of these studies, there was agreement on 80 percent. For the remaining studies that did not meet standards, the principal investigator examined the study review to validate the rating.

Characterization of study findings

Of the 76 studies identified through screening as eligible for evidence review, the reviewers determined that 23 studies met standards with or without reservations and 53 studies did not meet standards (see appendix D for a list of studies that did not meet standards and why).

For studies that met standards, researchers characterized each intervention effect into one of five mutually exclusive categories based on the effect size and statistical significance of the effect (table A4). Using the WWC benchmark for substantively important effect sizes (U.S. Department of Education, 2014b), the reviewers characterized effect sizes for academic outcomes that are equal to or greater than 0.25 as "substantively important positive effects." Effect sizes that are less than or equal to –0.25 were characterized as "substantively important negative effects" (U.S. Department of Education, 2014a, p. 26). "Statistically significant" effects are effects large enough, given the sample size, to likely not have occurred by chance. Effects that are deemed neither large enough to be substantively important nor statistically significant were characterized as indeterminate (U.S. Department of Education, 2014b). If a study did not provide enough information to calculate effect sizes, study findings could be characterized only if they were statistically significant. When an outcome had two or more intervention effects in the same domain, rules were applied to characterize the effect (see table A4).

Table A4. Criteria for characterizing formative assessment effects that met standards modeled after those used by the What Works Clearinghouse

Characterization	Criteria to meet rating, after applying any needed corrections
Statistically significant positive effect	• If one outcome, the estimated effect is positive and statistically significant. • If more than one outcome in the same domain, the average effect size across the domain's outcome measures is positive and statistically significant. • If more than one outcome in the same domain and information to calculate effect sizes is not available, in that domain at least half of the effects are positive and statistically significant and no effects are negative and statistically significant.
Substantively important positive effect	• If one outcome, the estimated effect is greater than 0.25 but not statistically significant. • If more than one outcome in the same domain, the average effect size across that domain's outcome measures is greater than 0.25 but not statistically significant.
Indeterminate effect	• If one outcome, the estimated effect is between 0.25 and –0.25 and not statistically significant. • If more than one outcome in the same domain, the average effect size across that domain's outcome measures is between 0.25 and –0.25 and not statistically significant.
Substantively important negative effect	• If one outcome, the estimated effect is less than –0.25 but not statistically significant. • If more than one outcome in the same domain, the average effect size across that domain's outcome measures is less than –0.25 but not statistically significant.
Statistically significant negative effect	• If one outcome, the estimated effect is negative and statistically significant. • If more than one outcome in the same domain, the average effect size across that domain's outcome measures is negative and statistically significant. • If more than one outcome and information to calculate effect sizes is not available, in that domain at least half of the effects are negative and statistically significant and no effects are positive and statistically significant.

Source: Authors' compilation based on U.S. Department of Education (2014b).

Analysis approach

To address the research questions about the magnitude of the effect of formative assessment, effect sizes were averaged across studies. This approach deviates from the WWC approach to combining results across studies of interventions (U.S. Department of Education, 2014b). Unlike WWC intervention reports, which focus on one discrete intervention, this report aims to summarize effects across a variety of interventions with similar characteristics to draw conclusions about the effectiveness of a broader approach to instruction. In this case, simple averages of effect size are more interpretable and easier to understand.

Appendix B. Detailed research findings

This appendix provides detailed findings. For the studies that reviewers determined met standards, table B1 presents descriptions of the interventions and tables B2 and B3 present review findings; in each table, information is identified by study citation listed alphabetically.

Table B1. Descriptions of formative assessment interventions in studies that met standards modeled after those used by the What Works Clearinghouse with or without reservations, by intervention type

Study citation	Description of study conditions (intervention and comparison groups)	Intervention type
Abrami, Venkatesh, Meyer, & Wade (2013)	Digital portfolio. This web-based system, ePEARL, prompts students and manages information through three phases of self-regulation: plan (set goals, plan strategies, create learning logs), perform (create work; self-examine through recordings, drafts, and learning log entries), and self-reflect (reflect on work, process, and feedback, and adjust future goals). Teachers and peers also review and provide feedback on work through entries in the portfolio system. Students in the comparison group did not use ePEARL; teachers used business-as-usual instruction.	Other directed
Bond & Ellis (2013)	Self-assessment. During the four weeks of the intervention, teachers delivered scripted lessons on probability and statistics. At the end of each class session, students in the intervention group wrote a sentence that started with "I learned," discussed their completed sentence with another student ("think aloud"), then revised their "I learned" sentence. This took five minutes. The intervention was intended to facilitate student reflection on what they learned; teachers collected the statements and submitted them to the researchers daily. Teachers in the control group delivered instruction using the same scripted lessons and spent the last five minutes of class reviewing material.	Student directed
Craven, Marsh, & Debus (1991)	Feedback. In small-group pull-out sessions or in the regular class, students were invited to identify their strengths and weaknesses in math and reading and attribute their math and reading accomplishments to ability and effort, followed by receiving contingent feedback that attributed their success or failure on reading and math tasks to ability or effort. The math and reading activities and feedback occurred for eight weeks. Teachers in the comparison group delivered business-as-usual instruction.	Other directed
Cripps (1995)	Flash-card assessment and progress monitoring with peers and curriculum-based measurement. Two types of peer tutoring interventions were compared to each other. In both groups peers used flash cards to help another student learn multiplication facts. In addition, the intervention group used curriculum-based measurement probes; progress graphing; discussion of the graph, which included a star to represent the student's goal; and encouragement to keep working toward the star.	Student directed
Fuchs, Butterworth, & Fuchs (1989)	Curriculum-based measurement. One intervention group used computer-assisted curriculum-based measurement, and the other group used pen-and-paper curriculum-based measurement. Teachers assessed each student's spelling performance at least twice weekly and graphed correct words and letter sequences. Decision rules were made using the graphed letter sequence scores to guide development of students' instructional programs. Teachers in the computer-assisted curriculum-based measurement group used software to enter data. The software created graphs, applied decision rules, and provided feedback statements to teachers summarizing the decisions. Teachers in the pen-and-paper group graphed scores and applied decision rules by hand. Teachers in the comparison group received training on how to support the improvement of students' spelling skills. The training did not include any specific procedures for using student data to guide instructional decisions.	Other directed
Fuchs, Fuchs, & Hamlett (1989)	Curriculum-based measurement. In a reading study, researchers identified two intervention groups by post hoc inspection of study records. Teachers who made at least one instructional modification in response to the curriculum-based measurement data became the "measurement + evaluation" curriculum-based measurement group. Teachers who implemented curriculum-based measurement but did not modify instruction became the "measurement only" curriculum-based measurement group. Teachers in the comparison group were instructed to write goals for students and to use their typical procedures for monitoring student progress.	Other directed

(continued)

Table B1. Descriptions of formative assessment interventions in studies that met standards modeled after those used by the What Works Clearinghouse with or without reservations, by intervention type *(continued)*

Study citation	Description of study conditions (intervention and comparison groups)	Intervention type
Fuchs, Fuchs, Hamlett, & Allinder (1991a)	Curriculum-based measurement. This study examined the effects of two types of curriculum-based measurement: curriculum-based measurement with computerized, expert system instructional consultation and curriculum-based measurement alone, with no expert consultation. In both curriculum-based measurement interventions, teachers determined end-of-year spelling goals for individual students, monitored progress toward the goal by having students complete an assessment at the computer twice weekly, evaluated the assessment data weekly by using software to graph individual scores, applied decision rules, accessed skills analyses results, and implemented instructional changes accordingly. In the curriculum-based measurement with expert consultation intervention, teachers entered additional information into the computerized system and received recommendations regarding instructional changes. In the curriculum-based measurement with no consultation, teachers determined instructional changes on their own. Teachers in the comparison group were instructed to use their typical procedures for monitoring student progress.	Other directed
Fuchs, Fuchs Hamlett, & Allinder (1991b)	Curriculum-based measurement. This study examined the effects of two types of curriculum-based measurement: curriculum-based measurement with a computerized, expert system of skills analysis and curriculum-based measurement alone, with no computerized expert system of skills analysis. In both curriculum-based measurement interventions, teachers determined end-of-year spelling goals for individual students, monitored progress toward the goal by having students complete an assessment at the computer twice weekly, evaluated the assessment data weekly by using software to graph individual scores, applied decision rules, and implemented instructional changes accordingly. Group 1 teachers received skills analysis, and group 2 teachers did not. Teachers in the comparison group did not use curriculum-based measurement. Instead, they set goals using standard individualized education plan forms and monitored progress toward those goals as they ordinarily would have.	Other directed
Fuchs, Fuchs, Hamlett, & Ferguson (1992)	Curriculum-based measurement. This study examined the effects of two types of curriculum-based measurement: curriculum-based measurement with computerized, expert system instructional consultation and curriculum-based measurement alone, with no consultation. In both curriculum-based measurement interventions, teachers determined end-of-year reading goals for individual students, monitored progress toward the goal by having students complete a passage reading assessment at the computer twice weekly, evaluated the assessment data weekly by using software to graph individual scores, applied decision rules, and implemented instructional changes or not accordingly. In the curriculum-based measurement with consultation intervention, teachers entered additional information into the computerized system and received recommendations regarding instructional changes. In the curriculum-based measurement with no consultation, teachers determined instructional changes on their own. Teachers in the comparison group delivered business-as-usual instruction.	Other directed
Fuchs, Fuchs, Hamlett, & Stecker (1991)	Curriculum-based measurement. This study examined the effects of two types of curriculum-based measurement: curriculum-based measurement with computerized, expert system instructional consultation and curriculum-based measurement alone, with no consultation. In both curriculum-based measurement interventions, teachers determined end-of-year math goals for individual students, monitored progress toward the goal by having students complete an assessment at the computer twice weekly, evaluated the assessment data weekly by using software to graph individual scores, applied decision rules, accessed skills analyses results, and implemented instructional changes accordingly. In the curriculum-based measurement with consultation intervention, teachers entered additional information into the computerized system and received recommendations regarding instructional changes. In the curriculum-based measurement with no consultation, teachers determined instructional changes on their own. Teachers in the comparison group delivered business-as-usual instruction.	Other directed
Graham & Harris (1989)	Self-regulation. Students in all groups were taught "self-instructional strategy training." Only the students in the self-regulation intervention engaged in the following formative assessment process: students received a teacher-review of their charted pretest performance, set goals in terms of story grammar elements, wrote a story, self-assessed the story elements in relation to their goals, and set the same or different goals for their next story.	Other directed

(continued)

Table B1. Descriptions of formative assessment interventions in studies that met standards modeled after those used by the What Works Clearinghouse with or without reservations, by intervention type *(continued)*

Study citation	Description of study conditions (intervention and comparison groups)	Intervention type
Iannuccilli (2003)	Feedback and progress monitoring. Every two weeks, teachers received a graph of students' individual performance on two Dynamic Indicators of Basic Early Literacy Skills subtests and descriptive feedback on the types of correct and incorrect responses each individual student made in phonemic segmentation and nonsense word fluency. Teachers in the comparison group delivered business-as-usual instruction.	Other directed
Johnson, Graham, & Harris (1997)	Self-regulation. The formative assessment intervention was added to lessons in which students were taught strategies for reading comprehension. The intervention had two variations: students who were taught to set strategy-use goals and monitor their progress toward attaining their goals, and students who were taught to set strategy-use goals and monitor their progress toward attaining their goals PLUS use self-instructional strategies (self-statements) to guide their use of the strategies they were taught. Students in the comparison group were received the same instruction on strategies for reading comprehension.	Student directed
Martens, Eckert, & Begeny (2007)	Curriculum-based measurement. In this curriculum-based measurement intervention, students stated a reading goal of reading 100 words correctly per minute, read a passage, graphed number of words read correctly per minute, and received contingent feedback on performance in relation to the goal. The feedback was as follows: "the experimenter stated either that the student met the goal and provided praise, or that the student did not meet the goal and could earn another ticket after practice or during the next training session" (Martens et al., 2007, p. 45). Depending on their reading performance, students either moved to a more difficult passage or corrected errors and practiced reading at the same level of passage difficulty. Students in the comparison group received business-as-usual instruction.	Other directed
McCurdy & Shapiro (1992)	Curriculum-based measurement. Students participated in curriculum-based measurement probes twice weekly. Students also reviewed graphs of their progress; the graphs tracked number of correctly read words per minute for every curriculum-based measurement probe. The intervention had three variations: peer-monitoring, in which the curriculum-based measurement was implemented by peers; self-monitoring, in which students themselves implemented the curriculum-based measurement, listening to an audio recording of their own reading and scoring and graphing their results themselves; and teacher-monitoring, in which the curriculum-based measurement was implemented by teachers. Teachers and peers were trained also to provide feedback, but no details are provided in the article about the nature of that aspect of the training. No training on how to interpret and use the assessment information was provided to the teachers, peers, or students in the self-monitoring condition. Teachers used business-as-usual approaches for monitoring the progress of students in the comparison group.	Student directed and other directed
Menesses & Gresham (2009)	Flash-card assessment and progress monitoring with peers. The formative assessment intervention was a form of classwide peer tutoring in which peer tutors used flash cards to assess, monitor progress, and provide feedback to another student (the target student) who was learning addition and subtraction or multiplication and division math facts. Although Menesses and Gresham's (2009) study had other experimental conditions, the classrooms assigned to the condition where students received peer tutoring that involved assessing, monitoring progress, and providing feedback was the intervention of interest for this review. In each session the peer tutor presented a math-facts flash card for three seconds to the target student, sorted cards into correct and incorrect piles depending on the student's response, and counted and recorded the number of correct facts with the target student. Students in the comparison group received typical math instruction with no peer tutoring.	Student directed
Meyer, Abrami, Wade, Aslan, & Deault (2010)	Digital portfolio. The intervention, ePEARL, is a web-based electronic portfolio tool that is designed to support self-regulation. The software uses three phases (forethought, performance, and self-reflection) that are designed to include metacognitive and motivational components. The software prompts students to set goals and share work to solicit feedback on drafts of work. Students are expected to use this feedback, along with self-reflection, to "adjust their goals for the next work" (Meyer et al., 2010, p. 86). In this study, students set learning targets, provided documentation of how their work progressed, and self-evaluated their work using the ePEARL system. Throughout this process, teachers provided feedback to students. Only "medium" and "high" implementing classrooms, defined by more frequent feedback, were included in the analysis. Students in the comparison group did not use ePEARL; teachers delivered business-as-usual instruction.	Other directed

(continued)

Table B1. Descriptions of formative assessment interventions in studies that met standards modeled after those used by the What Works Clearinghouse with or without reservations, by intervention type *(continued)*

Study citation	Description of study conditions (intervention and comparison groups)	Intervention type
Mostow et al. (2003)	Computer-facilitated assessment. The study compares human and computerized tutoring. The computerized tutor selected stories that were at students' level (based on how well they had read previous stories). It used voice recognition software to "listen" to students reading and provided support when students made errors. It supplied words, sounded words out, and drew attention to words that students skipped. The human tutors were supposed to provide responses similar to those of the computer tutor. Students worked with their computerized and human tutors for 20 minutes each day for a whole school year. Teachers in the comparison group delivered business-as-usual instruction.	Other directed
Null (1990)	Mastery learning. Teachers delivered lessons on decoding skills to students in a large-group format. In each lesson they administered formative assessments on the objectives of the lesson. Students who answered 80 percent or more of the items on the assessment correctly participated in enrichment activities while the rest of the students participated in reteaching activities. A second, parallel formative assessment was then administered. Teachers in the comparison group delivered business-as-usual instruction.	Other directed
Ross, Rolheiser, & Hoaboam-Gray (1998)	Self-assessment. Intervention teachers received a handbook and participated in in-service trainings that provided guidance on how to teach students to self-evaluate in math. Teachers in the comparison group delivered business-as-usual instruction.	Student directed
Sawyer, Graham, & Harris (1992)	Self-regulation. In this formative assessment intervention, students were taught self-regulation, including goal setting and self-monitoring. Both groups received writing instruction, using a multicomponent strategy instructional model to teach the number and kind of story grammar elements to include in a story. The intervention group also received the self-regulation intervention; it involved learning how to set a goal for the number of story elements to include in a story, examine the completed story and assess whether all elements were included, and graph the number of elements included.	Student directed
Wesson (1990)	Curriculum-based measurement. The formative assessment intervention was curriculum-based measurement. Teachers were trained to write oral reading goals for individual students and monitor individual progress using curriculum-based measurement. These curriculum-based measurement procedures involved assessing individual student oral reading using number of words read correctly per minute three times per week, graphing scores, and analyzing the data to make instructional decisions based on the amount of weekly increase in performance. Teachers in the comparison group received four hours of training on how to specify instructional plans. During the training, participants discussed ideas for monitoring progress. Teachers in this group wrote goals and instructional plans for their students and developed their own approaches to monitoring progress.	Other directed
Ysseldyke & Tardrew (2007)	Computer-facilitated assessment. Students took the computer adaptive STAR Math test. Based on the results of the test, teachers assigned students to appropriate "libraries" of material, based on their instructional objectives for each student. The computer generated worksheets containing random sets of problems that met the objectives for each student. These were completed on paper, and students recorded their answers on a scan sheet. When the sheet was scanned, the information regarding the student's score was added to the instructional management system. Teachers could receive daily reports about how students were doing relative to their customized instructional goals. The daily report flagged students who appeared to be experiencing the most difficulty so that teachers could design other interventions for these students. Teachers in the comparison group provided business-as-usual instruction.	Other directed

Source: Authors' analysis of studies published between 1988 and 2014; see appendix A for details.

Table B2. Student-directed formative assessment interventions, effect sizes, and information about samples in studies that met standards modeled after those used by the What Works Clearinghouse with or without reservations, by outcome domain

Study citation	Intervention name (group size)	Comparison condition (group size)	Study design	Outcome subdomain	Effect size	Characterization of findings	Student sample type	Student sample grade level	Study rating
Math									
Bond & Ellis (2013)	Metacognitive reflective assessment (47)	General review of lesson (48)	Randomized controlled trial	Data, probability, and/or statistics	0.49	Substantively important positive	Primarily general education students	5–6	Meets without reservations
Cripps (1995)	Flash-card assessment and progress monitoring with peers and curriculum-based measurement (27)	Flash-card assessment and progress monitoring with peers without curriculum-based measurement (26)	Randomized controlled trial	Math operations	0.46	Substantively important positive	Primarily general education students	3–4	Meets without reservations
Menesses & Gresham (2009)	Flash-card assessment and progress monitoring with peers (15)	Conventional classroom instruction (16)	Randomized controlled trial	Math operations	1.01	Substantively important positive	Primarily general education students	2–4	Meets without reservations
Ross, Rolheiser, & Hoaboam-Gray (1998)	Self-assessment (164)	No self-assessment (142)	Randomized controlled trial	Data, probability and/or statistics	–0.18	Indeterminate	Primarily general education students	5	Meets with reservations
Reading									
Johnson, Graham, & Harris (1997)	Reading strategy instruction plus goal setting/ self-monitoring plus self-instruction (13)	Self-instruction (9)	Randomized controlled trial	General reading achievement and reading comprehension	–0.16	Indeterminate	Students in special education	4–6	Meets without reservations
Johnson, Graham, & Harris (1997)	Reading strategy instruction plus goal setting/ self-monitoring (9)	Strategy instruction (11)	Randomized controlled trial	General reading achievement and reading comprehension	–0.46	Substantively important negative	Students in special education	4–6	Meets without reservations

(continued)

Table B2. Student-directed formative assessment interventions, effect sizes, and information about samples in studies that met standards modeled after those used by the What Works Clearinghouse with or without reservations, by outcome domain *(continued)*

Study citation	Intervention name (group size)	Comparison condition (group size)	Study design	Outcome subdomain	Effect size	Characterization of findings	Student sample type	Student sample grade level	Study rating
McCurdy & Shapiro (1992)	Curriculum-based measurement (peer monitoring) (10)	No curriculum-based measurement (11)	Randomized controlled trial	Oral reading fluency	–0.15	Indeterminate	Students in special education	2–5	Meets without reservations
McCurdy & Shapiro (1992)	Curriculum-based measurement (self-monitoring) (12)	No curriculum-based measurement (11)	Randomized controlled trial	Oral reading fluency	0.17	Indeterminate	Students low-achieving in a particular subject (math, reading, or spelling)	2–5	Meets without reservations
Writing									
Sawyer, Graham, & Harris (1992)	Self-regulation (11)	No self-regulation (10)	Randomized controlled trial	Narrative composition	0.63	Substantively important positive	Students in special education	5	Meets without reservations

Source: Authors' analysis of studies published between 1988 and 2014; see appendix A for details.

Table B3. Other-directed formative assessment interventions, effect sizes, and information about outcome measures and samples in studies that researchers determined met standards modeled after those used by the What Works Clearinghouse with or without standards, by outcome domain

Study citation	Intervention name (group size)	Comparison condition (group size)	Study design	Outcome subdomain	Effect size	Characterization of findings	Student sample type	Student sample grade level	Study rating
Math									
Craven, Marsh, & Debus (1991)	Attributional feedback: math (54)	No feedback (106)	Randomized controlled trial	General math achievement	—	—	Primarily general education students	3–6	Meets without reservations
Fuchs, Fuchs, Hamlett, & Stecker (1991)	Curriculum-based measurement without expert consultation (11)	No curriculum-based measurement (11)	Randomized controlled trial	Math operations	0.07	Indeterminate	Students in special education	2–8	Meets without reservations
Fuchs, Fuchs, Hamlett, & Stecker (1991)	Curriculum-based measurement with expert consultation (11)	No curriculum-based measurement (11)	Randomized controlled trial	Math operations	0.66	Substantively important positive	Students in special education	2–8	Meets without reservations
Ysseldyke & Tardrew (2007)	Computer-facilitated assessment (Accelerated Math) (231)	No use of computer-facilitated assessment (245)	Quasi-experimental design	General math achievement	0.33	Substantively important positive	Primarily general education students	3	Meets with reservations
Ysseldyke & Tardrew (2007)	Computer-facilitated assessment (Accelerated Math) (303)	No use of computer-facilitated assessment (311)	Quasi-experimental design	General math achievement	0.25	Substantively important positive	Primarily general education students	4	Meets with reservations
Ysseldyke & Tardrew (2007)	Computer-facilitated assessment (Accelerated Math) (335)	No use of computer-facilitated assessment (255)	Quasi-experimental design	General math achievement	0.36	Substantively important positive	Primarily general education students	5	Meets with reservations
Ysseldyke & Tardrew (2007)	Computer-facilitated assessment (Accelerated Math) (169)	No use of computer-facilitated assessment (157)	Quasi-experimental design	General math achievement	0.14	Indeterminate	Primarily general education students	6	Meets with reservations

(continued)

Table B3. Other-directed formative assessment interventions, effect sizes, and information about outcome measures and samples in studies that researchers determined met standards modeled after those used by the What Works Clearinghouse with or without standards, by outcome domain *(continued)*

Study citation	Intervention name (group size)	Comparison condition (group size)	Study design	Outcome subdomain	Effect size	Characterization of findings	Student sample type	Student sample grade level	Study rating
Reading									
Craven, Marsh, & Debus (1991)	Attributional feedback: reading (54)	No feedback (106)	Randomized controlled trial	General reading achievement and reading comprehension	—	—	Primarily general education students	3–6	Meets without reservations
Fuchs, Fuchs, & Hamlett (1989)	Curriculum-based measurement plus evaluation (21)	No curriculum-based measurement (18)	Quasi-experimental design	General reading achievement and reading comprehension	1.22	Statistically significant positive	Students in special education	2–9	Meets with reservations
Fuchs, Fuchs, Hamlett, & Ferguson (1992)	Curriculum-based measurement without expert consultation (11)	No curriculum-based measurement (11)	Randomized controlled trial	General reading achievement and reading comprehension	0.45	Substantively important positive	Students in special education	2–9	Meets without reservations
Fuchs, Fuchs, Hamlett, & Ferguson (1992)	Curriculum-based measurement with expert consultation (11)	No curriculum-based measurement (11)	Randomized controlled trial	General reading achievement and reading comprehension	0.79	Substantively important positive	Students in special education	2–9	Meets without reservations
Iannuccilli (2003)	Feedback to teachers (26)	No feedback (28)	Randomized controlled trial	Phonemic segmentation and decoding	0.06	Indeterminate	Primarily general education students	1	Meets without reservations
Martens, Eckert, & Begeny (2007)	Curriculum-based measurement (5)	No curriculum-based measurement (5)	Randomized controlled trial	Oral reading fluency	0.19	Indeterminate	Students in special education	2	Meets without reservations
Martens, Eckert, & Begeny (2007)	Curriculum-based measurement (10)	No curriculum-based measurement (10)	Randomized controlled trial	Oral reading fluency	0.15	Indeterminate	Students in special education	3	Meets without reservations

(continued)

Table B3. Other-directed formative assessment interventions, effect sizes, and information about outcome measures and samples in studies that researchers determined met standards modeled after those used by the What Works Clearinghouse with or without standards, by outcome domain *(continued)*

Study citation	Intervention name (group size)	Comparison condition (group size)	Study design	Outcome subdomain	Effect size	Characterization of findings	Student sample type	Student sample grade level	Study rating
McCurdy & Shapiro (1992)	Curriculum-based measurement (teacher monitoring) (10)	No curriculum-based measurement (11)	Randomized controlled trial	Oral reading fluency	0.02	Indeterminate	Students in special education	2–5	Meets without reservations
Mostow et al. (2003)	Human-facilitated assessment (34)	No facilitated assessment (39)	Quasi-experimental design	General reading achievement and reading comprehension	—	—	Students low-achieving in a particular subject (math, reading, or spelling)	1–3	Meets with reservations
Mostow et al. (2003)	Computer-facilitated assessment (58)	No facilitated assessment (39)	Quasi-experimental design	General reading achievement and reading comprehension	—	—	Students low-achieving in a particular subject (math, reading, or spelling)	1–3	Meets with reservations
Null (1990)	Mastery learning (80)	Conventional classroom instruction (88)	Randomized controlled trial	General reading achievement and reading comprehension	0.38	Substantively important positive	Primarily general education students	2–3	Meets without reservations
Writing									
Abrami, Venkatesh, Meyer, & Wade (2013)	Self-regulation (ePEARL) (154)	No self-regulation (165)	Quasi-experimental design	Writing in response to reading	0.34	Substantively important positive	Primarily general education students	4–6	Meets with reservations
Fuchs, Butterworth, & Fuchs (1989)	Curriculum-based measurement (computer-assisted) (9)	No curriculum-based measurement (10)	Randomized controlled trial	Spelling	—	—	Students in special education	4–8	Meets without reservations
Fuchs, Butterworth, & Fuchs (1989)	Curriculum-based measurement (paper/pencil) (10)	No curriculum-based measurement (10)	Randomized controlled trial	Spelling	—	—	Students in special education	4–8	Meets without reservations

(continued)

Table B3. Other-directed formative assessment interventions, effect sizes, and information about outcome measures and samples in studies that researchers determined met standards modeled after those used by the What Works Clearinghouse with or without standards, by outcome domain *(continued)*

Study citation	Intervention name (group size)	Comparison condition (group size)	Study design	Outcome subdomain	Effect size	Characterization of findings	Student sample type	Student sample grade level	Study rating
Fuchs, Fuchs, Hamlett, & Allinder (1991a)	Curriculum-based measurement without expert consultation (20)	No curriculum-based measurement (20)	Randomized controlled trial	Spelling	0.30	Substantively important positive	Students in special education	2–8	Meets without reservations
Fuchs, Fuchs, Hamlett, & Allinder (1991a)	Curriculum-based measurement with expert consultation (20)	No curriculum-based measurement (20)	Randomized controlled trial	Spelling	0.23	Indeterminate	Students in special education	2–8	Meets without reservations
Fuchs, Fuchs, Hamlett, & Allinder (1991b)	Curriculum-based measurement without skills analysis (35)	No curriculum-based measurement (18)	Quasi-experimental design	Spelling	0.09	Indeterminate	Students in special education	2–8	Meets with reservations
Fuchs, Fuchs, Hamlett, & Allinder (1991b)	Curriculum-based measurement with skills analysis (39)	No curriculum-based measurement (18)	Quasi-experimental design	Spelling	0.15	Indeterminate	Students in special education	2–8	Meets with reservations
Graham & Harris (1989)	Self-regulation (11)	No self-regulation (11)	Quasi-experimental design	Narrative composition	–0.20	Indeterminate	Students in special education	5–6	Meets without reservations
Meyer, Abrami, Wade, Aslan, & Deault (2010)	Self-regulation (ePEARL) (121)	No self-regulation (175)	Quasi-experimental design	Writing in response to reading	0.11	Indeterminate	Primarily general education students	4–6	Meets with reservations

— Information not available to compute effect size or characterize findings.

Source: Authors' analysis of studies published between 1988 and 2014; see appendix A for details.

Appendix C. Findings from studies that compared two different types of formative assessment

Results for studies that compared two different types of formative assessment are presented in table C1. In math there was only one study comparing two interventions (Fuchs, Fuchs, Hamlett, & Stecker, 1991). In that study curriculum-based measurement was more effective when teachers received computer-based expert consultation, which suggested instructional changes based on student assessment data, than when teachers were required to identify instructional changes on their own.

In reading there were six comparisons from four studies that involved two types of formative assessment. Of these, effect sizes for two studies were large enough to be considered substantively important. Similar to the math study described above, one of the two studies compared curriculum-based measurement with expert consultation to curriculum-based measurement alone (Fuchs, Fuchs, Hamlett, & Ferguson, 1992). As with the math study, curriculum-based measurement was more effective when teachers received computer-generated suggestions for instructional changes that were informed by the student assessment data than when teachers were required to determine instructional changes on their own. The second study found that curriculum-based measurement was more effective than teacher-designed goal setting and monitoring (Wesson, 1990). For one study the information to compute an effect size was not available (Mostow et al., 2003). For the remaining three comparisons, the effects were small, indicating that there was not enough evidence to conclude that one type of formative assessment was more effective than another.

In writing there were three comparisons from three studies that involved two types of formative assessment. For one study the information required to compute an effect size was not available (Fuchs, Butterworth, & Fuchs, 1989). For the other two the effect sizes were small, indicating that there was not enough evidence to conclude that one type of formative assessment was more effective than the other.

Table C1. Studies that compared two types of formative assessment, effect sizes, and information about outcome measures and samples in studies that met standards modeled after those used by the What Works Clearinghouse with or without reservations

Study citation	Intervention name (group size)	Comparison condition (group size)	Study design	Outcome subdomain	Effect size	Characterization of findings	Student sample type	Student sample grade level	Study rating
Math									
Fuchs, Fuchs, Hamlett, & Stecker (1991)	Curriculum-based measurement with expert consultation (11)	Curriculum-based measurement without expert consultation (11)	Randomized controlled trial	Math operations	0.72	Substantively important positive	Students in special education	2–8	Meets without reservations
Reading									
Fuchs, Fuchs, Hamlett, & Ferguson (1992)	Curriculum-based measurement with expert consultation (11)	Curriculum-based measurement without expert consultation (11)	Randomized controlled trial	General reading achievement and reading comprehension	0.44	Substantively important positive	Students in special education	2–9	Meets without reservations
McCurdy & Shapiro (1992)	Curriculum-based measurement (peer monitoring) (10)	Curriculum-based measurement (self-monitoring) (12)	Randomized controlled trial	Oral reading fluency	–0.22	Indeterminate	Students in special education	2–5	Meets without reservations
McCurdy & Shapiro (1992)	Curriculum-based measurement (teacher monitoring) (10)	Curriculum-based measurement (peer monitoring) (10)	Randomized controlled trial	Oral reading fluency	0.18	Indeterminate	Students in special education	2–5	Meets without reservations
McCurdy & Shapiro (1992)	Curriculum-based measurement (teacher monitoring) (10)	Curriculum-based measurement (self-monitoring) (12)	Randomized controlled trial	Oral reading fluency	–0.06	Indeterminate	Students in special education	2–5	Meets without reservations
Mostow et al. (2003)	Computer-facilitated assessment (58)	Human-facilitated assessment (39)	Quasi-experimental design	General reading achievement and reading comprehension	—	—	Students low-achieving in a particular subject (math, reading, or spelling)	1–3	Meets with reservations
Wesson (1990)	Curriculum-based measurement (29)	Teacher-designed goal setting and monitoring (26)	Randomized controlled trial	General reading achievement and reading comprehension	0.29	Substantively important positive	Students in special education	2–7	Meets without reservations

(continued)

Table C1. Studies that compared two types of formative assessment, effect sizes, and information about outcome measures and samples in studies that met standards modeled after those used by the What Works Clearinghouse with or without reservations *(continued)*

Study citation	Intervention name (group size)	Comparison condition (group size)	Study design	Outcome subdomain	Effect size	Characterization of findings	Student sample type	Student sample grade level	Study rating
Writing									
Fuchs, Butterworth, & Fuchs (1989)	Curriculum-based measurement (computer-assisted) (9)	Curriculum-based measurement (paper/pencil) (10)	Randomized controlled trial	Spelling	—	—	Students in special education	4–8	Meets without reservations
Fuchs, Fuchs, Hamlett, & Allinder (1991a)	Curriculum-based measurement with expert consultation (20)	Curriculum-based measurement without expert consultation (20)	Randomized controlled trial	Spelling	–0.09	Indeterminate	Students in special education	2–8	Meets without reservations
Fuchs, Fuchs, Hamlett, & Allinder (1991b)	Curriculum-based measurement with skills analysis (39)	Curriculum-based measurement without skills analysis (35)	Quasi-experimental design	Spelling	0.06	Indeterminate	Students in special education	2–8	Meets with reservations

— Information not available to compute effect size or characterize findings.

Source: Authors' analysis of studies published between 1988 and 2014; see appendix A for details.

Appendix D. Studies rated "does not meet standards"

This appendix lists the studies that were rated "does not meet standards." For each reference a brief description of the reason or reasons why the study did not meet standards is provided. These descriptions are modeled after the descriptions used by the What Works Clearinghouse (WWC).

Table D1. Studies that did not meet standards modeled after those used by the What Works Clearinghouse

Reference	Reason the study did not meet standards
Alitto, J. M. (2009). *The effects of peer-mediated goal setting and performance feedback on curriculum-based measurement indices of written expression.* Unpublished doctoral dissertation, Northern Illinois University, DeKalb.	It is a randomized controlled trial that does not provide information needed to determine the amount of attrition. In addition, the intervention and comparison groups are not shown to be equivalent at baseline.
Allinder, R. M., Bolling, R. M., Oats, R. G., & Gagnon, W. A. (2000). Effects of teacher self-monitoring on implementation of curriculum-based measurement and mathematics computation achievement of students with disabilities. *Remedial and Special Education, 21*(4), 219–226.	The intervention and comparison groups are not shown to be equivalent at baseline.
Baker, K. R. (2005). *The effects of self-monitoring with accuracy feedback versus self-monitoring with corrective feedback on students' performance in mathematics.* Unpublished master's thesis, Eastern Illinois University, Charleston.	It does not provide adequate information to determine whether it uses an outcome that is valid or reliable.
Baxter, J. B. (1996). *Providing technology-based formative evaluation support for mathematics programs using a subjective-probability assessment format to improve at-risk student attainment.* Unpublished doctoral dissertation, University of California, Los Angeles.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Beal, C. R., Arroyo, I. M., Cohen, P. R., & Woolf, B. P. (2010). Evaluation of AnimalWatch: An intelligent tutoring system for arithmetic and fractions. *Journal of Interactive Online Learning,* 9(1), 64–77.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions. In addition, the intervention and comparison groups are not shown to be equivalent at baseline.
Bohlin, S. L. (2000). *Effectiveness of instruction in rubric use in improving fourth-grade students' science open-response outcomes.* Unpublished doctoral dissertation, University of Massachusetts, Lowell.	The intervention and comparison groups are not shown to be equivalent at baseline.
Boys, C. J. (2003). *Mastery orientation through task-focused goals: Effects on achievement and motivation.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.	It is a randomized controlled trial that does not provide information needed to determine the amount of attrition. In addition, the intervention and comparison groups are not shown to be equivalent at baseline.
Brace, D. L. (1992). *A study of group-based mastery learning strategies.* Unpublished doctoral dissertation, Northern Illinois University, DeKalb.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Burdon, P. C., Flowers, J. D., & Manchak, S. C. (2011). *Impact of students' self-assessment and creation of personal learning targets on reading comprehension and attitudes in elementary schools*. http://eric.ed.gov/?id=ED529623	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Cabral-Marquez, C. (2012). *The effects of setting reading goals on reading motivation, reading achievement, and reading activity.* Unpublished doctoral dissertation, Northern Illinois University, DeKalb.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.

(continued)

Table D1. Studies that did not meet standards modeled after those used by the What Works Clearinghouse *(continued)*

Reference	Reason the study did not meet standards
Caputo, M. T. (2007). *A comparison of the effects of the Accelerated Math program and the Delaware procedural fluency workbook program on academic growth in grade six at X middle school.* Unpublished doctoral dissertation, Wilmington University, New Castle, DE.	The intervention and comparison groups are not shown to be equivalent at baseline.
Conte, K., & Hintze, J. M. (2000). The effects of performance feedback and goal setting on oral reading fluency within curriculum-based measurement. *Assessment for Effective Intervention, 25*(2), 85–98.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Delacruz, G. C. (2010). *Games as formative assessment environments: Examining the impact of explanations of scoring and incentives on math learning, game performance, and help seeking.* Unpublished doctoral dissertation, University of California, Los Angeles. Additional sources: Delacruz, G. C. (2011). *Games as formative assessment environments: Examining the impact of explanations of scoring and incentives on math learning, game performance, and help seeking* (CRESST Report 796). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Delacruz, G. C. (2012). *Impact of incentives on the use of feedback in educational video games* (CRESST Report 813). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).	The intervention and comparison groups are not shown to be equivalent at baseline.
Easterwood, C. A. (1996). *The effect of self-reflective portfolios on the writing achievement of third grade students.* Unpublished master's thesis, Central Missouri State University, Warrensburg.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions. In addition, the intervention and comparison groups are not shown to be equivalent at baseline.
Forbush, D. (2001). *Math Renaissance improves student achievement and attitudes in Idaho school* (Renaissance Independent Research Report 35). Madison, WI: Renaissance Learning.	The intervention and comparison groups are not shown to be equivalent at baseline.
Franco-Castillo, I. (2013). *The relationship between scaffolding metacognitive strategies identified through dialogue journals and second graders' reading comprehension, science achievement, and metacognition using expository text.* Unpublished doctoral dissertation, Florida International University, Miami.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Fuchs, L., Fuchs, D., Hamlett, C. L., & Stecker, P. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review, 19*, 6–22.	The intervention and comparison groups are not shown to be equivalent at baseline.
Furey, J. (2012). *Effects of curriculum-based measurement feedback to students on reading fluency, self-efficacy, and motivation.* Unpublished master's thesis, University of Rhode Island, Kingston.	It is a randomized controlled trial that does not provide information needed to determine the amount of attrition. In addition, the intervention and comparison groups are not shown to be equivalent at baseline.
Gaetz, T. M. (1992). *The effects of a self-monitoring checklist on elementary students' postreading question-answering performance.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.

(continued)

Table D1. Studies that did not meet standards modeled after those used by the What Works Clearinghouse *(continued)*

Reference	Reason the study did not meet standards
Gaskill, P. J. (2003). *Effects of a goal-setting strategy on second graders' self-efficacy for a listening task.* Unpublished doctoral dissertation, Ohio State University, Columbus.	It does not use an outcome that is valid or reliable.
Havens, P. F. (2000). *The relationships between the reactivity of self-monitoring and the hierarchical facets of self-concept.* Unpublished doctoral dissertation, Arizona State University, Tempe.	The intervention and comparison groups are not shown to be equivalent at baseline.
Hecht-Lewis, R. A. (1990). *The impact of locus-of-control upon self-control training.* Unpublished doctoral dissertation, George Washington University, Washington, DC.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Helseth, M. S. (2013). *What is the effect of teacher feedback in student notebooks on student achievement?* Unpublished master's thesis, Montana State University, Bozeman.	It does not provide adequate information to determine whether it uses an outcome that is valid or reliable.
Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning, 4*(2), 161–176.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Johnson, L. I. (2004). *The effects of reflective assessment on intermediate-grade student achievement in mathematics.* Unpublished doctoral dissertation, Seattle Pacific University, Seattle.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Johnson-Scott, P. L. (2006). *The impact of Accelerated Math on student achievement.* Unpublished doctoral dissertation, Mississippi State University, Starkville.	The estimates of effects did not account for differences in pre-intervention characteristics when using a quasi-experimental design.
Kariuki, P., & Wiseman, B. (2006). *The effects of self-assessment on kindergarten students learning of high frequency words*. http://eric.ed.gov/?id=ED495491	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
King, M. D. (2003). *The effects of formative assessment on student self-regulation, motivational beliefs, and achievement in elementary science.* Unpublished doctoral dissertation, George Mason University, Fairfax, VA.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions. In addition, the intervention and comparison groups are not shown to be equivalent at baseline.
Klentschy, M. P. (1992). *Designing instructional support and decision-making systems to service accelerated learning environments in large urban elementary schools.* Unpublished doctoral dissertation, University of California, Los Angeles.	It is a randomized controlled trial that does not provide information needed to determine the amount of attrition. In addition, the intervention and comparison groups are not shown to be equivalent at baseline.
Kline, F. M., Schumaker, J. B., & Deshler, D. D. (1991). Development and validation of feedback routines for instructing students with learning disabilities. *Learning Disability Quarterly, 14*(3), 191–207.	The intervention and comparison groups are not shown to be equivalent at baseline.
Knutson, J. S. (2005). *The effect of corrective feedback and individualized practice guided by formative evaluation on the reading performance of children who have not made adequate progress in early reading instruction.* Unpublished doctoral dissertation, University of Oregon, Eugene.	The intervention and comparison groups are not shown to be equivalent at baseline.
Krzmarzick, L. A. (1994). *An investigation of videotaped feedback and its effect on elementary students' public speaking effectiveness.* Unpublished master's thesis, St. Cloud University, St. Cloud, MN.	The intervention and comparison groups are not shown to be equivalent at baseline.
Leung, L. K. (1991). *The effects of goal setting on the academic achievement, motivation, and confidence of bright underachievers.* Unpublished master's thesis, University of Victoria, Victoria, British Columbia, Canada.	It does not use an outcome that is valid or reliable.

(continued)

Table D1. Studies that did not meet standards modeled after those used by the What Works Clearinghouse *(continued)*

Reference	Reason the study did not meet standards
Long, V. M. (1992). *Effects of mastery learning on mathematics achievement and attitudes.* Unpublished doctoral dissertation, University of Missouri, Columbia.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Malone, L. D., & Mastropieri, M. A. (1992). Reading comprehension instruction: Summarization and self-monitoring training for students with learning disabilities. *Exceptional Children, 58*(3), 270–279.	The intervention and comparison groups are not shown to be equivalent at baseline.
Niedo, J., Lee, Y.-L., Breznitz, Z., & Berninger, V. W. (2014). Computerized silent reading rate and strategy instruction for fourth graders at risk in silent reading rate. *Learning Disability Quarterly, 37*(2), 100–110.	It is a randomized controlled trial in which the combination of overall and differential attrition exceeds What Works Clearinghouse standards for this area, and subsequent analytic intervention and comparison groups are not shown to be equivalent.
Peltier, V. E., & Ross, J. S. (1995). *Improving student achievement and attitudes in elementary mathematics through written error-correcting feedback on tests.* Unpublished educational specialist thesis, University of Dayton, Dayton, OH.	It is a randomized controlled trial that does not provide information needed to determine the amount of attrition. In addition, the intervention and comparison groups are not shown to be equivalent at baseline.
Price, S. C. (1993). *The effects of group-based mastery learning on first-grade reading achievement.* Unpublished doctoral dissertation, Miami University, Oxford, OH.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Quinn, G. P. (1996). *Using goal setting and self-regulation to enhance reading achievement and study time.* Unpublished doctoral dissertation, Florida State University, Tallahassee.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Ritchie, D., & Thorkildsen, R. (1994). Effects of accountability on students' achievement in mastery learning. *Journal of Educational Research, 88*(2), 86–90.	It is a randomized controlled trial that does not provide information needed to determine the amount of attrition. In addition, the intervention and comparison groups are not shown to be equivalent at baseline.
Rosenthal, B. D. (2006). *Improving elementary-age children's writing fluency: A comparison of improvement based on performance feedback frequency.* Unpublished doctoral dissertation, Syracuse University, Syracuse, NY.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Ross, J. A., Hoaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5–6 mathematics effects on problem-solving achievement. *Educational Assessment, 8*(1), 43–58.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions. In addition, the intervention and comparison groups are not shown to be equivalent at baseline.
Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing Writing, 6*(1), 107–132.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions. In addition, the intervention and comparison groups are not shown to be equivalent at baseline.
Rudd, P., & Wade, P. (2006). *Evaluation of Renaissance Learning mathematics and reading programs in UK specialist and feeder schools*. Slough, UK: National Foundation for Educational Research.	The intervention and comparison groups are not shown to be equivalent at baseline.
Spicuzza, R., Ysseldyke, J., Lemkuil, A., Kosciolek, S., Boys, C., & Teelucksingh, E. (2001). Effects of curriculum-based monitoring on classroom instruction and math achievement. *Journal of School Psychology, 39*(6), 521–542.	The intervention and comparison groups are not shown to be equivalent at baseline.
Stecker, P. M. (1993). *Effects of instructional modifications with and without curriculum-based measurement on the mathematics achievement of students with mild disabilities.* Unpublished doctoral dissertation, Vanderbilt University, Nashville, TN.	The intervention and comparison groups are not shown to be equivalent at baseline.

(continued)

Table D1. Studies that did not meet standards modeled after those used by the What Works Clearinghouse *(continued)*

Reference	Reason the study did not meet standards
Tieso, C. (2005). The effects of grouping practices and curricular adjustments on achievement. *Journal for the Education of the Gifted, 29*(1), 60–89.	It is a randomized controlled trial that does not provide information needed to determine the amount of attrition. In addition, the intervention and comparison groups are not shown to be equivalent at baseline.
Vollands, S. R. (1996). *Experimental evaluation of computer assisted self-assessment of reading comprehension: Effects on reading achievement and attitude*. http://eric.ed.gov/?id=ED408567	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Whalen, A. J. (2002). *The effect of direct teacher involvement in formative evaluation of student progress on student attainment of critical early literacy outcomes.* Unpublished doctoral dissertation, University of Oregon, Eugene.	The intervention and comparison groups are not shown to be equivalent at baseline.
Wick, J. B. (2006). *Enhancing young readers' oral reading fluency and metacognitive sophistication: Evaluating the effectiveness of a computer mediated self-monitoring literacy tool.* Doctoral dissertation, University of Texas, Austin.	The measures of effectiveness cannot be attributed solely to the intervention—only one unit was assigned to one or both conditions.
Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Brandon, P. R., Furtak, E. M., et al. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education, 21*(4), 335–359.	The intervention and comparison groups are not shown to be equivalent at baseline.
Ysseldyke, J., & Bolt, D. M. (2007). Effect of technology-enhanced continuous progress monitoring on math achievement. *School Psychology Review, 36*(3), 453–467.	It is a randomized controlled trial that does not provide information needed to determine the amount of attrition. In addition, the intervention and comparison groups are not shown to be equivalent at baseline.[a]
Ysseldyke, J. E., Spicuzza, R., & McGill (2000). *Changes in mathematics achievement and instructional ecology resulting from implementation of a learning information system*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.	The intervention and comparison groups are not shown to be equivalent at baseline.

a. This study was also included in two reports on Accelerated Math published by U.S. Department of Education (2008, 2010). In those reviews, the What Works Clearinghouse was able to obtain all necessary information from the study authors to determine the amount of attrition. Readers are encouraged to consult those resources for additional information about this study.

Source: Authors' analysis of studies published between 1998 and 2014; see appendix A for details.

Notes

1. Curriculum-based measurement is a system of monitoring individual student progress toward annual goals using frequent brief assessments. Results are graphed over time as number of digits correct in computations, number of words read correctly, or number of letters correct in spelled words. When student performance remains flat, teachers make changes in the instruction and continue monitoring to determine whether the revisions are effective.
2. The approach to this review is modeled after the approach the WWC uses, with a few modifications, as explained in appendix A. WWC-certified reviewers examined all studies. However, this review was not conducted by the WWC and should not be considered an official WWC publication.
3. Nearly all studies of students in special education classes included students from multiple grades. Some of these studies included students through grade 9. When study results were not reported separately by grade, this review included these studies and their results. Specific grade levels of students who were part of each study's sample are reported in appendix B.
4. For more information on the attrition standards used by the WWC, see U.S. Department of Education (n.d.).
5. The Study Review Guide was developed by the U.S. Department of Education, Institute of Education Sciences through its WWC project and was used by the review team with permission (U.S. Department of Education, 2014c). All the reviewers were trained and certified by the WWC to conduct reviews. However, this review was not conducted by the WWC and should not be considered an official WWC publication.
6. According to the *What Works Clearinghouse Procedures and Standards Handbook:* "To be a separate study, the sampling errors must be independent. For randomized controlled trials, a study is defined by randomization. This definition excludes subgroups from being their own studies because they were randomized at the same time as the full sample. … For quasi-experimental designs, studies are separate only if they use independent samples" (U.S. Department of Education, 2014b, p. 7).

References

The 23 studies that met standards and contributed findings to this review are marked with an asterisk.

*Abrami, P. C., Venkatesh, V., Meyer, E. J., & Wade, C. A. (2013). Using electronic portfolios to foster literacy and self-regulated learning skills in elementary students. *Journal of Educational Psychology*, *105*(4), 1188–1209. http://eric.ed.gov/?id=EJ1054427

Andrade, H. L., & Cizek, G. J. (2010). *Handbook of formative assessment*. New York, NY: Routledge.

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213–238.

Baumert, J., Nagy, G., & Lehmann, R. (2012). Cumulative advantages and the emergence of social and ethnic inequality: Matthew effects in reading and mathematics development within elementary schools. *Child Development*, *83*(4), 1347–1367. http://eric.ed.gov/?id=EJ991707

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. http://eric.ed.gov/?id=EJ912798

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

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139–148. http://eric.ed.gov/?id=EJ575146

Black, P., & Wiliam, D. (2003). 'In praise of educational research': Formative assessment. *British Educational Research Journal*, *29*(5), 623–637. http://eric.ed.gov/?id=EJ770636

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment Evaluation and Accountability*, *21*(1), 5–31. http://eric.ed.gov/?id=EJ829749

Boekaerts, M., Pintrich, P. R., & Zeidner, M. (2000). *Handbook of self-regulation*. San Diego, CA: Academic Press.

*Bond, J. B., & Ellis, A. K. (2013). The effects of metacognitive reflective assessment on fifth and sixth graders' mathematics achievement. *School Science and Mathematics*, *113*(5), 227–234. http://eric.ed.gov/?id=EJ1011105

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., & Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice*, *31*(4), 13–17. http://eric.ed.gov/?id=EJ988994

Brookhart, S. M. (2008). *How to give effective feedback to your students*. Alexandria, VA: ASCD. http://eric.ed.gov/?id=ED509138

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

Brookhart, S. M. (2014). *The essence of formative assessment: Creating a common understanding of the formative assessment process (Part one of a three-part series)* (Archived webinar). Washington, DC: U.S. Department of Education. Retrieved July 11, 2015, from https://www.relcentral.org/news-and-events/the-essence-of-formative-assessment-creating-a-common-understanding-of-the-formative-assessment-process/.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–281.

Carreker, S., Neuhaus, G., Swank, P., Johnson, R., Monfils, M. J., & Montemayor, M. (2007). Teachers with linguistically informed knowledge of reading subskills are associated with a Matthew effect in reading comprehension for monolingual and bilingual students. *Reading Psychology*, *28*(2), 187–212. http://eric.ed.gov/?id=EJ763814

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, *24*(2), 205–249. http://eric.ed.gov/?id=EJ964039

*Craven, R. G, Marsh, H. W., & Debus, R. L. (1991). Effects of internally focused feedback and attributional feedback on enhancement of academic self-concept. *Journal of Educational Psychology*, *83*(1), 17–27. http://eric.ed.gov/?id=EJ436862

*Cripps, K. H. (1995). *The effects of peer monitoring with feedback on progress toward established goals in cross-age peer tutoring.* Unpublished doctoral dissertation, University of Oregon, Eugene.

Dynarski, M., & Kisker, E. (2014). *Going public: Writing about research in everyday language* (REL 2014–051). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved June 7, 2016, from http://ies.ed.gov/ncee/edlabs.

Filsecker, M., & Kerres, M. (2012). Positioning formative assessment from an educational assessment perspective: A response to Dunn & Mulvenon (2009). *Practical Assessment, Research & Evaluation*, *17*(16), 1–9. http://eric.ed.gov/?id=EJ990690

*Fuchs, L. S., Butterworth, J. R., & Fuchs, D. (1989). Effects of ongoing curriculum-based measurement on student awareness of goals and progress. *Education and Treatment of Children*, *12*(1), 63–72.

*Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Effects of instrumental use of curriculum-based measurement to enhance instructional programs. *Remedial and Special Education*, *10*(2), 43–52. http://eric.ed.gov/?id=EJ390644

*Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Allinder, R. M. (1991a). Effects of expert system advice within curriculum-based measurement on teacher planning and student achievement in spelling. *School Psychology Review*, *20*(1), 49–66.

*Fuchs, D., Fuchs, L. S., & Hamlett, C. L., & Allinder, R. M. (1991b). The contribution of skills analysis to curriculum-based measurement in spelling. *Exceptional Children*, *57*(5), 443–452.

*Fuchs, D., Fuchs, L. S., Hamlett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement, using a reading maze task. *Exceptional Children*, *58*(5), 436–450. http://eric.ed.gov/?id=EJ444443

*Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal*, *28*(3), 617–641. http://eric.ed.gov/?id=EJ433857

*Graham, S., & Harris, K. R. (1989). Components analysis of cognitive strategy instruction: Effects on learning disabled students' compositions and self-efficacy. *Journal of Educational Psychology*, *81*(3), 353–361.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Heritage, M. (2010a, September). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* Paper prepared for the Council of Chief State School Officers, Washington, DC. http://eric.ed.gov/?id=ED543063

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

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

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.

*Iannuccilli, J. A. (2003). Monitoring the progress of first-grade students with Dynamic Indicators of Basic Early Literacy Skills. *Dissertation Abstracts International: Humanities and Social Sciences*, *64*(8-A), 2824.

*Johnson, L., Graham, S., & Harris, K. R. (1997). The effects of goal setting and self-instruction on learning a reading comprehension strategy: A study of students with learning disabilities. *Journal of Learning Disabilities*, *30*(1), 80–91. http://eric.ed.gov/?id=EJ542685

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. http://eric.ed.gov/?id=EJ951173

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*, *31*(4), 18–19. http://eric.ed.gov/?id=EJ988828

Kingston, N., & Nash, B. (2015). Erratum. *Educational Measurement: Issues and Practice*, *34*(2), 55.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284.

*Martens, B. K., Eckert, T. L., & Begeny, J. C. (2007). Effects of a fluency-building program on the reading performance of low-achieving second and third grade students. *Journal of Behavioral Education*, *16*(1), 39–54. http://eric.ed.gov/?id=EJ757777

Marzano, R. J. (2010). *Formative assessment & standards-based grading.* Bloomington, IN: Marzano Research.

*McCurdy, B. L., & Shapiro, E. S. (1992). A comparison of teacher-, peer-, and self-monitoring with curriculum-based measurement in reading among students with learning disabilities. *The Journal of Special Education*, *26*(2), 162–180. http://eric.ed.gov/?id=EJ451623

McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement: So much more is needed. *Practical Assessment, Research & Evaluation*, *18*(2), 1–15. http://eric.ed.gov/?id=EJ1005135

*Menesses, K. F., & Gresham, F. M. (2009). Relative efficacy of reciprocal and nonreciprocal peer tutoring for students at-risk for academic failure. *School Psychology Quarterly*, *24*(4), 266–275. http://eric.ed.gov/?id=EJ866092

*Meyer, E., Abrami, P. C., Wade, C. A., Aslan, O., & Deault, L. (2010). Improving literacy and metacognition with electronic portfolios: Teaching and learning with ePEARL. *Computers & Education*, *55*(1), 84–91. http://eric.ed.gov/?id=EJ877999

Moss, C. M., & Brookhart, S. M. (2009). *Advancing formative assessment in every classroom: A guide for instructional leaders.* Alexandria, VA: ASCD.

*Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., et al. (2003). Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research*, *29*(1), 61–117. http://eric.ed.gov/?id=EJ773580

National Council of Teachers of Mathematics. (2007). *Benefits of formative assessment: Assessment research clips.* Reston, VA: Author. Retrieved July 11, 2016, from http://www.nctm.org/Research-and-Advocacy/research-brief-and-clips/Benefits-of-Formative-Assessment/.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Noyce, P. E., & Hickey, D. T. (2011). *New frontiers in formative assessment*. Cambridge, MA: Harvard Education Press. http://eric.ed.gov/?id=ED527526

*Null, L. D. H. (1990). *The effects of learning for mastery on first and second-grade decoding skill and general reading achievement*. Unpublished doctoral dissertation, Indiana University, Bloomington.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, *28*(3), 5–13. http://eric.ed.gov/?id=EJ853799

Popham, W. J. (2013). Waving the flag for formative assessment. *Education Week*, *32*(15), 29.

*Ross, J. A., Rolheiser, C., & Hoaboam-Gray, A. (1998, April). *Impact of self-evaluation training on mathematics achievement in a cooperative learning environment*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, California. http://eric.ed.gov/?id=ED422381

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

*Sawyer, R. J., Graham, S., & Harris, K. R. (1992). Direct teaching, strategy instruction, and strategy instruction with explicit self-regulation: Effects on the composition skills and self-efficacy of students with learning disabilities. *Journal of Educational Psychology*, *84*(3), 340–352. http://eric.ed.gov/?id=EJ452404

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14. http://eric.ed.gov/?id=EJ615905

Shepard, L. (2005). Linking formative assessment to scaffolding. *Educational Leadership*, *63*(3), 66–70. http://eric.ed.gov/?id=EJ745460

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. http://eric.ed.gov/?id=EJ787077

U.S. Department of Education. (2008, September). *Accelerated Math (middle school math) intervention report*. Washington, DC: Author. http://eric.ed.gov/?id=ED502796

U.S. Department of Education. (2010, September). *Accelerated Math (elementary school math) intervention report*. Washington, DC: Author. http://eric.ed.gov/?id=ED511797

U.S. Department of Education. (2014a). *Review protocol for beginning reading interventions,* version 3.0. Washington, DC: Author. Retrieved February 4, 2015, from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_br_protocol_v3.0.pdf.

U.S. Department of Education. (2014b). *What Works Clearinghouse procedures and standards handbook,* version 3.0. Washington, DC: Author.

U.S. Department of Education. (2014c). *WWC study review guide template*. Washington, DC: Author. Retrieved August 1, 2014, from http://ies.ed.gov/ncee/wwc/DownloadSRG.aspx.

U.S. Department of Education. (n.d.). *WWC Standards Brief for Attrition*. Washington, DC: Author. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_brief_attrition_080715.pdf.

*Wesson, C. L. (1990). Curriculum-based measurement and two models of follow-up consultation. *Exceptional Children*, *57*(3), 246–256. http://eric.ed.gov/?id=EJ421434

Wiliam, D. (2011). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 18–40). New York, NY: Routledge.

*Ysseldyke, J., & Tardrew, S. (2007). Use of a progress monitoring system to enable teachers to differentiate mathematics instruction. *Journal of Applied School Psychology*, *24*(1), 1–28. http://eric.ed.gov/?id=EJ783511

The Regional Educational Laboratory Program produces 7 types of reports

Making Connections
Studies of correlational relationships

Making an Impact
Studies of cause and effect

What's Happening
Descriptions of policies, programs, implementation status, or data trends

What's Known
Summaries of previous research

Stated Briefly
Summaries of research findings for specific audiences

Applied Research Methods
Research methods for educational settings

Tools
Help for planning, gathering, analyzing, or reporting data or research