Purpose:
The purpose of this study is to systematically review the evidence for the effectiveness of formative assessment interventions in the elementary grades. This study addresses the following four research questions:

1. How effective is formative assessment for elementary school students?
2. Does the effectiveness of formative assessment vary depending on whether it is student-directed or other-directed?
3. Is formative assessment more effective for some subject areas than others?
4. Are different types of formative assessment more effective for particular subject areas?

Background:
Assessment experts and educational leaders have promoted formative assessment as a necessary complement to summative accountability. Prior reviews of evidence for the impact of formative assessment on student achievement have used varying definitions of formative assessment and have produced widely different estimates of its effectiveness (Black & Wiliam, 1998a; 1998b; Kingston & Nash, 2011). We define formative assessment as a process dedicated to both gathering and using assessment information about students learning within a four-week period of time.

The current study improves upon previous reviews by considering the extent to which the studies of formative assessment interventions were conducted rigorously enough to have confidence that the observed effects on student outcomes were indeed caused by the formative assessment intervention. This study used an approach modeled after the What Works Clearinghouse (WWC) (U.S. Department of Education, 2014) evidence standards and procedures to identify studies that support causal inferences.

Teachers report they lack time to both interpret information gathered during formative assessment and plan adjustments based on the interpretation. One approach to minimizing the formative assessment burden on teachers is to more actively involve students in the formative assessment process (Black & Wiliam, 1998a). To shed light on the effectiveness of different approaches to formative assessment, this study examines whether student-directed formative assessment interventions are as effective as interventions that involve other agents, such as educators or software programs, in the formative assessment process.

Research Design:
The study team conducted a systematic and comprehensive search for research on a range of different interventions that met the definition of formative assessment. Two features of the interventions examined in each of the studies were recorded. First, each intervention was classified according to the formative assessment intervention type based on who was the primary agent implementing the process of gathering and using evidence to improve learning: students themselves or others, such as teachers or computer software programs. Second, the academic subject that the intervention addressed was recorded. The studies that met standards addressed three academic subjects: mathematics, reading, and writing.

The search process identified 76 studies. Nine reviewers assessed the quality of the screened-in studies against WWC standards for group designs (U.S. Department of Education, 2014). Twenty-two studies met standards with or without reservations and are included in this paper. Although the authors searched for studies in five core academic areas (mathematics, reading, writing, science, and social studies), no studies...
met standards that focused on science or social studies. Therefore, this paper focuses only on results for mathematics, reading, and writing.

Findings:

**How effective is formative assessment for elementary school students?** Combining across all of the studies, formative assessment has a positive effect on student achievement. The overall average of 30 effect sizes from 19 studies was 0.26 standard deviation. The 30 effect sizes ranged from a low of −0.46 to a high of 1.22.

**Does the effectiveness of formative assessment vary depending on whether it is student-directed or other-directed?** The average effect size for student-directed formative assessment interventions was one-fifth of a standard deviation. The average effect size for other-directed interventions was larger, over a quarter of a standard deviation (table 1).

**Is formative assessment more effective for some subject areas than others?** The average effect size for formative assessment interventions in mathematics was about a third of a standard deviation. The average effect sizes for reading and writing were smaller, slightly greater than a fifth of a standard deviation (table 2).

**Are different types of formative assessment more effective for particular subject areas?** Seven studies examined formative assessment in the academic content area of mathematics. These studies tested 10 comparisons for which effect sizes could be calculated. For student-directed interventions, the average effect size was nearly half of a standard deviation. For other-directed, it was nearly a third of a standard deviation (table 3).

Nine studies examined formative assessments in the academic content area of reading. These studies tested 12 comparisons for which effect sizes could be calculated. Four comparisons examined the impact of student-directed formative assessment interventions, and the remaining eight examined the impact of other-directed formative assessment interventions. The average effect size for student-directed interventions was small and negative (table 4).

Seven studies examined formative assessment interventions in writing. Only one of these studies examined a student-directed formative assessment intervention. With so few student-directed formative assessment interventions, it is impossible to examine the differential effectiveness of the two types. The one student-directed intervention (Sawyer, Graham, & Harris, 1992) had the largest effect observed for a writing formative assessment intervention (0.63), suggesting that more research should be conducted on student-directed formative assessment interventions in writing to determine if this finding is isolated or can be confirmed with future research. The average effect size for the other-directed formative assessment interventions was low (0.15), with effect sizes ranging from −0.20 to 0.34.

Conclusions:
The results of this study provide evidence that formative assessment interventions have a positive impact on student achievement. Both student- and other-directed interventions in mathematics appear to be effective for improving student outcomes. When formative assessment interventions are used in reading, other-directed interventions appear to be more effective. There was not enough evidence to determine the effectiveness of student-directed interventions in writing.

The results of this study confirm the overall positive effect of formative assessment that has been reported in earlier reviews (Black & Wiliam, 1998a, 1998b; Kingston & Nash, 2011, 2015). The results of the study
can help teachers and administrators identify approaches to formative assessment that are appropriate for particular subject areas in the elementary grades.
Table 1. Mean effect sizes for formative assessment interventions, by type

<table>
<thead>
<tr>
<th>Type of formative assessment intervention</th>
<th>Number of studiesa</th>
<th>Number of effect sizesb</th>
<th>Mean effect size</th>
<th>Standard deviation</th>
<th>Minimum effect size</th>
<th>Maximum effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student-directed</td>
<td>7</td>
<td>9</td>
<td>0.20</td>
<td>0.48</td>
<td>-0.46</td>
<td>1.01</td>
</tr>
<tr>
<td>Other-directed</td>
<td>13</td>
<td>21</td>
<td>0.29</td>
<td>0.30</td>
<td>-0.20</td>
<td>1.22</td>
</tr>
</tbody>
</table>

Note: Effect sizes greater than 0.25 or less than -0.25 are bolded.

a The actual column sum for number of studies (20) does not equal the total number of studies reviewed (22) because three studies did not provide enough information to calculate effect sizes and one study (McCurdy & Shapiro, 1992) examined both student-directed and other-directed formative assessment interventions.

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, & Ferguson, 1992; Fuchs, Fuchs, Hamlett, & Allinder, 1991a, 1991b; Fuchs, Fuchs, Hamlett, & Stecker, 1991; Johnson, Graham, & Harris, 1997; Martens, Eckert, & Begeny, 2007; McCurdy & Shapiro, 1992; Ysseldyke & Tardrew, 2007).

Table 2. Mean effect sizes for formative assessment interventions, by subject area

<table>
<thead>
<tr>
<th>Subject area</th>
<th>Number of studiesa</th>
<th>Number of effect sizesb</th>
<th>Mean effect size</th>
<th>Standard deviation</th>
<th>Minimum effect size</th>
<th>Maximum effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>7</td>
<td>10</td>
<td>0.36</td>
<td>0.33</td>
<td>-0.18</td>
<td>1.01</td>
</tr>
<tr>
<td>Reading</td>
<td>7</td>
<td>12</td>
<td>0.22</td>
<td>0.45</td>
<td>-0.46</td>
<td>1.22</td>
</tr>
<tr>
<td>Writing</td>
<td>6</td>
<td>8</td>
<td>0.21</td>
<td>0.24</td>
<td>-0.20</td>
<td>0.63</td>
</tr>
<tr>
<td>Spelling</td>
<td>2</td>
<td>4</td>
<td>0.19</td>
<td>0.09</td>
<td>0.09</td>
<td>0.30</td>
</tr>
<tr>
<td>Composition</td>
<td>4</td>
<td>4</td>
<td>0.22</td>
<td>0.35</td>
<td>-0.20</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Note: Effect sizes greater than 0.25 or less than −0.25 are bolded.

a The actual column sum for number of studies (20) does not equal the total number of studies reviewed (22) because three studies did not provide enough information to calculate effect sizes and one study (Craven, Marsh, & Debus, 1991) examined formative assessment interventions for both reading and mathematics.

b The number of effect sizes is greater than the number of studies because two studies of mathematics (Fuchs, Fuchs, Hamlett, & Stecker, 1991; Ysseldyke & Tardrew, 2007), four studies of reading (Fuchs, Fuchs, Hamlett, & Ferguson, 1992; Johnson, Graham, & Harris, 1997; Martens, Eckert, & Begeny, 2007; 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.
Table 3. Mean effect sizes for formative assessment interventions focused on mathematics, by type

<table>
<thead>
<tr>
<th>Type of formative assessment intervention</th>
<th>Number of studies</th>
<th>Number of effect sizes</th>
<th>Mean effect size</th>
<th>Standard deviation</th>
<th>Minimum effect size</th>
<th>Maximum effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student-directed</td>
<td>4</td>
<td>4</td>
<td>0.45</td>
<td>0.49</td>
<td>-0.18</td>
<td>1.01</td>
</tr>
<tr>
<td>Other-directed</td>
<td>3</td>
<td>6</td>
<td>0.30</td>
<td>0.21</td>
<td>0.07</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Note: Effect sizes greater than 0.25 and less than -0.25 are bolded.

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

Table 4. Mean effect sizes for formative assessment interventions focused on reading, by type

<table>
<thead>
<tr>
<th>Type of formative assessment intervention</th>
<th>Number of studies</th>
<th>Number of effect sizes</th>
<th>Mean effect size</th>
<th>Standard deviation</th>
<th>Minimum effect size</th>
<th>Maximum effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student-directed</td>
<td>2</td>
<td>4</td>
<td>-0.15</td>
<td>0.26</td>
<td>-0.46</td>
<td>0.17</td>
</tr>
<tr>
<td>Other-directed</td>
<td>8</td>
<td>8</td>
<td>0.41</td>
<td>0.41</td>
<td>0.02</td>
<td>1.22</td>
</tr>
</tbody>
</table>

Note: Effect sizes greater than 0.25 are bolded.

a The actual column sum for number of studies (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 interventions.

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


