Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: https://www.tandfonline.com/loi/hame20

The Effectiveness and Features of Formative Assessment in US K-12 Education: A Systematic Review

Hansol Lee, Huy Q. Chung, Yu Zhang, Jamal Abedi & Mark Warschauer

To cite this article: Hansol Lee, Huy Q. Chung, Yu Zhang, Jamal Abedi & Mark Warschauer (2020) The Effectiveness and Features of Formative Assessment in US K-12 Education: A Systematic Review, Applied Measurement in Education, 33:2, 124-140, DOI: 10.1080/08957347.2020.1732383

To link to this article: https://doi.org/10.1080/08957347.2020.1732383

Published online: 02 Mar 2020.

Submit your article to this journal

Article views: 3858

View related articles

View Crossmark data

Citing articles: 23 View citing articles

Full Terms & Conditions of access and use can be found at https://www.tandfonline.com/action/journalInformation?journalCode=hame20

APPLIED MEASUREMENT IN EDUCATION
2020, VOL. 33, NO. 2, 124–140
https://doi.org/10.1080/08957347.2020.1732383

Routledge
Taylor & Francis Group

Check for updates

The Effectiveness and Features of Formative Assessment in US K-12 Education: A Systematic Review

Hansol Lee [a], Huy Q. Chung[b], Yu Zhang[c], Jamal Abedi[d], and Mark Warschauer [b]

[a]Department of English, Korea Military Academy; [b]School of Education, University of California, Irvine; [c]Research & Evaluation Department, Seattle Public Schools; [d]School of Education, University of California, Davis

ABSTRACT

In the present article, we present a systematical review of previous empirical studies that conducted formative assessment interventions to improve student learning. Previous meta-analysis research on the overall effects of formative assessment on student learning has been conclusive, but little has been studied on important features of formative assessment interventions and their differential impacts on student learning in the United States' K-12 education system. Analysis of the identified 126 effect sizes from the selected 33 studies representing 25 research projects that met the inclusion criteria (e.g., included a control condition) revealed an overall small-sized positive effect of formative assessment on student learning (d = .29) with benefits for mathematics (d = .34), literacy (d = .33), and arts (d = .29). Further investigation with meta-regression analyses indicated that supporting student-initiated self-assessment (d = .61) and providing formal formative assessment evidence (e.g., written feedback on quizzes; d = .40) via a medium-cycle length (within or between instructional units; d = .52) were found to enhance the effectiveness of formative assessments.

In the present study, we systematically reviewed previous formative assessment interventions in the United States' K-12 education system to investigate the overall effect of formative assessment on student learning and to identify features of formative assessment interventions and how they influenced learning outcomes. Although the definitions of formative assessment are varied (see Wiliam, 2018 for a review), in her review Brookhart (2018) stated that the different approaches and ideas share the basic concept that assessment information comes to a student as instructional feedback to facilitate student learning. In this way, both the teachers and students can benefit from formative assessment such that teachers can make decisions in their instructions based on students' achievement (Black & Wiliam, 2009; Ruiz-Primo & Brookhart, 2018) and students can understand the current status of their learning to make improvements (Hattie & Timperley, 2007; Sadler, 1989). For this reason, formative assessment is thus believed to have overall positive effects on student achievement and there have been several meta-analytic efforts to quantify its effectiveness (Black & Wiliam, 1998; Graham, Hebert, & Harris, 2015; Kingston & Nash, 2011; Klute, Apthorp, Harlacher, & Reale, 2017). As an extension of such an effort, we conducted a meta-analysis using multilevel regression modeling to yield more accurate descriptive and inferential estimates than previous traditional meta-analytic studies. In this way, we could offer evidence-based suggestions for implementing successful formative assessment strategies in classrooms.

1. Background

According to recent reviews of formative assessment (Brookhart, 2018; Wiliam, 2018), feedback is believed to be at the heart of how formative assessment can be effective in improving student

CONTACT Hansol Lee hansol6461@gmail.com 574, Hwarang-ro, Nowon-gu, Seoul 01805, Republic of Korea
© 2020 Taylor & Francis Group, LLC

learning as the ultimate outcome. Since the term "feedback" was first used in the field of psychology and education in the mid-20th century (e.g., Gagné, 1954; Jenkins, 1948; Roseborough, 1953; Wilson, High, & Beem, 1954), its effectiveness has been assessed from a few different perspectives. For example, Kluger and DeNisi's (1996) historical review identified that most of the studies about instructional feedback conducted in the twentieth century (i.e., from 1905 to 1995) were based on behaviorist and associationist views of learning and had overall small effects on student learning (d = .41). In other words, according to Wiliam (2018) the early studies were limited to perceiving feedback as mere rewards and punishments to increase or decrease learning (see Thorndike, 1927 for the law of effect from a behaviorist view).

By the late-20th century, the conception of feedback changed along with the cognitive and constructivist view of learning (Brookhart, 2018), where learners are posited as subjects of their learning, and meaning construction (Adie, Willis, & Van der Kleij, 2018). At around this turning point, as Black and Wiliam (2006) suggested, feedback began to be conceived as an integral part of learning processes because feedback on student work reflects information about students' performance or understanding (see also Hattie & Timperley, 2007). Along the same lines, the use of assessment to elicit evidence of learning and to provide feedback that moves students' learning forward (i.e., assessment for learning, or formative assessment; see Bennett, 2011) began to appear (Wiliam, 2018).

According to the formative assessment framework presented by Wiliam (2010) based on previously suggested principles of assessment (Broadfoot et al., 2002) and definitions of feedback (Ramaprasad, 1983), the key aspects of formative assessment include the following three processes in learning: (1) making goals, (2) making progress toward the goals, and (3) making better progress. To facilitate these processes, he further suggested several formative assessment strategies not only for teachers but also for both learners and their peers (e.g., student-initiated self-assessment; Andrade & Valtcheva, 2009), to increase the important roles they play in feedback (Brookhart, 2018), thus reflecting the cognitive and constructivist view of learning. Moreover, Clark's (2012) extensive review of 199 sources on assessment, learning, and motivation revealed that teacher feedback based on evidence of learning coming from formative assessment helps to unveil learning processes, thus promoting not only students' self-assessment skills but also their self-regulated learning (SRL), both of which lead to improving their understanding and performance (Andrade, 2010; Andrade & Brookhart, 2016). Overall, over the past decade, there has been a shift in the focus of formative assessment work from the teacher to the learner.

We think that previous meta-analysis research could not fully resolve Wiliam's (2010) concerns about a meta-analysis on formative assessment in general, namely that (1) statistical measures are needed to control for differences across studies when aggregating outcomes from different studies and (2) researchers' attention needs to be shifted toward designs of effective formative assessment in learning environments. Moreover, in the conclusion of her review, Brookhart (2018) raised a concern that formative assessment practices in the field may not follow the changing view about the role of learners in formative assessment, based on empirical evidence (e.g., Boud & Molloy, 2013; Ruiz-Primo & Brookhart, 2018). We found that these concerns were not thoroughly addressed in previous meta-analysis research and thus conducted the present study to address this limitation. In the following section, we analyzed the most recent three meta-analysis studies (i.e., Graham et al., 2015; Kingston & Nash, 2011; Klute et al., 2017) to verify our interpretation.

In brief, it was clear that the previous meta-analysis research found consistent positive effects of formative assessment on student learning. First, Kingston and Nash (2011) analyzed 13 studies and reported a small-sized average impact (d = .20). Moreover, they found that the average impact varied by subject, such as mathematics (d = .17), English language arts (d = .32), and science (d = .09). Second, focusing on formative writing assessments, Graham et al. (2015) collected 27 studies that involved feedback (not necessarily part of formative assessment) and computed a medium-sized overall impact (d = .61). Moreover, they identified that feedback from adults (d = .87), peers (d = .58), the self (d = .62), and computers (d = .38) provided different magnitudes of impact on the learners' writing quality. Lastly, Klute et al. (2017) identified 30 effect sizes from 19 eligible studies for their meta-analysis and revealed an overall positive effect of formative assessment (d = .26). Further, they found that the effectiveness varied by subject area, such as mathematics (d = .36), reading (d = .22), and writing (d = .21).

To address the concerns raised by Wiliam (2010) about aggregating existing differences across studies, Kingston and Nash (2011) computed effect sizes by different subject areas (i.e., mathematics, English language arts, and science), Graham et al. (2015) only focused on one subject (i.e., English writing), and Klute et al. (2017) calculated mean effect sizes by subject areas (i.e., math, reading, writing, spelling, and composition). Except for Graham et al.'s (2015) study where a meta-regression was additionally conducted to determine how the quality of studies, grade level, and feedback structure influenced the overall average effect size estimates, the other studies did not go further to control for differences across studies as Wiliam (2010) had suggested. Regarding the concern about the changed role of learners in formative assessment, although Kingston and Nash (2011) and Klute et al. (2017) included additional analyses of moderators (e.g., grade levels and five treatment types in Kingston & Nash, 2011; two intervention types in Klute et al., 2017), they were not able to statistically compare differences across the values of moderators (i.e., features of formative assessment interventions), limiting their interpretations of their findings. Overall, we found that delving into design features of formative assessment interventions that may influence their effectiveness and assessing indications of the paradigm shift from the teacher to the learner in formative assessment practices, in terms of the source of the feedback, would be an important addition to the previous literature.

2. Present Study

We conducted a meta-analysis with a multilevel modeling approach to ensure higher accuracy in synthesizing the overall and differential effects of formative assessment on student learning. The rationale behind our use of such an approach is to statistically accommodate differences and similarities across the included studies. For example, we found that there were studies that stemmed from the same research projects (e.g., Chen & Andrade, 2018, Chen, Lui, Andrade, Valle, & Mir, 2017; Valle, 2015 came from the Arts Achieve project supported by the New York City Department of Education), and there were studies that had sub-samples and multiple measurements that require separate effect size calculations. Thus, our calculated effect sizes were nested in studies coming from different research projects, making the data set multilevel (i.e., level 1: effect sizes; level 2: studies; and level 3: projects). In this way, we could code various features of formative assessment interventions along with computing the effectiveness without losing statistical power to take their nested data structures into consideration. Further, by using a multilevel regression analysis for effect sizes via hierarchical linear modeling for moderator estimation (i.e., meta-regression), we could not only systematically synthesize the overall impact of formative assessment on student learning, but also investigate features of formative assessment interventions and their differential roles to enhance the effectiveness of formative assessment. In doing so, we believe that the present study will provide the cumulative evidence of positive effects of formative assessment on student learning and contribute to further understanding of features of formative assessment practices that improve student learning. The following are the two research questions that guide these research goals. In what follows we describe the methods we adopted and utilized in order to answer the research questions. The findings and implications are further discussed.

(1) What is the impact of formative assessment on student learning?
(2) What are the features of formative assessment and how do they differentially influence the effectiveness of the formative assessment processes?

3. Methods

To perform a meta-analysis, we went through multiple steps to locate, include/exclude, evaluate, and analyze previous formative assessment interventions. This section describes the methods used in the current study, including the literature search, inclusion/exclusion criteria, effect size (ES, henceforth) calculation, and coding schemes for the meta-analysis, and the data analysis plan.

3.1. *Literature Search*

To locate previous literature (up to June 2018) in the literature review process more objectively than just selecting prominent or frequently cited studies, we conducted keyword searches (i.e., either formative, formative assessment, formative evaluation, or formative test) in the main databases, such as Google Scholar, ProQuest Databases (i.e., ERIC, Dissertations & Theses A&I: Social Sciences, and PsycINFO), and Web of Science. In order to narrow down the search results more precisely, we specified filters when possible for each database, such as language (English), publication date, location (US), subject (excluded non-related topics; e.g., higher education, community college education, religion, etc.), and index terms (excluded non-related keywords; e.g., climate change, parenting, bullying, etc.). Moreover, we checked the related major journals in the field of education and assessment. In addition, the reference lists of the identified studies as well as the previously introduced three meta-analysis studies (i.e., Graham et al., 2015; Kingston & Nash, 2011; Klute et al., 2017) were further reviewed as part of this effort.

3.2. *Inclusion Criteria*

Studies had to meet the following six inclusion criteria: The study (1) was designed to implement a formative assessment intervention, (2) was a true experiment (with a random assignment procedure) or a quasi-experimental study (with a statistical adjustment to check baseline differences) including a control condition, (3) focused on improving student learning, (4) was presented in English, (5) involved students who were in grades kindergarten through 12 in the United States, and (6) included necessary statistics to compute ESs. Figure 1 illustrates the PRISMA flow chart describing how we identified, screened, analyzed, and included studies in our meta-analysis. First, we identified 3,730 records through database searching and 57 studies included in previous meta-analysis research or identified via manual searches. After removing duplicates, we found a total of 3,541 records. Second, during the screening phase we reviewed titles and abstracts and were able to exclude 3,432 records that were not relevant to the main topic. Third, we found that 109 studies were eligible for a full-scale review, of which 76 articles were excluded for the following reasons: (1) 32 studies were not designed to implement formative assessment interventions (studies that provided feedback without formative assessments were excluded); (2) 6 studies did not focus on student achievement; (3) 13 studies were conducted in countries other than the United States; (4) 13 studies were neither experiments nor quasi-experiments; (5) 9 studies did not have comparable control groups; and (6) 3 studies did not have necessary statistics to compute ESs for intervention impacts. In conclusion, only 33 studies remained and were included in our meta-analysis.

We then analyzed the selected 33 studies to compute ESs to investigate the effectiveness of formative assessment and code variables to understand the features of formative assessment interventions. During this phase, we found that some studies came from the *same* research projects, such as Criteria-Reference formative assessment in Arts Achieve (Chen & Andrade, 2018; Chen et al., 2017; Valle, 2015), Classroom Assessment for Student Learning (CASL) program (Randel, Apthorp, Beesley, Clark, & Wang, 2016; Randel et al., 2011), POWERSOURCE© formative assessment (Phelan et al., 2012; Phelan, Choi, Vendlinski,

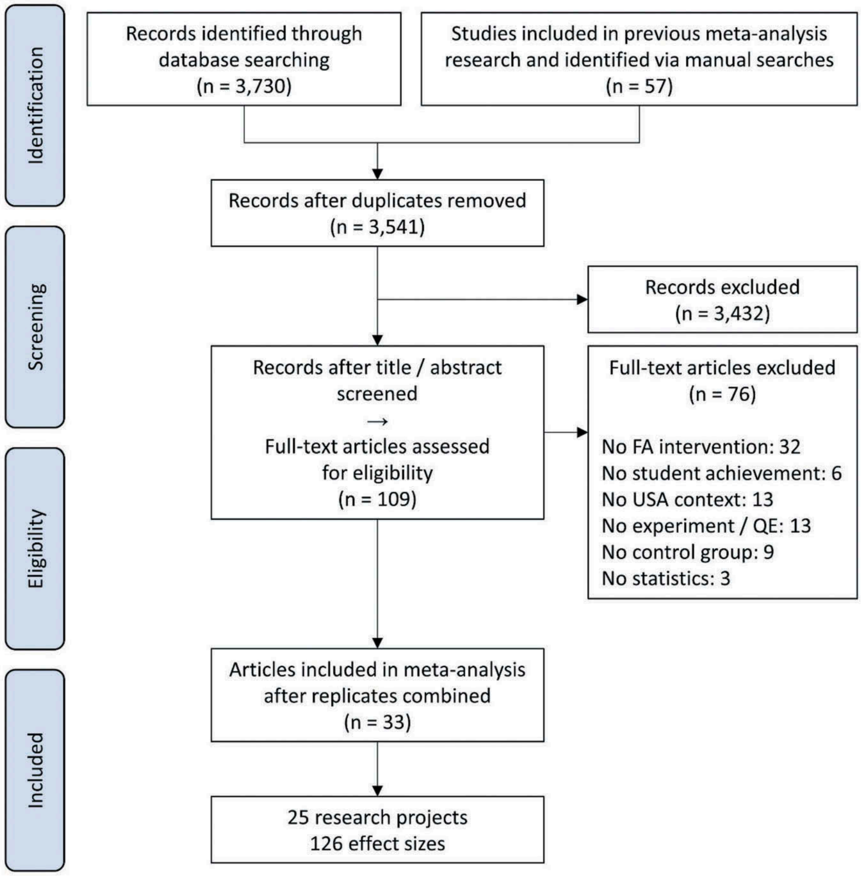

Figure 1. PRISMA flow chart (adapted from Moher, Liberati, Tetzlaff, Altman, & PRISMA Group, 2009).

Baker, & Herman, 2011), Dynamic Indicators of Basic Early Literacy Skills (DIBELS) assessment (Brookhart, Moss, & Long, 2008; Iannuccilli, 2003), Foundational Approaches in Science Teaching (FAST) curriculum I (Tomita, 2008; Yin, 2005), and the 6 + 1 Trait Writing Model (Adler, 1998; Coe, Hanita, Nishioka, & Smiley, 2011; Kozlow & Bellamy, 2004). Table 1 represents the list of studies included in the current meta-analysis.

Furthermore, we found that some studies had multiple *independent* groups of participants (e.g., Brookhart et al., 2008; Koedinger, McLaughlin, & Heffernan, 2010; Kozlow & Bellamy, 2004; LaVenia, 2016; Tomita, 2008; Valle, 2015; Ysseldyke & Tardrew, 2007), and the majority of the collected studies had multiple assessments (e.g., different subjects, different measurements). For these reasons, a study could have more than one ES, and we were able to compute a total of 126 ESs from the 33 studies. Thus, our data set consists of 126 ESs from 33 studies representing 25 research projects, as shown in Table 1, making its structure multilevel (i.e., three levels).

3.3. *Effect Size Calculation*

We used unbiased d (also known as Hedges' g) for the ES calculation to measure the standardized mean differences between formative assessment treated and non-treated groups. Based on the most frequently used Cohen's d ES calculation, unbiased d is computed by multiplying a correction factor (J) to eliminate the upward bias in Cohen's d estimates particularly for small

Table 1. List of studies included in the meta-analysis.

Included study	Search source	Number of ESs	Subject	Project
Adler (1998)	Graham et al. (2015)	4	Literacy	The 6 + 1 Trait Writing Model
Coe et al. (2011)	Graham et al. (2015)	7	Literacy	The 6 + 1 Trait Writing Model
Kozlow and Bellamy (2004)	Graham et al. (2015)	4	Literacy	The 6 + 1 Trait Writing Model
Andrade et al. (2010)	Manual search	1	Literacy	
Benson (1979)	Graham et al. (2015)	1	Literacy	
Bond and Ellis (2013)	Klute et al. (2017)	2	Mathematics	
Brookhart et al. (2008)	Kingston and Nash (2011)	2	Literacy	DIBELS
Iannuccilli (2003)	Klute et al. (2017)	1	Literacy	DIBELS
Butler (2014)	ProQuest	4	Mathematics	
Chen and Andrade (2018)	Manual search	3	Arts	Criteria-Reference FA in Arts Achieve
Chen et al. (2017)	ProQuest	1	Arts	Criteria-Reference FA in Arts Achieve
Valle (2015)	ProQuest	8	Arts	Criteria-Reference FA in Arts Achieve
DeWeese (2012)	ERIC	1	Mathematics	
Fox (2013)	ERIC	3	Literacy	
Franzke, Kintsch, Caccamise, Johnson, and Dooley (2005)	Graham et al. (2015)	12	Literacy	
Irving et al. (2016)	ERIC	5	Mathematics	
Koedinger et al. (2010)	Kingston and Nash (2011)	2	Mathematics	
LaVenia (2016)	ProQuest	2	Mathematics	
Manuel (2015)	PsycINFO	1	Mathematics	
McCurdy and Shapiro (1992)	Klute et al. (2017)	6	Literacy	
Meyen and Greer (2010)	Manual search	6	Mathematics	
Null (1990)	Klute et al. (2017)	2	Literacy	
Phelan et al. (2011)	ERIC	1	Mathematics	POWERSOURCE© FA
Phelan et al. (2012)	ERIC	7	Mathematics	POWERSOURCE© FA
Randel et al. (2011)	ERIC	1	Mathematics	CASL Program
Randel et al. (2016)	ERIC	1	Mathematics	CASL Program
Tomita (2008)	Kingston and Nash (2011)	10	Science	FAST curriculum I
Yin (2005)	Kingston and Nash (2011)	4	Science	FAST curriculum I
Tuominen (2008)	Kingston and Nash (2011)	5	Mathematics	
VanEvera (2003)	Kingston and Nash (2011)	2	Science	
Witmer et al. (2014)	ERIC	3	Literacy	
Yin, Tomita, and Shavelson (2014)	ERIC	2	Science	
Ysseldyke and Tardrew (2007)	Klute et al. (2017)	12	Mathematics	

sample sizes (e.g., $n < 50$; see Hedges & Olkin, 1985). The equations for the ES calculation are shown below.

$$ES_n = J_{Correction\ Factor} \times \frac{Mean_{treated} - Mean_{control}}{\sqrt{\left[\frac{(n_T-1)SD^2_{treated}+(n_C-1)SD^2_{control}}{n_{treated}+n_{control}-2}\right]}}$$

$$SE_n = J_{Correction\ Factor} \times \sqrt{\frac{1}{n_{treated}} + \frac{1}{n_{control}} + \frac{\text{Cohen}'sd^2}{2 \times (n_{treated} + n_{control})}}$$

$$J_{Correction\ Factor} = 1 - \frac{3}{\{4 \times (n_{treated} + n_{control} - 2) - 1\}}$$

3.4. Coding Scheme

To build a data set for our meta-analysis, we developed the following coding scheme. First, based on Bennett's (2011) idea of formative assessment components and McMillan, Venable, and Varier's (2013) suggested formative assessment characteristics, we identified 15 variables regarding features of formative assessment interventions. In addition, we followed a study review protocol suggested by *What Works Clearinghouse Procedures and Standards Handbook, Version 4.0* (What Works Clearinghouse, 2017) about the standards and procedures of experiments. With students' learning outcome as the dependent variable, a total of six variables regarding features of formative assessment interventions as well as two variables regarding contextual factors emerged from our collected data for the coding scheme, as shown in Table 2.

With respect to the features of formative assessment interventions, first, we coded the main sources of the formative assessment feedback (either by helping the teacher to implement formative assessment practices or aiding students to formally assess themselves, or mixed; see Andrade, 2010). In particular, 43 ESs (34%) came from interventions that focused on teachers' formative assessment practices, 29 ESs (23%) came from interventions that focused on student-initiated formative assessments, and the remaining 54 ESs (43%) came from interventions that had both teachers and students as their main source for formative assessment feedback.

Second, we coded the level of formality of formative assessment evidence as formal (e.g., written feedback on quizzes), informal (e.g., oral comments on student's class work), or mixed (see Brookhart, 2018). We found that a majority of ESs (93; 74%) came from interventions that used both formal and informal formative assessment evidence, while only 14 ESs (11%) and 19 ESs (15%) came from interventions that used either formal or informal formative assessment evidence, respectively.

Third, we coded cycle lengths of formative assessment feedback (either within and between instructional units – medium-cycle – or within and between lessons – short-cycle; see Wiliam, 2010). A majority (108 ESs; 86%) came from interventions that provided formative assessment feedback within and between lessons (short-cycle length) while only nine ESs (7%) came from studies that provided formative assessment feedback within and between instructional units (medium-cycle length).

Table 2. Coding scheme and descriptive statistics of variables.

Features of formative assessment		Descriptive statistics # of ESs	# of Studies	# of Projects
Source of FA feedback	(1) Teacher FA practice	43	13	9
	(2) Student-initiated self-assessment	29	7	7
	(3) Mixed	54	13	11
Formality of FA evidence	(1) Formal FA evidence (e.g., written feedback)	14	6	5
	(2) Informal FA evidence (e.g., oral feedback)	19	3	2
	(3) Mixed	93	24	18
Cycle length of FA feedback	(1) Medium-cycle length	9	3	3
	(2) Short-cycle length	108	26	19
	(3) Not defined	9	4	3
Professional development	(1) No PD	26	9	9
	(2) One-time PD	50	12	10
	(3) On-going supports	50	13	10
Computer	(1) Paper-based	85	26	18
	(2) Computer-based	41	7	7
Instructional adjustment	(1) No adjustment	23	7	7
	(2) Planned adjustment	26	3	2
	(3) Unplanned adjustment	16	3	2
	(4) Mixed	61	21	16
School	(1) K-6	60	19	14
	(2) 7–12	66	16	13
Classroom	(1) Regular classroom	111	30	23
	(2) Special education classroom	15	4	4

Fourth, we coded the inclusion of professional development (PD), and three values were identified: no PD, one-time training, and ongoing support. We found that 26 ESs (21%) came from interventions with no PD, 50 ESs (40%) had one-time training, and another 50 ESs (40%) were from programs that provided ongoing professional development.

Fifth, we coded whether the treatment was paper-based or computer-based. We found that 85 ESs (67%) came from paper-based formative assessment interventions and 41 ESs (33%) came from computer-based formative assessment interventions.

Sixth, we coded if teachers made instructional adjustments after formative assessment results were reviewed. Four values emerged: no instructional adjustment, planned instructional adjustment, unplanned instructional adjustment, and mixed, based on 23 ESs (18%), 26 ESs (21%), 16 ESs (13%), and 61 ESs (48%), respectively.

Last, in regard to contextual factors, we identified and coded the school and classroom variables. A total of 60 ESs (48%) were based on formative assessment interventions conducted in K-6 (elementary school) classrooms, while 66 ESs (52%) were in grades 7–12 (middle and high school) classrooms. In terms of types of classrooms for formative assessment interventions, we found that 111 ESs (88%) came from interventions conducted in regular classrooms and 15 ESs (12%) in special education classrooms.

3.5. *Publication Bias*

After the ES calculations, we examined if there was publication bias among the ESs we computed. Given that studies with non-significant findings and small ESs are unlikely to be published in peer-reviewed journals, collecting, and analyzing published studies may yield a biased estimate of an overall treatment effect (see Egger, Smith, Schneider, & Minder, 1997). In general, Egger's test is widely used to statistically check publication bias by testing whether an intercept in a regression line of standardized ES estimate (ES/its standard error) on precision (1/standard error of ES) is different from zero. The results of the test revealed that the bias intercept was 1.57 (SE = .32, p > .05), and it was different from zero according to its 95% confidence interval [.94 ~ 2.20], indicating a possible publication bias in our ES calculation. For this reason, we then used a trim-and-fill approach to handle possible publication bias (Duval & Tweedie, 2000). This method removes studies that cause asymmetry and fills in hypothetical studies to re-estimate ESs. However, the results of this approach showed that no trimming was performed, and the data were unchanged, meaning that the aforementioned possible publication bias could be negligible.

3.6. *Data Analysis*

We used multilevel modeling to compute overall average ES and to conduct moderator analyses to answer the research questions. Because the data set we built for the meta-analysis has a hierarchical structure, such that each ES (level 1) is nested in a unique sample (level 2), which is nested in a research project (level 3), we used a multilevel regression modeling approach. First, we computed multiple ESs for a study, and these ESs were not independent from each other. A multilevel regression analysis retains the cluster membership of each ES by distinguishing different levels of variance components (see Lee, Warschauer, & Lee, 2019 for rationales for a multi-level meta-analysis model), while an ordinary least squares (OLS) regression analysis treats each ES as an independent observation without taking into consideration our nested design. Moreover, by recognizing the membership of ESs this approach allows us to compute more accurate estimates with our data set where multiple studies came from the same research project. In other words, reporting more ESs (or samples) will not contribute more to the estimation of the average ES (see Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013 for rationales for a three-level meta-analysis model).

What makes a multilevel meta-analysis different from regular multilevel modeling is that we already computed sampling errors of the dependent variable (i.e., standard errors of the ESs). For this reason, we included them into the multilevel regression analysis instead of estimating level 1 variance, and this is called a *variance-known* model (see Lee et al., 2019). We used STATA 14 and its command *meglm* to accommodate this unique version of multilevel regression analysis, according to the theoretical suggestion of Hox, Moerbeek, and van de Schoot (2010) and practical guidance of Lee et al. (2019), and the results were replicated in HLM 7.

For the first research question, the following equation was used

$$Effect\ size_{ijk} = \gamma_{000} + r_{0jk} + u_{00k} \quad (1)$$

where $Effect\ size_{ijk}$ is a computed ES i among multiple ESs in sample j coming from research project k, γ_{000} is the weighted mean, r_{0jk} is the level 2 variance, and $u_{00\ k}$ is the level 3 variance. In the equation, the level 1 variance is omitted because it is replaced by the variance of ESs. Similarly, to answer the second research question, we first identified moderator variables during the coding process and used the following Equation (2) to compute average ESs for each moderator:

$$Effect\ size_{ijk} = \gamma_{000} + \gamma_{100} \times Moderator\ variable_{ijk} + r_{0jk} + u_{00k} \quad (2)$$

where $\gamma_{000} + \gamma_{100}$ is the weighted mean of the moderator variable included in the equation. As a result, the adjusted means of each variable as well as the contrasts between these values are reported; therefore, the values should be interpreted after keeping other variables at their averages (or holding others constant).

4. Results

4.1. *RQ #1. Impact of Formative Assessment on Student Learning*

After collecting, analyzing, and coding formative assessment interventions, we generated a total of 126 ESs from 33 studies, which came from 25 research projects. Table 3 represents the mean ES estimation, including the mean ES estimates and there standard errors. As a result, for the first research question, we found an overall small-sized ES across all the studies (d = .29, p < .001).

For each subject area, we calculated 46 ESs (37%) from interventions focused on literacy (12 studies, 9 projects), 50 ESs (40%) from mathematics (14 studies, 12 projects), 18 ESs (14%) from science (4 studies, 3 projects), and 12 ESs (9%) from arts (3 studies, 1 project). Their mean ES indicated that the average impact of formative assessment interventions varied by subject: literacy (d = .33, p < .001), mathematics (d = .34, p < .001), science (d = .13, p < .001), and arts (d = .29, p < .001). In particular, we found that science had a relatively lower average ES (i.e., marginal size; d < .20), and a similar issue has been previously identified by Kingston and Nash (2011), who conjectured that science (and math in their case) tasks could be cognitively more complex than other tasks.

Table 3. Mean effect size estimation.

Estimation	# of ESs	# of Studies	# of Projects	Mean ES	SE
Overall	126	33	25	.29***	(.05)
Literacy	46	12	9	.33***	(.08)
Mathematics	50	14	12	.34***	(.09)
Science	18	4	3	.13***	(.00)
Arts	12	3	1	.29***	(.08)

*** p < .001

Table 4. Analyses of moderator variables (Meta-regression).

Moderator variables		Predicted mean	SE	Difference
Source of FA feedback	(1) Teacher FA practice	0.18*	(0.09)	(2) > (1) ≈ (3)
	(2) Student-initiated self-assessment	0.61***	(0.16)	
	(3) Mixed	0.13	(0.08)	
Formality of FA evidence	(1) Formal FA evidence	0.40**	(0.13)	No difference
	(2) Informal FA evidence	0.18	(0.20)	
	(3) Mixed	0.26***	(0.05)	
Cycle length of FA feedback	(1) Medium-cycle length	0.52***	(0.15)	No difference
	(2) Short-cycle length	0.24***	(0.04)	
	(3) Not defined	0.24	(0.19)	
Professional development	(1) No PD	0.18	(0.11)	No difference
	(2) One-time PD	0.27***	(0.07)	
	(3) On-going supports	0.30***	(0.07)	
Computer	(1) Paper-based	0.29***	(0.06)	No difference
	(2) Computer-based	0.21*	(0.10)	
Instructional adjustment	(1) No adjustment	0.16	(0.18)	No difference
	(2) Planned adjustment	0.35*	(0.16)	
	(3) Unplanned adjustment	0.15	(0.13)	
	(4) Mixed	0.29***	(0.06)	
School	(1) K-6	0.29***	(0.05)	No difference
	(2) 7–12	0.23***	(0.06)	
Classroom	(1) Regular classroom	0.25***	(0.04)	No difference
	(2) Special education classroom	0.35**	(0.12)	

Values in light gray and dark gay are small and medium effect sizes, respectively; this only applies to statistically significant values.
* $p < .05$, ** $p < .01$, *** $p < .001$.

4.2. *RQ #2. Features of Formative Assessment and Their Differential Impacts*

To answer the second research question, Table 4 describes the predicted means of moderator variables, representing their differential impacts on student learning after controlling for other features of formative assessment interventions. Also, the contrast results from multiple regression analyses are reported, which statistically compared the predicted means between different categorical values of each moderator variable.

First, we examined the *source of formative assessment feedback* variable. The results from meta-regression showed that when other features of formative assessment interventions are equal, interventions focusing on supporting students' self-assessment as part of formative assessment had an average ES of .61 ($p < .001$), while improving teachers' formative assessment practices as the focus had an average ES of .18 ($p < .05$). For the interventions that focused on both the teacher and student as the main source of the feedback had an average ES of .13 ($p > .05$). Moreover, the positive impact of providing student-initiated formative assessments was significantly higher ($p < .05$) than the two other categories, whose effectiveness were similar with each other ($p > .05$). Perhaps formative assessment interventions that *particularly* focused on supporting student-initiated self-assessment were more directly related to promoting learners' active role in feedback than other formative assessment intervention types.

Second, with respect to the *formality of formative assessment evidence*, we found predicted mean ESs of .40 ($p < .01$) and .18 ($p > .05$), for providing formal formative assessment evidence and informal formative assessment evidence, respectively. Interventions that provided both formal and informal formative assessment evidence and those that did not report related information had an average ES of .26 ($p < .001$).

Third, we identified different *cycle lengths of formative assessment feedback*, such as medium- and short-cycle length. The predicted mean ESs of .52 ($p < .001$) and .24 ($p < .001$) were calculated for medium- and short-cycle length, respectively. The interventions that did not report related information had an average ES of .24 ($p > .05$).

Fourth, in regard to *professional development for formative assessment*, three values were coded. The predicted mean ESs indicated that after controlling for other features of formative assessment, providing ongoing teacher training support (d = .30, p < .001) could be the most effective way to enhance the effectiveness of formative assessment, and providing one-time PD (d = .27, p < .001) could be more effective than not providing PD (d = .18, p > .05), though the ES differences were not statistically significant.

Fifth, whether *formative assessment interventions were paper-based or computer-based* was examined. The predicted mean ESs indicated similar sizes of intervention impact across the categories: mean ESs of .29 (p < .001) and .21 (p < .05) for paper-based and computer-based formative assessment interventions, respectively.

Sixth, the *instructional adjustment for formative assessment* was coded with four categories: no instructional adjustment, planned instructional adjustment, unplanned instructional adjustment, and mixed. The results indicated that only formative assessment interventions with the planned instructional adjustment component (i.e., the planned instructional adjustment and mixed categories) were effective in promoting student learning outcomes after keeping other features of formative assessment constant. That is, the planned instructional adjustment and mixed categories had ESs of .35 (p < .05) and .29 (p < .001), respectively, while the no instructional adjustment and unplanned instructional adjustment categories had marginal ESs (d = .16 & .15, respectively; p > .05).

Last, in regard to *contextual factors*, we identified and coded the school and classroom variables. The predicted mean ESs showed that there were small-sized intervention impacts (.23 ≤ d ≤ .35) across elementary and middle/high schools and across regular and special classrooms. In regards to classroom types, we could not find any supporting evidence of one of the widespread beliefs that formative assessment interventions might be more successful for students with lower achievement or special needs (e.g., Fuchs & Fuchs, 1986).

5. Discussion

In the present meta-analysis, we calculated a total of 126 ESs from 33 studies (coming from 25 projects) and found an overall mean ES of .29 (SE = .05, p < .001) with slightly different means for each subject category, such as mathematics, arts, literacy, and science, extending previous findings from meta-analyses (Graham et al., 2015; Kingston & Nash, 2011; Klute et al., 2017) and the critiques about how to define formative assessment components articulated in Bennett (2011). Furthermore, the key findings from our meta-analysis provide a more nuanced picture of the impact of formative assessments on student learning by considering important features of these interventions, such as the main sources of formative assessment feedback, the formality of formative assessment evidence, the cycle lengths of formative assessment feedback, the professional development provided to teachers, and the instructional adjustments made after data were reviewed. Based on the findings, we believe the present study significantly contributes to extending our understanding of how to effectively operationalize formative assessment practices to improve student learning in US K-12 education. In this section, we discuss our two key findings, such as (1) student-initiated self-assessment as part of formative assessment and (2) implementation of effective formative assessment practices. Our discussions of the limitations of the present study and future suggestions follow.

5.1. *Student-Initiated Self-Assessment in Formative Assessment*

We believe that the most important finding of the present meta-analysis is related to the main sources of formative assessment feedback. Above all, results indicated that formative assessment interventions were most effective when focused on providing student-initiated formative assessments (i.e., medium-sized impact; d = .61). As the only significant contrast from the meta-regression analysis for moderator variables, the impact of such an approach was statistically larger than the interventions designed to promote teachers' formative assessment practices.

What makes this finding so important is that it is empirical evidence supporting the cognitive and constructivist view of learning for formative assessment, where the learners' active role is considered essential to successful formative assessment (Andrade & Brookhart, 2019; Brookhart, 2018; Clark, 2012; Wiliam, 2018). For example, "understanding learning intentions and criteria for success" and "activating students as the owners of their own learning" (p. 63) are two of the five key strategies for effective formative assessment proposed by Wiliam and Thompson (2008), which are the roles of the learners in the three processes of formative assessment: (1) what are the goals (where the learner is going); (2) what progress is being made toward the goal (where the learner is right now); and (3) how to make better progress (how to get there; adapted from Wiliam, 2010; Wiliam & Thompson, 2008). In other words, as Andrade and Brookhart (2016), Clark (2012), and Butler and Winne (1995) suggested, feedback based on formative assessment evidence offered to the learner may lead to their self-regulated learning, which would be a complementary process to self-assessment in improving their learning outcomes.

For example, in Bond and Ellis (2013) study, where students in the treatment condition (n = 46) were taught a formative self-assessment strategy to help themselves to delve deeper into their thinking about their current learning (see Shepard, 2008), they found that the four-week-long intervention had a significant positive effect on students' mathematics achievement. Their idea was that focusing on students themselves may help them to gauge "a sense of their progress" (Bond & Ellis, 2013, p. 227), which corresponds to the aforementioned two key strategies for formative assessment presented by Wiliam and Thompson (2008). Similarly, in the context of middle-school English writing, Andrade, Du, and Mycek (2010) confirmed the positive effect of providing students with rubrics as self-assessment tools along with reading a model essay and generating criteria on effective writing. Based on the cognitive and constructivist view of learning for formative assessment (Black & Wiliam, 1998; Brookhart, 2018; Stiggins, 2001), Andrade et al. concluded that student-centered approach to formative assessment may help students to be active in understanding the quality of their writing and developing their writing skills.

Likewise, our findings indicated that when the learner is active in their own learning the effectiveness of formative assessment interventions was further enhanced (Andrade & Valtcheva, 2009). Further, we believe this finding indicates the need for more processes that encourage students to take a more proactive role in their own learning progressions, which is in line with the recent paradigm shift in education from teacher-centered to student-centered instruction (Kaufman, 2004). On a side note, we found evidence that Brookhart's (2018) concern about a possible persistence of behaviorist and associationist views of learning in instructional practices in the field was not without basis, provided that only seven studies (21%) out of 33 studies included in the meta-analysis focused on student-initiated formative assessments.

5.2. *Implementation of Effective Formative Assessment Practices*

Another important finding of the present study is that we provided evidence-based implications for implementing effective formative assessment practices. The formative assessment process has been considered as an effective way to help learners develop targeted skills, standards, and outcomes when harnessed appropriately and with the right contextual conditions (e.g., Heritage, 2010; Moss & Brookhart, 2009; Popham, 2008; Wiliam, 2011). As learning from teaching (Hiebert, Morris, Berk, & Jansen, 2007) is an iterative process of constantly monitoring learning goals (e.g., Daro, Mosher, & Corcoran, 2011; McManus, 2008; Popham, 2007, 2008; Stylianides & Stylianides, 2007), collecting evidence of learning, analyzing teaching and learning from said evidence (e.g., Hattie & Timperley, 2007; Heritage, 2010; Sadler, 1989, 1998; Shute, 2007), and making decisions based on this analysis (e.g., Popham, 2008), so to is the process of effective formative assessment in the classroom.

Along these lines, the findings from our meta-regression represented a dynamic interaction of the identified features of formative assessment interventions to promote student learning, in addition to supporting student-initiated self-assessment as part of formative assessment, which was addressed in the previous section. For example, we found that providing formal formative assessment feedback, such as written feedback on quizzes (e.g., DeWeese, 2012; Phelan et al., 2011, 2012), within and between instructional units (medium-cycle length; e.g., Butler, 2014; Witmer, Duke, Billman, & Betts, 2014), would be more effective than just providing informal feedback, such as oral feedback on classroom work (e.g., Tomita, 2008; Tuominen, 2008; Yin, 2005), within and between lessons (short-cycle length; e.g., Coe et al., 2011; Fox, 2013; Kozlow & Bellamy, 2004). In particular, for instructional adjustment, educators are encouraged to adjust their instruction in light of the assessment evidence, and this should be carefully planned ahead of their teaching (e.g., Tomita, 2008; Yin, 2005; Ysseldyke & Tardrew, 2007). Further, the findings suggest that researchers and policymakers should be aware that providing the necessary professional development and ongoing support for formative assessment is crucial to its success, provided that considerable knowledge of implementing formative assessment practices effectively in their classrooms is needed for teachers (Bennett, 2011).

5.3. *Limitations and Future Directions*

The present study is not without limitations that should be addressed in future studies. In order to resolve the previous concerns about aggregating different outcome measures from different disciplines, we computed multiple ESs for a study and used multilevel modeling to accommodate the complex structure of our data set. By using a meta-regression, we could control for various intervention and contextual features of formative assessment to compute average ESs for each moderator variable. However, we could not explain why the formative assessment interventions in science had a lower average ES than other disciplines. Kingston and Nash (2011) presumed that it may lie in the nature of content areas, similar to Wiliam's (2018) concern that outcomes in different subjects may vary in their sensitivity to the effects of formative assessment interventions. Another potential explanation is found in Yin (2005), where a formative assessment intervention examined if the formative assessment can improve students' achievement and lead to conceptual change in science. Results indicated non-significant effects of the intervention, and Yin hinted that the quality of implementation (also known as fidelity) could be a possible reason. In other words, for disciplines where a cognitive approach (e.g., restructuring or reorganizing existing knowledge) is a research goal, formative assessment interventions may not be as impactful as we assumed, if the fidelity of implementation is not ensured. Still, we could not find any supporting evidence, so future studies are encouraged to resolve this issue. In line with the same concern, Wiliam (2018) suggested that learning outcomes can be distinguished between performance and learning to mitigate the sensitivity issue. We agree that this is a very important factor which will offer meaningful research and pedagogical implications. Nevertheless, we could not follow this suggestion when reviewing and coding the included studies due to practical reasons; therefore, future studies are encouraged to do so. If possible, one can differentiate the effects of formative assessment interventions in accordance with their different foci either on students' improvement on an assignment or on students' learning new materials or skills. We conjecture that this issue may be closely related to the unanswered questions regarding why interventions in science had lower ESs than other disciplines.

In summary, we conducted a meta-analysis in the present study to systematically review previous formative assessment interventions conducted in US K-12 education. Our aims were to assess the overall effectiveness of this type of assessment in improving students' academic achievement and to identify essential features of formative assessment and their differential contributions to its effectiveness. We found an overall small-sized positive effect of formative assessment on student learning, and the results of meta-regression indicated a complex interaction of the features of formative assessment for successful teaching and learning. Although the included empirical studies demonstrated important potential features of formative assessment, overall incomplete presentations of interventions in the identified studies placed a limit on our coding scheme. As Bennett (2011) states, we agree that formative assessment "is both conceptually and practically still a work-in-progress" (p.

21). For this reason, if future empirical studies would report detailed descriptions of their formative assessment interventions, we can expect more accurate and comprehensive meta-analysis research to better understand the effectiveness and features of formative assessment. Finally, meta-analysis efforts should be continued to obtain more cumulative evidence of the positive effects of formative assessment on students' learning outcomes.

Acknowledgments

An earlier version of the article was presented at the American Educational Research Association (AERA) 2018 Annual Meeting, April 17, New York City, NY. This research was supported by the National Science Foundation (NSF) under award number 1020393. The opinions expressed are those of the authors and do not represent views of the NSF.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Hansol Lee http://orcid.org/0000-0002-6912-7128
Mark Warschauer http://orcid.org/0000-0002-6817-4416

References

Adie, L. E., Willis, J., & Van der Kleij, F. M. (2018). Diverse perspectives on student agency in classroom assessment. *The Australian Educational Researcher*, *45*(1), 1–12. doi:10.1007/s13384-018-0262-2

*Adler, M. (1998). *The effects of instruction in 6+1 trait writing on third grade students' writing abilities and attitudes toward writing* (Unpublished master's thesis). Emporia State University, Emporia, KS.

Andrade, H. (2010). Students as the definite source of formative assessment: Academic self-assessment and the self-regulation of learning. In G. J. Cizek & H. L. Andrade (Eds.), *Handbook of formative assessment* (pp. 90–105). New York, NY: Routledge.

Andrade, H., & Brookhart, S. M. (2016). The role of classroom assessment in supporting self-regulated learning. In D. Laveault & L. Allal (Eds.), *Assessment for learning: Meeting the challenge of implementation* (pp. 293–309). Heidelberg, Germany: Springer.

Andrade, H., & Brookhart, S. M. (2019). Classroom assessment as the co-regulation of learning. *Assessment in Education: Principles, Policy & Practice*. Advance Online Publication. doi:10.1080/0969594X.2019.1571992

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory into Practice*, *48*(1), 12–19. doi:10.1080/00405840802577544

*Andrade, H., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 199–214. doi:10.1080/09695941003696172

Bennett, R. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. doi:10.1080/0969594X.2010.513678

*Benson, N. L. (1979). *The effects of peer feedback during the writing process on writing performance, revision behavior, and attitude toward writing* (Unpublished doctoral dissertation). University of Colorado, Boulder, CO.

Black, P., & Wiliam, D. (2006). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 81–100). London, UK: Sage.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. doi:10.1080/0969595980050102

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*, 5–31. doi:10.1007/s11092-008-9068-5

*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. doi:10.1111/ssm.12021

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment and Evaluation in Higher Education*, *38*, 698–712. doi: 10.1080/02602938.2012.691462

Broadfoot, P. M., Daugherty, R., Gardner, J., Harlen, W., James, M., & Stobart, G. (2002). *Assessment for learning: 10 principles*. Cambridge, UK: University of Cambridge School of Education.

Brookhart, S. M. (2018). Summative and formative feedback. In A. Lipnevich & J. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 52–78). Cambridge, UK: Cambridge University Press. doi:10.1017/9781316832134.005

*Brookhart, S. M., Moss, C. M., & Long, B. A. (2008). Teacher inquiry into formative assessment practices in remedial reading classrooms. *Assessment in Education: Principles, Policy & Practice*, *17*, 41–58. doi:10.1080/09695940903565545

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*, 245–281. doi:10.3102/00346543065003245

*Butler, M. D. (2014). *The effects of embedding formative assessment measures in a problem-based learning mathematics curriculum for middle school students* (Unpublished doctoral dissertation). University of Kentucky, Lexington, KY.

*Chen, F., Lui, A. M., Andrade, H., Valle, C., & Mir, H. (2017). Criteria-referenced formative assessment in the arts. *Educational Assessment, Evaluation and Accountability*, *29*(3), 297–314. doi:10.1007/s11092-017-9259-z

*Chen, F., & Andrade, H. (2018). The impact of criteria-referenced formative assessment on fifth-grade students' theater arts achievement. *The Journal of Educational Research*, *111*(3), 310–319. doi:10.1080/00220671.2016.1255870

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, *24*(2), 205–249. doi:10.1007/s10648-011-9191-6

*Coe, M., Hanita, M., Nishioka, V., & Smiley, R. (2011). *An investigation of the impact of the 6+1 trait writing model on grade 5 student writing achievement* (NCEE 2012-4010). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Daro, P., Mosher, F., & Corcoran, T. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction* (Consortium for Policy Research in Education Report #RR-68). Philadelphia, PA: Consortium for Policy Research in Education.

*DeWeese, S. V. (2012). *The effects of mastery learning correctives on academic achievement and student affect* (Unpublished doctoral dissertation). Mercer University, Macon, GA.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. doi:10.1111/j.0006-341X.2000.00455.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*(7109), 629–634. doi:10.1136/bmj.315.7109.629

*Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary street: Computer support for comprehension and writing. *Journal of Educational Computing Research*, *33*, 53–80. doi:10.2190/DH8F-QJWM-J457-FQVB

*Fox, J. M. (2013). *The effects of write score formative assessment on student achievement* (Unpublished doctoral dissertation). Tennessee State University, Knoxville, TN.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, *53*(3), 199–208. doi:10.1177/001440298605300301

Gagné, R. M. (1954). Training devices and simulators: Some research issues. *American Psychologist*, *9*(3), 95–107. doi:10.1037/h0062991

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, *115*(4), 523–547. doi:10.1086/681947

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. doi:10.3102/003465430298487

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York, NY: Academic Press.

Heritage, M. (2010). *Formative assessment*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Hiebert, J., Morris, A. K., Berk, D., & Jansen, A. (2007). Preparing teachers to learn from teaching. *Journal of Teacher Education*, *58*(1), 47–61. doi:10.1177/0022487106295726

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

*Iannuccilli, J. A. (2003). *Monitoring the progress of first-grade students with dynamic indicators of basic early literacy skills* (Unpublished doctoral dissertation). Indiava University of Pennsylvania, Indiana, PA.

*Irving, K., Pape, S., Owens, D., Abrahamson, L., Silver, D., & Sanalan, V. (2016). Classroom connectivity and Algebra 1 achievement: A three-year longitudinal study. *Journal of Computers in Mathematics and Science Teaching*, *35*(2), 131–151. Retrieved from https://www.learntechlib.org/primary/p/149118/

Jenkins, D. H. (1948). Feedback and group self-evaluation. *Journal of Social Issues*, *4*(2), 50–60.

Kaufman, D. (2004). Constructivist issues in language learning and teaching. *Annual Review of Applied Linguistics*, *24*, 303–319. doi:10.1017/S0267190504000121

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. doi:10.1111/j.1745-3992.2011.00220.x

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*, 254–284. doi:10.1037/0033-2909.119.2.254

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: Regional Educational Laboratory Central.

*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. doi:10.1177/002246699202600203

*Kozlow, M., & Bellamy, P. (2004). *Experimental study on the impact of the 6+1 trait writing model on student achievement in writing*. Portland, OR: Northwest Regional Educational Laboratory.

*Koedinger, K. R., McLaughlin, E. A., & Heffernan, N. T. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Educational Computing Research*, *43*(4), 489–510. doi:10.2190/EC.43.4.d

*Manuel, A. K. (2015). *The effects of immediate feedback using a student response system on math achievement of eleventh grade students* (Unpublished doctoral dissertation). Mercer University, Macon, GA.

*LaVenia, M. (2016). *Mathematics formative assessment system: Testing the theory of action based on the results of a randomized field trial* (Unpublished doctoral dissertation). Florida State University, Tallahassee, FL.

Lee, H., Warschauer, M., & Lee, J. H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*. *40*(5), 721–753. doi:10.1093/applin/amy012

McManus, S. (2008). *Attributes of effective formative assessment*. Washington, DC: Council of Chief State School Officers.

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. Retrieved from https://pareonline.net/

*Meyen, E. L., & Greer, D. L. (2010). Applying technology to enhance STEM achievement for students with disabilities: The blending assessment with instruction program. *Journal of Special Education Technology*, *25*(3), 49–63. doi:10.1177/016264341002500306

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264–269. doi:10.7326/0003-4819-151-4-200908180-00135

Moss, C., & Brookhart, S. (2009). *Advancing formative assessment in every classroom: A guide for instructional leaders*. Alexandria, VA: Association for Supervision and Curriculum Development.

*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, IN.

*Phelan, J., Choi, K., Vendlinski, T., Baker, E., & Herman, J. (2011). Differential improvement in student understanding of mathematical principles following formative assessment intervention. *The Journal of Educational Research*, *104*(5), 330–339. doi:10.1080/00220671.2010.484030

*Phelan, J. C., Choi, K., Niemi, D., Vendlinski, T. P., Baker, E. L., & Herman, J. (2012). The effects of POWERSOURCE© assessments on middle-school students' math performance. *Assessment in Education: Principles, Policy & Practice*, *19*(2), 211–230. doi:10.1080/0969594X.2010.532769

Popham, W. J. (2007). All about accountability/The lowdown on learning progressions. *Educational Leadership*, *64*(7), 83–84. Retrieved from http://www.ascd.org

Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, *28*(1), 4–13. doi:10.1002/bs.3830280103

*Randel, B., Beesley, A. D., Apthorp, H., Clark, T. F., Wang, X., Cicchinelli, L. F., & Williams, J. M. (2011). *Classroom assessment for student learning: The impact on elementary school mathematics in the central region* (NCEE 2011-4005). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

*Randel, B., Apthorp, H., Beesley, A. D., Clark, T. F., & Wang, X. (2016). Impacts of professional development in classroom assessment on teacher and student outcomes. *The Journal of Educational Research*, *109*(5), 491–502. doi:10.1080/00220671.2014.992581

Roseborough, M. E. (1953). Experimental studies of small groups. *Psychological Bulletin*, *50*(4), 275–303. doi: 10.1037/h0059491

Ruiz-Primo, M. A., & Brookhart, S. M. (2018). *Using feedback to improve learning*. London, UK: Routledge.

Sadler, D. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144. doi:10.1007/BF00117714

Sadler, D. (1998). Formative assessment: Revisiting the territory. *Assessment in Education*, *5*(1), 77–84. doi:10.1080/0969595980050104

Shepard, L. (2008). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 279–303). New York, NY: Lawrence Erlbaum Associates.

Shute, V. (2007). *Focus on formative feedback* (ETS Research Report RR-07-11). Princeton, NJ: Educational Testing Service.

Stiggins, R. J. (2001). *Student-involved classroom assessment* (3rd ed.). Upper Saddle River, NJ: Merrill/Prentice-Hall.
Stylianides, A. J., & Stylianides, G. J. (2007). Learning mathematics with understanding: A critical consideration of the learning principle in the principles and standards for school mathematics. *The Mathematics Enthusiast*, *4*(1), 103–114. Retrieved from https://scholarworks.umt.edu/tme/vol4/iss1/8
Thorndike, E. L. (1927). The law of effect. *American Journal of Psychology*, *39*, 212–222. doi:10.2307/1415413
*Tomita, M. K. (2008). *Examining the influence of formative assessment on conceptual accumulation and conceptual change* (Unpublished doctoral dissertation). Stanford University, Stanford, CA.
*Tuominen, K. R. (2008). *Formative assessment and collaborative teaming with support involving middle school mathematics teachers* (Unpublished doctoral dissertation). UCLA, Los Angeles, CA.
*Valle, C. (2015). *Effects of criteria-referenced formative assessment on achievement in music* (Unpublished doctoral dissertation). State University of New York, Albany.
Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*(2), 576–594. doi:10.3758/s13428-012-0261-6
*VanEvera, W. C. (2003). *Achievement and motivation in the middle school science classroom: The effects of formative assessment feedback* (Unpublished doctoral dissertation). Harvard University, Cambridge, MA.
What Works Clearinghouse. (2017). *WWC procedures and standards handbook* (version 4.0). Retrieved from https://ies.ed.gov
Wiliam, D., & Thompson, M. (2008). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Mahwah, NJ: Lawrence Erlbaum Associates.
Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. Andrade & G. Cizek (Eds.), *The handbook of formative assessment* (pp. 17–40). New York, NY: Routledge.
Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. doi:10.1016/j.stueduc.2011.03.001
Wiliam, D. (2018). Feedback: At the heart of – But definitely not all of – Formative assessment. In A. Lipnevich & J. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 3–28). Cambridge, UK: Cambridge University Press. doi:10.1017/9781316832134.003
*Yin, Y. (2005). *The influence of formative assessments on student motivation, achievement, and conceptual change* (Unpublished doctoral dissertation). Harvard University, Cambrige, MA.
*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. doi:10.1300/J370v24n01_01
*Witmer, S. E., Duke, N. K., Billman, A. K., & Betts, J. (2014). Using assessment to improve early elementary students' knowledge and skills for comprehending informational text. *Journal of Applied School Psychology*, *30*(3), 223–253. doi:10.1080/15377903.2014.924454
*Yin, Y., Tomita, M. K., & Shavelson, R. J. (2014). Using formal embedded formative assessments aligned with a short-term learning progression to promote conceptual change and achievement in science. *International Journal of Science Education*, *36*(4), 531–552. doi:10.1080/09500693.2013.787556
Wilson, R. C., High, W. S., Beem, H. P., & Comrey, A. L. (1954). A factor-analytic study of supervisory and group behavior. *Journal of Applied Psychology*, *38*(2), 89–92. doi: 10.1037/h0062593