frontiers | Frontiers in Psychology

TYPE Review
PUBLISHED 22 August 2022
DOI 10.3389/fpsyg.2022.990196

OPEN ACCESS

EDITED BY
Cheng Yong Tan,
The University of Hong Kong,
Hong Kong SAR, China

REVIEWED BY
Qiuxian Joy Chen,
Shanxi University, China
Harpreet Kaur Dhir,
Arizona College of Nursing,
United States

*CORRESPONDENCE
Alan Cheung
alancheung@cuhk.edu.hk

SPECIALTY SECTION
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

RECEIVED 09 July 2022
ACCEPTED 29 July 2022
PUBLISHED 22 August 2022

CITATION
Xuan Q, Cheung A and Sun D (2022) The effectiveness of formative assessment for enhancing reading achievement in K-12 classrooms: A meta-analysis.
Front. Psychol. 13:990196.
doi: 10.3389/fpsyg.2022.990196

The effectiveness of formative assessment for enhancing reading achievement in K-12 classrooms: A meta-analysis

Qianying Xuan, Alan Cheung* and Dan Sun

The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR, China

This quantitative synthesis included 48 qualified studies with a total sample of 116,051 K-12 students. Aligned with previous meta-analyses, the findings suggested that formative assessment generally had a positive though modest effect (ES = + 0.19) on students' reading achievement. Meta-regression results revealed that: (a) studies with 250 or less students yielded significantly larger effect size than large sample studies, (b) the effects of formative assessment embedded with differentiated instruction equated to an increase of 0.13 SD in the reading achievement score, (c) integration of teacher and student directed assessment was more effective than assessments initiated by teachers. Our subgroup analysis data indicated that the effect sizes of formative assessment intervention on reading were significantly different between Confucian-heritage culture and Anglophone culture and had divergent effective features. The result cautions against the generalization of formative assessment across different cultures without adaptation. We suggest that effect sizes could be calculated and intervention features be investigated in various cultural settings for practitioners and policymakers to implement tailored formative assessment.

KEYWORDS

reading achievement, K-12 students, differentiated instruction, meta-analysis, formative assessment

Introduction

In an era of reconfiguring the relationship between learning and assessment, spurred by quantitative and qualitative evidence, formative assessment is proffered to meet the goals of lifelong learning and promote high-performance and high equity for all students (OECD, 2008). It has gained momentum among researchers and practitioners in various culture contexts. In an oft-cited 'configurative review' (Sandelowski et al., 2012) on formative assessment, Black and Wiliam (1998) reported that effect sizes of formative assessment of student achievements were between 0.4 and 0.7 ranging over age groups from 5-year-olds to university undergraduates. The impact of teachers' formative evaluation on student achievement was ranked third with an effect size of 0.9 in 138 learning activities influencing student achievement (Hattie, 2009).

Also, feedback, as an essential part of formative assessment, has been found to positively enhance students' learning (Hattie and Timperley, 2007; Hattie, 2009; Wisniewski et al., 2019). The large *prima facie* effect sizes found to raise the standards of learning laid a foundation for future evidence-based assessment policy reform. Formative assessment has gained an ever-widening array of attentions in various countries and regions.

In the past three decades, only four comprehensive reviews have reported the positive effect sizes of formative assessment on reading achievement which ranged from +0.22 to +0.7 (Fuchs and Fuchs, 1986; Black and Wiliam, 1998; Kingston and Nash, 2011; Klute et al., 2017). Yet in a literature review of 15 studies commissioned by the Australian Institute of Teaching and School Leadership (AITSL), the researchers stated that the impact of formative assessment on reading achievement was discouraging due to no effective tools could be identified and some programs integrated with technologies (Lane et al., 2019). The interpretations from the prior meta-analyses and literature review seems to be conflicting. Different school subjects require domain-specific effective formative assessment interventions (Wiliam, 2011). Arguably, whether how formative assessment enhances students' reading achievement remains unclear, this problem can be addressed by an updated and comprehensive meta-analysis. Lane et al. (2019) concerned that it could not distinguish the effect of formative assessment from digital technology on reading if they were mixed in a program. This issue can be settled by setting the involvement of digital technology in formative assessment practices as a moderator, which compares the programs with or without technology. Given the importance of formative assessment and the need for further statistical evidence on the reading subject (Clark, 2010; Van der Kleij et al., 2017; Black and Wiliam, 2018; Andrade et al., 2019), the purpose of this review, included literature in English and Chinese up to 2021, is to assess evidence from rigorous evaluations to determine the magnitude of experiment effects of formative assessment on students' reading performance and identify what features influenced its effectiveness. Noticeably, performing international comparison of formative assessment practices requires culture sensitivity (Shimojima and Arimoto, 2017). In this meta-analysis, we set three factors suggested by Cheung et al. (2021) to frame the features of formative assessment: substantive factors (student characteristics, grade level, type of intervention, digital technology, program duration, differentiated instruction), methodological factors (sample size, research design), and other factors (publication type, cultural setting).

Working definition of formative assessment

Since the term formative assessment has been used widely and diversely in the literature and because its classroom practice can vary within different educational settings, it is important to provide a working definition of the term to guide this review. Given the nebulous nature of formative assessment, a working definition of formative assessment is proposed based on the prior definitions in the past three decades. The essential statements of the 19 definitions (shown in **Supplementary material**) were compiled aligned with a succinct framework. Jönsson (2020) suggested that definition of formative assessment should include evaluative judgment (qualitative judgment) occurring in daily teacher-student interactions and a psychometric understanding of assessment depending on aggregating evidence of student learning collected by teachers. To follow this advice and identify potential studies, this review culls the more comprehensive descriptions under each element of the suggested definitions. Formative assessment in this review is broadly defined as

> *an active and intentional process with formal and informal classroom practices/activities harvesting evidence of students' learning progress by evaluative/qualitative judgment and a psychometric understanding of various assessments (what) during teaching and learning (when), in which teachers (who) continuously and systematically elicit, interpret and use evidence about students' learning and conceptual organization (how) to guide their pedagogical plans (why), and/or students (who) work with/without teachers or peers to adjust their current learning tactics (how) with an effort to improve their achievements and self-regulate their learning (why)* (Popham, 2008; Black and Wiliam, 2009; Chappius, 2009; Moss and Brookhart, 2009; Cizek et al., 2019).

The evaluative judgment refers to the daily teacher-student interactions eliciting evidence about learners' progress, for instance, feedback, discussions, presentations, and other students' artifacts. Psychometric assessments entail some quizzes, tests or indirect measurement that necessitates interpretation of outcomes (Jönsson, 2020). In this sense, some formative utilities of benchmark assessments and summative assessments (Wiliam, 2011) would be included if they met all the selection criteria in this review. Considering that formative assessments are classroom practices to identify students' learning gaps and improve their learning, participants can be teachers, students or their peers, as well as the integration of teachers and students. This review would clarify types of intervention to compare the effectiveness of different participants' engagement.

Alternative terminologies have emanated from different emphases to serve a common underlying formative purpose (Kingston and Nash, 2011). It is worth mentioning, the term assessment for learning (AfL) is often used interchangeably with formative assessment to emphasize the function of formative assessment to improve student learning (Heritage, 2010;

Bennett, 2011). The term AfL, first used by Harry Black (Black and Wiliam, 1986), was advocated by the Assessment Reform Group (ARG) in United Kingdom. Another term assessment as learning (AaL), was also phrased to signal the active role students play in the formative assessment process (Earl, 2012). Regarding assessment for, as, and of learning, each delineates the purpose for which the assessment was carried out. Differently, formative assessment and summative assessment are clarified by the functions they actually serve (Wiliam, 2011). Bennett (2011) suggested that it was not instructive to equate AfL with formative assessment and assessment of learning with summative assessment. However, a thorough exploration of the nuances between these two distinctions is beyond the scope of this paper. To include potentially qualified studies as broadly as possible, albeit labeled by alternative terms of assessment, the terms "formative evaluation," "feedback," "AfL," and "assessment as learning," were used as the key words in this review.

Previous reviews of formative assessment on reading achievement

From the literature review, eight major reviews on formative assessment were found in this area (Fuchs and Fuchs, 1986; Kluger and DeNisi, 1996; Black and Wiliam, 1998; Hattie, 2009; Kingston and Nash, 2011; Heitink et al., 2016; Klute et al., 2017; Sanchez et al., 2017). However, only four out of the eight comprehensive reviews encompassed the effect sizes of formative assessment in reading achievement (Fuchs and Fuchs, 1986; Black and Wiliam, 1998; Kingston and Nash, 2011; Klute et al., 2017). These four reviews indicated the positive effects of formative assessment on reading achievement, with effect sizes that ranged widely, from +0.22 to +0.7 (**Table 1**).

Fuchs and Fuchs' (1986) generated 96 effect sizes from 21 controlled studies, with an average weighted effect size of +0.70. The authors described that 8 of the 21 investigations focused solely on reading, 4 on reading and math, and 1 on reading, math and spelling, with no specific effect size calculated for reading. This meta-analysis focused upon special education as 83% of the 3,835 investigated subjects belonged to the special educational needs (SEN) population. It is inappropriate to generalize the findings to population of students at large. Secondly, 96 effect sizes generated from the 21 controlled studies were derived from analyses of divergent quality as the authors acknowledged. 69 effect sizes were of fair quality and 8 of poor quality which accounted for around 80% of all the effect sizes. Thus, the average effect size of 0.70 from the 21 studies examined was from research that was methodologically unsound (Dunn and Mulvenon, 2009). The limitation of specialized sample groups and the quality of the studies reviewed cast doubts on the validity of the large effect size.

Black and Wiliam's (1998) review of more than 250 articles related to formative assessment was a seminal piece to prove the positive effects of formative assessment on student achievement. The authors presented eight articles to support their conclusions pertaining to the efficacy of formative assessment without performing any quantitative meta-analysis techniques. The effect size that ranged from 0.40 to 0.70 concluded from their analysis was equivocal and inadequate to be applied in different contexts (Dunn and Mulvenon, 2009; Kingston and Nash, 2011). This review did not clarify the subject-based effect sizes. Hence, no substantiated effect sizes on reading achievement could be retrieved. Nevertheless, this 'configurative review' (Sandelowski et al., 2012) did encourage more widespread empirical research in the area of formative assessment (Black and Wiliam, 2018).

Kington and Nash (2011) screened out 13 of over 300 studies in grades K-12 to reexamine the effects between 0.40 and 0.70. Their moderator analyses indicated that the effect size of formative assessment in English language arts (ES = + 0.32) was larger than those in mathematics (ES = + 0.17) and science (ES = + 0.09). Briggs et al. (2012) commented that one of the marked flaws that threatened Kingston and Nash's conclusion was their study retrieval and selection approach. It might explain the paucity of Kingston and Nash's research base (Kingston and Nash, 2012). This problem could be solved by referring to some subset of the studies suggested by Black and Wiliam (1998).

The latest meta-analysis involving reading achievement was conducted by the US Department of Education

TABLE 1 Summary of major meta-analysis on effects of formative assessment on reading achievement.

Authors	Years covered	Types of publication	Subjects covered	Grades	Number of studies (reading)	Effect size
Fuchs and Fuchs	1971–1984	Journal article	Reading and a variety of subjects	Elementary, middle/high	13	+ 0.7 (for all subjects)
Black and Wiliam	Unspecified-1998	Journal article	Reading and a variety of subjects	5 years old to university undergraduates	Unspecified	+ 0.4–0.7 (for all subjects)
Kingston and Nash	1988–2011	Journal article	Reading and a variety of subjects	elementary, middle/high	12	+0.32
Klute et al.	1988–2014	Report	Reading and a variety of subjects	Elementary	9	+0.22

(Klute et al., 2017). The research team identified 23 rigorous studies on reading, math and writing in elementary level to demonstrate the positive effects of formative assessment interventions on student outcomes from 1988 to 2014. Though it was stated that the review identified studies published between 1988 and 2014, their finalized list for reading only updated to 2007. Of the 23 studies in various subject areas, nine focused on reading with an average effect size of +0.22. Interestingly, their report revealed that other-directed formative assessment was more effective (ES = + 0.41) than student-directed formative assessment (ES = −0.15). Other-directed formative assessment encompassed educators or computer software programs, whilst, student-directed formative assessment referred to self-assessment, self-regulation and peer assessment. The novice categories of formative assessment provided new insights into the moderator analyses. This report was rigorous with stringent controls on selection criteria. However, it solely covered the elementary level and restricted the geographical research location in Anglophone countries.

Moderator variables

To warrant the quality of a meta-analysis, a rationale for the coding schema should be provided (Pigott and Polanin, 2019). Three factors suggested by Cheung et al. (2021) were set to frame the features of formative assessment.

Methodological factors

Methodological factors describe research design and sample size. One possible factor that might cause variance is the research design of divergent studies (Abrami and Bernard, 2006). Two groups of research designs were identified in this review: RCT (Randomized Control Trial) and QED (Quasi-experimental design). Particularly of concern is that cluster (school-level, classroom-level, and teacher-level) randomized control trial with student-level outcome measure would be coded as quasi-experimental studies. Another potential source of variation may lie in the sample size which was reported to be negatively correlated with effect sizes in studies of reading program (Slavin and Smith, 2009). Following the tradition in a previous meta-analysis (Cheung and Slavin, 2016), this review coded studies with 250 students or less as small sample, the others were taken as large sample.

Substantive factors

Substantive factors depict the background of a study such as population, context and duration. Six program features identified from some seminal meta-analyses on reading and formative assessment (Klute et al., 2017) were included in this review.

Student characteristics, grade level and program duration

Students in the included studies were categorized into at-risk or mainstream students. At-risk students referred to students who had reading difficulties or of low performance in common classrooms, others were coded as mainstream students. Grade level was divided into kindergarten, elementary and middle/high levels. Program duration set 1 year as a threshold to classify long and short programs. Programs that lasted for less than 1 year were coded as short; the rest were long.

Differentiated instruction

Formative assessment is a "gap minder" (Roskos and Neuman, 2012) enabling teachers and students to identify the gap between where students are and where they need to go in their reading development (Wiliam and Thompson, 2007). Consequently, teachers can stay alert to these gaps and differentiate their instruction to various students. Differentiated instruction is taken as an optional component in the formative assessment practice. In our review, there were several teacher's practices that were coded as "without interventions." For instance, teachers who kept track of students' learning gaps without changing their teaching plan, or who just monitored the interim/benchmark assessment results without further action on differentiating or individualizing their teaching to different students aligned with the data from assessment.

Type of intervention

Tethered to main sources of formative assessment practices (Andrade et al., 2019) in two latest integrated formative assessment meta-analyses (Klute et al., 2017; Lee et al., 2020), type of intervention was coded as teacher-directed, student-directed or integrated (teacher and student assessment). Specifically, teacher-directed assessment referred to teachers who provided feedback, interim/benchmark assessment or other resources to gauge students' learning, be it computer-based or paper-based, and/or conducted individualizing or differentiating instruction to students' classroom learning. Student-directed assessment mainly manifested in the forms of peer- or self- assessment, and young learners' meaning-focused group reading activity (Connor et al., 2009). Integrated practices involved both teacher and student in the assessment process.

Digital technology

Various digital technologies have been explored and applied in K-12 formative assessment practice in the 21st century (Spector et al., 2016). A newly published article suggested digital technology could be conducive to reading for young children not but for older children (See et al., 2021). This review will cross check this result by including more rigorous studies on reading from various cultural contexts.

Other factors

Other factors are the external variables that might influence the variance of effect sizes. We included publication type and cultural settings which were never assessed in previous meta-analyses on formative assessment.

Publication type

Validity of the results from a meta-analysis is often reported to be threatened by the presence of publication bias. To put it succinctly, publication bias refers to studies with large or statistically significant effects compared to studies with small or null effects being prone to publication. This meta-analysis included both published and unpublished literature (technical reports, dissertations and conference reports).

Cultural settings

Formative assessment was introduced and developed in Anglophone culture represented by United Kingdom and United States. In light of previous reviewed policies in Asia-Pacific regions, it is safe to assume that formative assessment has been introduced and implemented in Asia, especially in countries or regions heavily influenced by Confucian-heritage culture (CHC) which was heavily influenced by exam-orientation (Biggs, 1998). Teachers from CHC culture are often burdened with high-stake test pressure. It might be more demanding for teachers in CHC classrooms to believe that formative assessment is to facilitate learning rather than accredit it (Crossouard and Pryor, 2012). This review, as a first of its kind, attempted to compare the interventions in Anglophone and Confucian-heritage culture. Studies conducted in Anglophone culture are from Barbados (1), Germany (3), Spain (1), Sweden (1), United Kingdom (1) and United States (30), while studies in CHC settings are from Hong Kong (4), South Korea (1) and Taiwan (3). To note, although Germany, Spain and Sweden are not English-speaking countries, we still categorized them into Anglophone culture in stark contrast to the exam-driven CHC. Surprisingly, few studies from Mainland China could be located to meet our inclusion criteria. The reasons for this were threefold. First, some marginally qualified studies were carried out by only one teacher in two classes so the teacher effect could not be evened out. Second, some studies did not report results of reading achievement as they were not statistically significant, which was explicitly stated by the authors. Besides, the majority of formative assessment projects in China were based in higher education. The culling process implied some new directions for future research and reviews.

Methodological and other factors are mainly the extrinsic factors that can be applied to meta-analyses in other research fields. The substantive factors include intrinsic features that are commonly seen in formative assessment activities. These moderators provide a comparatively holistic set of features that might influence the effect of formative assessment on students' reading achievement.

Rationale for present review

Due to the paucity of studies, lack of stringent selection criteria and limitation of samples, the aforementioned comprehensive reviews encouraged more rigorous studies to be investigated to reveal the latest effect size for the subject–reading. The existing subject-based reviews have covered mathematics (Gersten et al., 2009; Burns et al., 2010; Wang et al., 2016; Kingston and Broaddus, 2017), writing (Graham et al., 2015; Miller et al., 2018), and science (Hartmeyer et al., 2018), but not reading.

To provide a more comprehensive understanding of the effectiveness of formative assessment for enhancing reading achievement, this study attempted to elicit exemplary formative assessment practices by applying rigorous, consistent inclusion criteria to identify high-quality studies. Our review, in an effort to sketch a comprehensive picture of the effects of formative assessment on reading, statistically consolidated the effect sizes of qualified studies in terms of methodological and substantive features. The present study attempts to address two research questions:

(1) What is the effect size of formative assessment on K-12 reading programs?
(2) What study and research features moderate the effects of formative assessment interventions on student reading achievement?

Method

The present review employed meta-analytic techniques suggested by Glass et al. (1981) and Lipsey and Wilson (2001). Comprehensive Meta-analysis Software Version 3.0

(Borenstein et al., 2013) was adopted to compute effect sizes and to carry out various meta-analytical tests. The following steps were taken during meta-analytic procedures: (1) scan potential studies for inclusion using preset criteria; (2) Locate all possible studies; (3) code all qualified studies based on their methodological and substantive features; (4) calculate effect sizes for all selected studies for additional combined analyses; (5) perform comprehensive statistical analyses encompassing both average effects and the relationships between effects and study features.

Criteria for inclusion

To be included in this review, the following inclusion criteria were preset.

(1) Studies that examined the effects of formative assessment or AfL on students' reading outcomes.
(2) Studies can be directed by a single party, be it teacher or student (peer- or self- assessment), or by collaboration of teachers and students.
(3) Classroom practices align with the definition of formative assessment in this review.
(4) The studies involved students in kindergarten, elementary and secondary education.
(5) Reading programs included English as a native or a foreign language in their reading courses, or reading courses in students' mother tongue.
(6) Studies could have taken place in any country or region, but the report had to be available in English or Chinese.
(7) Treatment/experiment group(s) embedded with formative assessment activities was/were compared with control group(s) using standard/traditional methods (aka business-as-usual groups).
(8) Pretest data had to be provided (What Works Clearinghouse, 2020), unless studies used random assignment of at least 30 units (individuals, classes, or schools) and no indications of initial inequality were reported, which were set aligned with ESSA (Every Student Succeeds Act) evidence standards (ESSA, 2015). Studies with pretest differences of more than 50% of a standard deviation were excluded because, large pretest differences could not be adequately managed as underlying distributions may be fundamentally different even with analyses of covariance (Shadish et al., 2002).
(9) Two teachers (each in one classroom) should be involved in each treatment group to even out the teacher effect in treatment effects. Of note, some studies which only examined the students' roles in formative assessment with only one teacher in each group were included.
(10) Studies interventions had to be replicable in realistic school settings (i.e., in usual classroom setting, students with their usual teacher, controlled experiments). Studies equipping experimental groups with extraordinary amounts of aids (e.g., additional staff to ensure proper implementation) where the Hawthorn effect would be generated were excluded.

Literature search procedures

All qualified studies from the current review come from three main sources. (1) Previous reviews; Analyzed studies from the previous reviews were further examined. (2) Electronic searches; A comprehensive literature search of articles written up to 2021 was conducted to screen out qualifying studies. Electronic searches were carried out through educational databases (e.g., ERIC, EBSCO, JSTOR, Psych INFO, ScienceDirect, Scopus, Dissertation Abstracts, ProQuest, WorldCat, CNKI), web-based repositories (e.g., Google, Google Scholar), and gray literature databases (e.g., OpenGrey, OpenDOAR). The key words for the search included 'formative assessment,' 'formative evaluation,' 'feedback,' 'assessment for learning,' 'assessment as learning,' 'curriculum-based assessment,' 'differentiated instruction,' 'portfolio assessment,' 'performance assessment,' 'process assessment,' 'progress monitoring,' 'response to intervention' (Gersten et al., 2020), as well as the subset forms under the formative assessment umbrella suggested by Klute et al. (2017) (e.g., self-monitoring, self-assessment, self-direct, peer assessment). (3) Relevant contextualized assessments. The following contextualized assessment projects and systems were included in the searching procedure: learning-oriented assessment (Carless, 2007), A2i (Assessment to instruction) (Connor et al., 2007), SLOA (Self-directed Learning Oriented Assessment) (Mok, 2012), LPA (learning progress assessment) (Förster and Souvignier, 2014), DIALANG (Diagnostic Language Assessment) (Zhang and Thompson, 2004) and CoDiAs (Cognitive Diagnostic Assessment System) (Leighton and Gierl, 2007).

Articles found in the databases were primarily screened by the lead author at the title and abstract level if the purpose of the study matched the independent (formative assessment intervention program) and dependent (reading outcome) variables guiding this meta-analysis. Records identified through database searching numbered 8,048. Additionally, 21 studies were found from previous meta-analysis (Kingston and Nash, 2011; Klute et al., 2017) and a literature review (Lane et al., 2019). Seven studies were included from two formative assessment projects: A2i (Assessment to instruction) (Connor et al., 2007, 2011, 2013; Al Otaiba et al., 2011) and LPA (learning progress assessment) (Förster and Souvignier, 2014, 2015; Förster et al., 2018; Peters et al., 2021). The screening of titles resulted in the retention of 8076 articles at the title and abstract levels that were further examined for eligibility

and inclusion in this study. In the first round of screening, we mainly parsed out studies that were not experiments and irrelevant to reading. Then, 113 articles were retained for full-text examination. By applying the inclusion criteria in this review, full-text articles were excluded for the following reasons: without a control group (e.g., Topping and Fisher, 2003), no pre-test (e.g., Cain, 2015), with over 0.50 SD in pre-test (e.g., Hall et al., 2014), only focusing on spelling or vocabulary (e.g., Faber and Visscher, 2018), without reading achievement (e.g., Marcotte and Hintze, 2009), with sample size less than 30 participants (e.g., Chen et al., 2021), students in special education (e.g., Fuchs et al., 1992), and at tertiary level (e.g., Palmer and Devitt, 2014). The numbers in each category can be seen in **Figure 1**.

Coding scheme

To assess the relationship between effects and studies' methodological and substantive features, studies were coded. Methodological features referred to research design and sample size. Substantive features entailed types of publication, grade levels, types of intervention, program duration, implementation, cultural settings, year of publication, students' characteristics, online technology. The study features were categorized as follows:

(1) Students' characteristics: Mainstream or at-risk students.
(2) Grade levels: Kindergarten, Elementary (Grade 1–6), Middle/High (7–12).
(3) Types of intervention: teacher-directed (feedback to teacher, response to intervention), student-directed (peer- or self- assessment), integration of teacher and student assessment.
(4) Digital technology: with or without.
(5) Program duration: short (less than 1 year), long (≥1 year).
(6) Differentiated instruction: with or without, and not applicable for those studies only involved peer- and self- assessment that did not describe teachers' instruction adjustment.

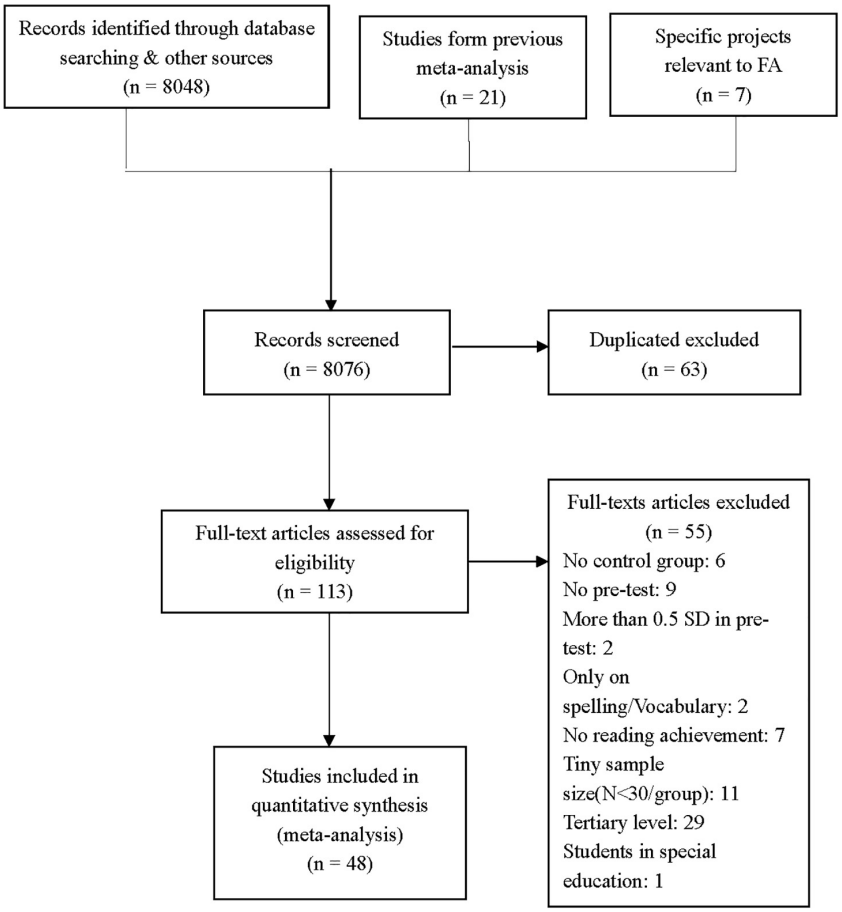

FIGURE 1
PRIMA flow chart (Moher et al., 2009).

(7) Research design: QED (quasi-experimental design) or RCT (randomized control trial).
(8) Sample size: Small ($N \leq 250$ students) or large ($N > 250$).
(9) Publication type: published or unpublished.
(10) Cultural settings: Anglophone culture (Australia, Canada, Ireland, New Zealand, United Kingdom, and United States), CHC (Mainland China, Hong Kong SAR, Taiwan, Singapore, Japan, Korea).

The coding of all characteristics was processed by two researchers independently. Inter-rater reliability was calculated by selecting 20 percent of randomly selected studies. Reliability was 87.21 percent. Disagreements were discussed and rectified in light of the definition proposed. All features of formative assessment are presented in **Table 2** and descriptive data of qualified studies can be found in the **Supplementary material**.

Effect size calculations and statistical analyses

In general, effect sizes were calculated as the difference between experimental and control student posttests after adjusting for pretests and other covariates, divided by the unadjusted posttest pooled standard deviation. When unadjusted pooled standard deviation was not available, as when the only standard deviation presented was already adjusted for covariates or when solely gain score standard deviations were available, procedures proposed by Sedlmeier and Gigerenzer (1989) and Lipsey and Wilson (2001) were used to estimate effect sizes. Provided that pretest and posttest means and standard deviations were presented but adjusted means were not, effect sizes for pretests were subtracted from effect sizes for posttests. An overall average effect size was produced for each study as these outcome measures were not independent. Comprehensive Meta-Analysis software was employed to carry out all statistical analyses, such as Q statistics and overall effect sizes.

Results

Overall effects

A total of 48 qualifying studies was included in the final analysis with a total sample size of 116, 051 K-12 students: 9 kindergarten studies (N = 2,040), 28 elementary studies (N = 107,919), 11 middle/high studies (N = 6,092). The overall effect sizes were calculated in fixed and random effect models. The large Q value (Q = 313.56, df = 47, $p < 0.000$) indicated that the distribution of effect sizes in this scope of studies is highly heterogeneous. In other words, the variance of study effect seizes is larger than can be explained by simple sampling error. Thus, a random effects model was adopted (DerSimonian and Laird, 1986; Borenstein et al., 2009; Schmidt et al., 2009). As shown in **Table 3**, the overall weighted effect size is + 0.18 with confident interval between 0.14 and 0.22. In an attempt to interpret this variance, key methodological features (sample size, research design), substantive features (student characteristics, program duration, types of intervention, grade level, digital technology involvement) and extrinsic features (publication type, culture) were used to model some of the variances. An overview of the effect sizes can be seen in **Figure 2** that provides a graphical representation of the estimated results of all included studies.

Subgroup analysis

The heterogeneity in the overall effect calculation implies that the large differences between these 48 included studies might be related to researchers' choice of methodology, samples'

TABLE 2 Coding scheme features.

Categories of features	Features of FA	Variables
Substantive factors	Student characteristic	(1) Mainstream students
		(2) At-risk students
	Grade level	(1) Kindergarten
		(2) Elementary (1–6)
		(3) Middle/High (7–12)
	Type of intervention	(1) Teacher-directed
		(2) Student-directed (self-assessment)
		(3) Integrated
	Digital technology	(1) Yes
		(2) No
	Program duration	(1) Short (<1 year)
		(2) Long (≥1 year)
	Differentiated instruction	(1) Yes
		(2) No
		(3) Not applicable (student-directed assessment only)
Methodological factors	Research design	(1) RCT (randomized controlled trial)
		(2) QED (quasi-experimental design)
	Sample size	(1) Large ($N > 250$)
		(2) Small ($N \leq 250$)
Other factors	Publication type	(1) Published
		(2) Unpublished
	Cultural setting	(1) Anglophone culture
		(2) Confucian-heritage culture (CHC)

TABLE 3 Overall effect size.

	k	ES	SE	95% confidence interval		Test of mean		Test of heterogeneity in effect size		
				Lower limit	Upper limit	*Z*-value	*p*-value	*Q*-value	df (Q)	*p*-value
(1) Fixed	48	0.07	0.01	0.05	0.08	10.78	0.00	313.56	47	0.000
(2) Random	48	0.19	0.02	0.15	0.23	8.65	0.00			

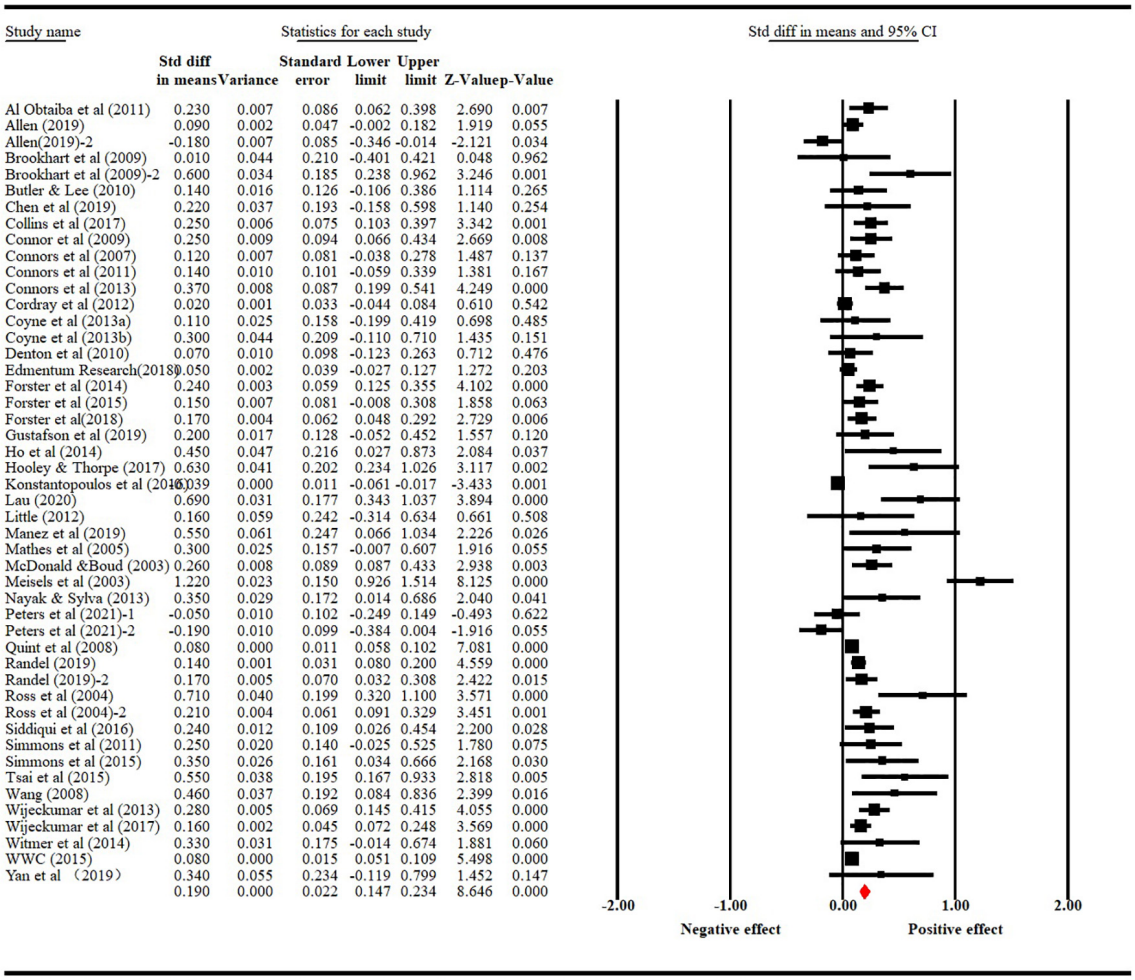

FIGURE 2
Forest plot of formative assessment effect size on K-12 students' reading achievement.

substantive features and other factors. Before our meta-regression analysis, we first estimated the comparisons in subgroups as shown in Table 4. Variation yielded significant differences in effect sizes that were from seven moderators: grade level, type of intervention, program duration, differentiated instruction, sample size, publication type and cultural setting. The effect sizes comparisons of subcategories in rest three moderators, student characteristics, digital technology and research design were non-significant.

The reason we performed subgroup analysis was to provide basic descriptive data of the 'constructed' features in formative assessment. Six substantive features have been investigated in previous reviews (Kingston and Nash, 2011; Klute et al., 2017; Lee et al., 2020). In this review, we added methodological factors (research design and sample size) and other factors (publication types and cultural settings) to examine how effect size of formative assessment on reading would vary in these categories. Hence, after we added new moderators into the

TABLE 4 Subgroup analysis results.

	Features	Studies included (*k*)	Effect size	*p*
Student characteristic	Mainstream students	37	0.18	0.156
	At-risk students	11	0.27	
Grade level*	Kindergarten	9	0.28	0.038
	Elementary (1–6)	28	0.16	
	Middle/High (7–12)	11	0.27	
Type of intervention*	Integrated	19	0.20	0.011
	Teacher-directed	19	0.12	
	Student-directed	10	0.31	
Digital technology*	No	19	0.32	0.011
	Yes	29	0.15	
Duration	Long	25	0.16	0.110
	Short	23	0.21	
Differentiated instruction***	Yes	29	0.24	0.000
	No	9	0.050.36	
	n.a.	10		
Research design	QED	36	0.18	0.382
	RCT	12	0.22	
Sample size***	Large	30	0.13	0.000
	Small	18	0.45	
Publication type***	Published	38	0.26	0.000
	Unpublished	10	0.09	
Cultural setting**	Anglophone	40	0.17	0.005
	Confucian-heritage culture	8	0.38	

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

meta-regression model, the subgroup analysis data would help us decipher why some of the results were contrary to previous findings.

Meta-regression

To address the second research question, we regressed all the moderator variables in a model presented in Table 5 to describe the predicted different standard deviation (coefficient) of comparing the categories in each moderator after controlling for other features of formative assessment interventions. It is assumed that the effects from meta-regression are more confidently reliable than results from subgroup analysis by taking account of the iterative influences from different moderators. In our proposed model, only three pairs of moderator categories comparison were significant.

First, the effect size from different *sample sizes* varied substantially. As aforementioned, we set $N = 250$ as a cut-off point. Clearly, small sample size studies yielded a significantly larger effect size ($d = 0.33$, $p < 0.001$) than large sample studies.

Next, three *types of intervention* were examined, namely, teacher-directed, student-directed and integration of teacher and student assessment. Results indicated that, when other features of formative assessment were controlled, formative assessment only engaged by teachers had a significantly smaller effect size than integrating teacher and students' assessment in the intervention for reading ($d = -0.12$, $p < 0.001$), whereas, student-directed assessment (self-assessment) showed no significant difference ($d = 0.03$, $p = 0.769$).

Third, in regard to *differentiated instruction*, we coded formative assessment in reading with or without differentiated instruction. Some interventions that only involved student

TABLE 5 Results of meta-regression.

Random effects	Coefficient	SE	*p*
Intercept	0.07	0.08	0.409
Sample size (Small) ***	0.33	0.06	0.000
Type of intervention (Teacher-directed) ***	−0.12	0.03	0.000
Type of intervention (Student-directed)	0.03	0.09	0.769
Differentiated instruction (Yes) ***	0.13	0.03	0.000
Differentiated instruction (n.a.)	0.17	0.11	0.124
Research design (RCT)	−0.04	0.06	0.465
Digital technology (Yes)	0.001	0.04	0.978
Grade level (Elementary)	0.02	0.07	0.790
Grade level (Middle/High)	−0.03	0.07	0.696
Student characteristics (at-risk students)	−0.001	0.08	0.983
Duration (Short)	0.02	0.04	0.576
Cultural settings (CHC)	−0.10	0.09	0.297
Publication type (unpublished)	0.01	0.03	0.701
Q	99.73		0.000
Df	13		
R^2 analog	0.95		

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

directed assessment without teachers' instructional adjustment were coded as not applicable (n.a.). Results, as we hypothesized, favored teachers who used differentiated instruction during or after their formative assessment on students' reading. If a teacher adopted differentiated or individualized instruction during or after formative assessment, students' reading achievement would be significantly higher than those of their peers taught by a teacher only applied formative assessment (d = 0.13, p < 0.001). When formative assessment was directed by the student, the effect size on reading achievement was larger than formative assessment with teachers' differentiated instruction, albeit not significantly (d = 0.17, p = 0.124).

Apart from the three pairs of contrast, the rest of the moderator variables comparisons were non-significant, although some showed significant results in subgroup analysis.

Given that *research design* might influence the effect size, we categorized all studies into randomized controlled studies (RCT) and quasi-experiments (QED). Results from the regression model indicated that effect sizes generated from RCT were smaller than QED design, but not significant (d = –0.04, p = 0.465).

Digital technology involvement was examined. Surprisingly, students' reading achievement was not influenced significantly by formative assessment with digital technology (d = 0.001, p = 0.978).

Formative assessment in reading classrooms seemed to exert a similar impact on students of different *grade level*s. The effect sizes in elementary level were slightly larger than those for kindergarten studies (d = 0.02, p = 0.790). Effect sizes for middle/high school studies were also slightly smaller than those for kindergarten studies (d = –0.03, p = 0.696).

Student characteristics were coded into mainstream students and at-risk students. The estimated effect size of formative assessment on mainstream students was slightly higher than that on at-risk students, albeit not significant (d = 0.001, p = 0.982).

With respect to *program duration*, studies that lasted less than 1 year were coded as short programs, others were coded as long ones. Programs in one-year or longer showed smaller effect size than short-term ones, but the difference was non-significant (d = –0.02, p = 0.576).

Different from the result of the subgroup analysis result regarding *cultural setting*, the effect size of formative assessment in CHC appeared to be smaller than those in Anglophone culture, though not significant (d = –0.10, p = 0.297).

Results of publication type revealed that no significant differences were found between published and unpublished articles (p = 0.701), indicating that no publication bias existed in this review.

The pseudo R^2 value in this meta-regression model estimated the moderators accounted for 95% of heterogeneity. The predictive power of this value is reliable as the number of studies (k) in this review exceeded the minimum number of 40 as suggested by López-López et al. (2014).

Discussion

Overall effect size

The findings of this review indicate that formative assessment produce a positive effect (ES = + 0.18) on reading achievement. The magnitude could be interpreted as a *small* effect aligned with the oft-cited indication of small (d = 0.2), medium (d = 0.5), and large (d = 0.8) effect size (Cohen, 1988). However, Kraft (2020), taking study features, program costs and scalability into account, proposed a new benchmark frame for effect size from causal studies of pre K-12 education intervention, namely small (d < 0.05), medium (0.05 to <0.20), and large (≥0.20). Accordingly, the overall aggregated effect size in this review could be taken as a medium effect size. Compared with effect sizes (from +0.22 to +0.70) from previous meta-analyses, the weighted average effect size reported in this review was the smallest one. Two potential factors may explain this. First, some early reviews set comparatively looser criteria for inclusion, which often inflates effect size estimates. Pertaining to our set of stricter inclusion criteria, five studies in Klute et al. (2017) review with less than 30 participants in each group (Fuchs et al., 1989; McCurdy and Shapiro, 1992; Johnson et al., 1997; Iannuccilli, 2003; Martens et al., 2007) were ruled out in the present review. Second, 35 of our selected studies were conducted after 2010 whereas a latest previous meta-analysis (Klute et al., 2017) only included studies till 2007. As publication bias might be mitigated over time (Guan and Vandekerckhove, 2016), more insignificant or even negative findings were reported. In this review, two large scale studies (Konstantopoulos et al., 2016; Allen, 2019) involving over 35,000 student reported negative effects of formative assessment on reading achievement.

The effects of moderators

The meta-regression results indicate that sample size, differentiated instruction and type of intervention suffice to account for the heterogeneity of the effect sizes. Additionally, we intend to discuss some implications from the results of cultural settings, digital technological and publication bias.

Sample size

Prior research indicated that studies with small sample sizes tend to yield much larger effect size than do large ones (Liao, 1999; Cheung and Slavin, 2016). In this review, sample size was a crucial variable that might influence the effect size of formative assessment on reading achievement. Two explanations could be put forward for this result. First, intuitively, small-scale studies are more likely able to be implemented with high fidelity. Teachers might find it easier to give more support

for students and monitor their progress. Researchers are more likely to purposefully recruit motivated teachers and schools. In this sense, they tend to produce larger effect size than large-scale studies. Next, researchers using small samples would be apt to more design self-developed outcome measures (Wang, 2008; Tsai et al., 2015; Lau, 2020; Yan et al., 2020)., which might be more sensitive to treatments than standardized studies (Cheung and Slavin, 2016).

Differentiated instruction

One of the key findings in our review was the positive effects of differentiated instruction during or after formative assessment on reading achievement for K-12 students. This significant result is in accord with the findings from an influential U.S. data-driven reform model on state assessment program. Slavin et al. (2013) found that, for fifth-grade reading, those schools and teachers adjusting reading instruction produced educationally important gains in achievement, while others did not if they merely understood students' data without further action on instructional adjustment. Formative assessment was analogous to taking a patient's temperature, while differentiated instruction was analogous to providing a treatment (Slavin et al., 2013).

In a study included in our review with a comparatively promising large effect size ($d = +0.63$) on an early literacy program designed for students at-risk, the researchers concluded that "if one practices formative assessment seriously, one will necessarily end up differentiating instruction" (Brookhart et al., 2010, p. 50). In a recent review on formative assessment (Lee et al., 2020), the research team coded a similar moderator "instructional adjustment" and revealed no significant contrast between their four moderator variables: no adjustment, planned adjustment, unplanned adjustment and mixed. We assumed that the effects might be ameliorated if too many variables were coded which led to the insufficient numbers in each category. Additionally, in our own model, primarily we added professional development as a moderator. However, this moderator was highly correlated with "differentiated instruction." Meta-regression could not be computed due to the collinearity. It is worth mentioned, in our qualified studies, 94% (34/36) of the interventions embedded with differentiated instruction were coupled with professional development for teachers. The evidence in turn implied that professional development is vital in fostering high fidelity of implementing formative assessment on reading programs.

Type of intervention

The types of intervention result indicated that an integration of teacher-directed and student-directed would be more effective than formative assessment in reading program directed by teacher or student alone in K-12 settings. In a previous meta-analysis, the research teamed concluded that other-directed formative assessment that encompassed educators or computer software programs was more effective than student-directed formative assessment (Klute et al., 2017). They included nine studies in their review, six of which were designed for students with special education needs. These participants might be less capable of making self- or peer-directed formative assessment. In the present review, a more holistic picture was depicted for general population was obtained advocating an integrated usage of teacher-directed and student-directed assessment.

The results of our review suggested that integrating teacher and student in formative assessment might be more effective than teacher- or student- directed assessment to enhance students' reading achievement. We attempted to explained this based on linguistic theory (Kintsch and Van Dijk, 1978). Some production-based subject like writing might be more effective when the formative assessment was student-centered (Black and Wiliam, 1998), but reading is a comprehension-based subject that requires explicit instruction necessitated by teachers' guidance (McLaughlin, 2012). Also, feedback messages require students' active construction on deciphering with the help of teachers (Ivanic et al., 2000; Higgins et al., 2001). But we were given a caveat that it was not a "one size fits for all" suggestion from our screening on studies in Anglophone and Confucian-heritage cultures.

Cultural setting

The subgroup analysis comparison result of interventions in two cultures were significant. Studies conducted in Confucian heritage culture yielded ostensibly much larger sample sizes than those in Anglophone culture. Nevertheless, the non-significant data in meta-regression indicated that it was influenced by other variables. By drawing on the data and the evidence we collected, we found it hard not to associate the impact with sample size. All the qualified studies in CHC were of small sample size. Sample size was reported to be one of the significant moderators which contributed to the variance of effect sizes in this review. After controlling for other moderators, no significant differences were found between the interventions in these two cultures.

Though it was provisional to conclude that there was no difference between the studies in Confucian-heritage culture and Anglophone culture, our screening process and descriptive data in subgroup analysis might render us some hints for the interpretation of the results.

Only eight qualified studies were set in CHC, while 38 for Anglophone culture. The limited number of experimental studies from CHC settings might be associated with the barriers of formative assessment intervention in CHC. Teachers from

CHC (Mainland China, Hong Kong SAR, Taiwan, Singapore, Japan, Korea) are often challenged by large class sizes (Hu, 2002) and high-stake test pressure (Berry, 2011), which gives rise to teachers' psychological burden on assessment (Chen et al., 2013). These sociocultural factors drastically hinder the translation (local adaptation of an educational policy) (Steiner-Khamsi, 2014) of formative assessment. When a school advocates formative assessment for teachers without appropriate professional development, they take it as a "villain of workload" (Black, 2015). Teachers in a test-driven culture would inevitably take formative assessment as a "test" instead of instruction.

Next, particularly of concern is the hint we obtained from the promising results of our included studies in CHC. Researchers in CHC contexts have started to explore alternative ways to implement formative assessment. Six out of eight studies in CHC in our review were self-assessment (Wang, 2008; Butler and Lee, 2010; Tsai et al., 2015; Chen et al., 2017; Lau, 2020; Yan et al., 2020). This renders us a new direction that self- assessment might be alternatives for reading teachers to implement formative assessment as part of their teaching in CHC classrooms. But we are far from confident to conclude it's the most effective way based upon the data we reported in this review.

Digital technology

Previous meta-analysis findings revealed that mobile devices (Sung et al., 2016) and educational technology (Slavin, 2013) do not exert significant differences on students' academic achievement, and digitally delivered formative assessment is only conducive to reading for young school-age children but not for older children (See et al., 2021). In line with those reviews, our findings also indicated that formative assessment with digital technology does not significantly influence students' reading achievement compared with traditional paper-pen intervention. The findings caution that digital technology is not the kernel of formative assessment. Nevertheless, our findings still advocate technology-enhanced formative assessment as it can provide an evidence-based platform to scaffold students' learning by generating and deploying formative feedback. From the methodological perspective, computer-based formative assessment systems are generally more accessible for teachers and students than traditional methods (Tomasik et al., 2018). Of note, lessons can be drawn from the undesirable effect sizes of those digital formative assessment programs: (1) A digital formative assessment program can be promisingly effective when teachers in intervention group differentiate their instructional practices based on the evidence feedbacked by the digital program. Researchers from the benchmark or interim assessment with small (Cordray et al., 2013) or even negative effect sizes (Konstantopoulos et al., 2016) reflected that teachers might need further support to adjust their teaching as their classroom schedules were quite crowded. (2) Professional development and training for teachers participating in the digital formative assessment are irreplaceable prerequisites for the quality of practice. Support for teachers to understand the concept and provision of technical assistance are essential for their instructional change (Connor et al., 2009; Kennedy, 2009).

Publication bias

To mitigate the threat of publication bias, we included 10 unpublished studies in this review. Traditional methods to assess publication bias included a visual inspection of symmetric dispersion of a funnel plot (Sterne et al., 2005), "fail safe N" statistics (Orwin, 1983), trim-and fill method (Duval and Tweedie, 2000) and setting publication bias as a moderator to test the differences in mean effect sizes between published and unpublished studies (Polanin and Pigott, 2015). As the latter method was comparatively straightforward and more objective than eyeballing evaluation, we took publication bias as a moderator in meta-regression to compare the mean effect sizes of published and unpublished studies by controlling other factors. No significant difference was found between the two groups of studies. We believe that publication bias is not a concern for the current meta-analysis.

Conclusion

This review has revealed that, without publication bias, formative assessment is making a positive and modest difference in enhancing students' reading achievement in diverse settings. The average weighted effect over all included studies was 0.19. The exact size a researcher finds may deviate considerably depending on the sample size, teachers' differentiated instruction and type of intervention. Studies involving a large sample size with over 250 students led to low and attenuated estimate of formative assessment. The implementation of teachers' differentiated instruction is linked to much stronger effects than intervention without differentiated instruction. Also, our results suggested that collaboration of teachers and students in formative assessment would be more effective than formative assessment merely initiated by teachers. Findings suggest that teachers are strongly encouraged to adjust their reading instruction in terms of content, process and product catering to student diversity (Tomlinson, 2001) during the formative assessment in the cooperation with students themselves. Studies with differentiated instruction coupled with teacher's professional development has a positive and modest effect on reading outcome. To enhance students' reading achievement and upskill teachers, future studies designs should focus more on

effective components that facilitate differentiated instruction and professional development.

This meta-analysis contributes to the existing understanding about formative assessment in K-12 reading program in three significant ways. First, it systematically records the critical components of formative assessment pertaining to reading program for frontline teachers to refer to by catering for learner diversity (Snow, 1986). Second, it affords a new cross-cultural perspective by comparing western and eastern formative assessment practices for school administrator and policy makers to tailor effective programs in their unique cultural contexts. Lastly, it substantiates the discipline-specific characteristics in reading to conceptualize formative assessment for K-12 reading program (Bennett, 2011), which is pivotal to a next-generation definition of domain-dependent formative assessment (Cizek et al., 2019).

It is vital to mention several limitations of this review merely focusing on the quantitative measurement of reading achievement. Evidence-based education advocates the insightful and irreplaceable findings from qualitative research (Slavin and Cheung, 2019). There is much to learn from non-experimental studies that can interpret the effects of formative assessment on students' reading. Next, this review centered on a standardized test of reading achievement. However, other outcomes maybe of great value to policymakers and practitioners. Third, student-directed assessment is often referred to peer- or self-assessment. Third, the qualified studies in this meta-analysis only include self-assessment. We are aware the value of peer-assessment and strongly suggest future review could locate more qualified studies concerning this type of assessment. Lastly, the culture settings in this study merely include Anglophone or CHC as we could not locate acceptable studies from other cultures temporarily. Studies setting at all cultures were equally important and should be included if possible. Further studies could explore research from other cultures.

Educational borrowing from other countries is not a simple case of duplicating the successful tales, inasmuch as extrapolation and recontextualization of educational interventions are embedded with cultural and historical stories (Luke et al., 2013). Our subgroup analysis indicated cultural settings might be a potential moderator. As a wealth of large-scale formative assessment initiatives have been advanced in classrooms heavily influenced by CHC, synthesized effect sizes in CHC settings are encouraged to be reported to ensure the continuity of formative assessment with cultural script (Stigler and Hiebert, 1998, 2009). Future reviews can apply narrative synthesis methods to explore the factors that advance or hinder the development of formative assessment on reading in CHC. Considering the complicated implementation of formative assessment on reading (Lane et al., 2019), teachers in CHC classrooms are suggested to explore their own ways to effectively "import" (Xu and Harfitt, 2018) and "translate" high-quality formative assessment (Black and Wiliam, 1998).

Author contributions

QX conceived of the presented idea. QX and AC performed the analytic calculations, performed the numerical simulations and contributed to the final version of the manuscript. DS contributed to important intellectual content in the revised version and the final version of the manuscript.

Acknowledgments

This manuscript would not have been possible without the exceptional support of my supervisor AC. His expertise and encouragement have been an inspiration and kept my work on track. We are grateful for the insightful comments offered by the peer reviewers. We also thank our family, who provide unending support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.990196/full#supplementary-material

References

Abrami, P. C., and Bernard, R. M. (2006). Research on distance education: In defense of field experiments. *Distance Educ.* 27, 5–26.

Al Otaiba, S., Connor, C. M., Folsom, J. S., Greulich, L., Meadows, J., and Li, Z. (2011). Assessment data–informed guidance to individualize kindergarten reading instruction: Findings from a cluster-randomized control field trial. *Elem. Sch. J.* 111, 535–560. doi: 10.1086/659031

Allen, J. (2019). *Does adoption of act aspire periodic assessments support student growth?*. Iowa, IA: ACT, Inc.

Andrade, H. L., Bennett, R. E., and Cizek, G. J. (2019). "Formative assessment: History, definition, and progress," in *Handbook of formative assessment in the disciplines*, eds H. L. Andrade, R. E. Bennett, and G. J. Cizek (New York, NY: Routledge), 3–19. doi: 10.4324/9781315166933-1

Bennett, R. E. (2011). Formative assessment: A critical review. *Assess. Educ. Princ. Policy Pract.* 18, 5–25. doi: 10.1080/0969594X.2010.513678

Berry, R. (2011). Assessment trends in Hong Kong: Seeking to establish formative assessment in an examination culture. *Assess. Educ. Princ. Policy Pract.* 18, 199–211. doi: 10.1080/0969594X.2010.527701

Biggs, J. (1998). Learning from the confucian heritage: So size doesn't matter? *Int. J. Educ. Res.* 29, 723–738. doi: 10.1016/S0883-0355(98)00060-3

Black, H., and Wiliam, D. (1986). "Assessment for learning," in *Assessing educational achievement*, ed. D. L. Nuttall (London: Falmer Press), 7–18.

Black, P. (2015). Formative assessment – an optimistic but incomplete vision. *Assess. Educ. Princ. Policy Pract.* 22, 161–177. doi: 10.1080/0969594X.2014.999643

Black, P., and Wiliam, D. (1998). Assessment and classroom learning. *Assess. Educ. Princ. Policy Pract.* 5, 7–74. doi: 10.1080/0969595980050102

Black, P., and Wiliam, D. (2009). Developing the theory of formative assessment. *Educ. Assess. Eval. Acc.* 21, 5–31. doi: 10.1007/s11092-008-9068-5

Black, P., and Wiliam, D. (2018). Classroom assessment and pedagogy. *Assess. Educ. Princ. Policy Pract.* 25, 551–575. doi: 10.1080/0969594X.2018.1441807

Borenstein, M., Cooper, H., Hedges, L., and Valentine, J. (2009). "Effect sizes for continuous data," in *The handbook of research synthesis and meta-analysis*, 2nd Edn, eds H. Cooper, L. V. Hedges, and J. C. Valentine (New York, NY: Russell Sage Foundation), 221–235.

Borenstein, M., Hedges, L., Higgins, J., and Rothstein, H. (2013). *Comprehensive meta-analysis version 3*. Englewood, CO: Biostat.

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., and Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educ. Meas.* 31, 13–17. doi: 10.1111/j.1745-3992.2012.00251.x

Brookhart, S. M., Moss, C. M., and Long, B. A. (2010). Teacher inquiry into formative assessment practices in remedial reading classrooms. *Assess. Educ. Princ. Policy Pract.* 17, 41–58. doi: 10.1080/09695940903565545

Burns, M. K., Codding, R. S., Boice, C. H., and Lukito, G. (2010). Meta-analysis of acquisition and fluency math interventions with instructional and frustration level skills: Evidence for a skill-by-treatment interaction. *Sch. Psychol. Rev.* 39, 69–83. doi: 10.1080/02796015.2010.12087791

Butler, Y., and Lee, J. (2010). The effects of self-assessment among young learners of English. *Lang. Test.* 27, 5–31. doi: 10.1177/0265532209346370

Cain, M. L. (2015). *The impact of the reading 3D program as a component of formative assessment. Doctoral dissertation*. Charlotte, NC: Wingate University.

Carless, D. (2007). Learning-oriented assessment: Conceptual bases and practical implications. *Innov. Educ. Teach. Int.* 44, 57–66. doi: 10.1080/14703290601081332

Chappius, J. (2009). *Seven strategies for assessment for learing*. Portland, OR: Pearson Assessment Training Institute.

Chen, C., Chen, L., and Horng, W. (2021). A collaborative reading annotation system with formative assessment and feedback mechanisms to promote digital reading performance. *Interact. Learn. Environ.* 29, 848–865. doi: 10.1080/10494820.2019.1636091

Chen, C., Wang, J., and Lin, M. (2017). Enhancement of English learning performance by using an attention-based diagnosing and review mechanism in paper-based learning context with digital pen support. *Univers. Access Inf. Soc.* 18, 141–153. doi: 10.1007/s10209-017-0576-2

Chen, Q., Kettle, M., Klenowski, V., and May, L. (2013). Interpretations of formative assessment in the teaching of English at two Chinese universities: A sociocultural perspective. *Assess. Eval. High Educ.* 38, 831–846. doi: 10.1080/02602938.2012.726963

Cheung, A. C. K., and Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educ. Res.* 45, 283–292. doi: 10.3102/0013189X16656615

Cheung, A. C. K., Xie, C., Zhuang, T., Neitzel, A. J., and Slavin, R. E. (2021). Success for all: A quantitative synthesis of US evaluations. *J. Res. Educ. Eff.* 14, 90–115. doi: 10.1080/19345747.2020.1868031

Cizek, G. J., Andrade, H. L., and Bennett, R. E. (2019). "Formative assessment: History, definition, and progress," in *Handbook of formative assessment in the disciplines*, eds L. A. Heidi and J. C. Gregory (New York, NY: Routledge), 3–19.

Clark, I. (2010). Formative assessment: 'There is nothing so practical as a good theory'. *Aust. J. Educ.* 54, 341–352. doi: 10.1177/000494411005400308

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge.

Connor, C. M., Jakobsons, L., Crowe, E., and Meadows, J. (2009). Instruction, differentiation, and student engagement in reading first classrooms. *Elem. Sch. J.* 109, 221–250. doi: 10.1086/592305

Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., and Underwood, P. (2007). The early years. Algorithm-guided individualized reading instruction. *Science* 315, 464–465. doi: 10.1126/science.1134513

Connor, C. M., Morrison, F. J., Fishman, B., Crowe, E. C., Al Otaiba, S., and Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychol. Sci.* 24, 1408–1419. doi: 10.1177/0956797612472204

Connor, C. M., Morrison, F. J., Schatschneider, C., Toste, J., Lundblom, E., Crowe, E. C., et al. (2011). Effective classroom instruction: Implications of child characteristics by reading instruction interactions on first graders' word reading achievement. *J. Res. Educ. Eff.* 4, 173–207. doi: 10.1080/19345747.2010.510179

Cordray, D. S., Pion, G. M., Brandt, C., and Molefe, A. (2013). *The impact of the measures of academic progress (MAP) program on student reading achievement. Final report*. Washington, DC: NCEE.

Crossouard, B., and Pryor, J. (2012). How theory matters: Formative assessment theory and practices and their different relations to education. *Stud. Philos. Educ.* 31, 251–263.

DerSimonian, R., and Laird, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* 7, 177–188. doi: 10.1016/0197-2456(86)90046-2

Dunn, K., and Mulvenon, S. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Pract. Assess. Res. Eval.* 14, 1–11. doi: 10.4324/9780203462041_chapter_1

Duval, S., and Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *J. Am. Stat. Assoc.* 95, 89–98. doi: 10.1080/01621459.2000.10473905

Earl, L. M. (2012). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin press.

ESSA (2015). *Evidence for ESSA: Standards and procedures*. Available online at: https://content.evidenceforessa.org/sites/default/files/On%20clean%20Word%20doc.pdf (accessed June 20, 2022).

Faber, J. M., and Visscher, A. (2018). The effects of a digital formative assessment tool on spelling achievement: Results of a randomized experiment. *Comput. Educ.* 122, 1–8. doi: 10.1016/j.compedu.2018.03.008

Förster, N., and Souvignier, E. (2014). Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept. *Learn. Instr.* 32, 91–100. doi: 10.1016/j.learninstruc.2014.02.002

Förster, N., and Souvignier, E. (2015). Effects of providing teachers with information about their students' reading progress. *Sch. Psychol. Rev.* 44, 60–75. doi: 10.17105/SPR44-1.60-75

Förster, N., Kawohl, E., and Souvignier, E. (2018). Short- and long-term effects of assessment-based differentiated reading instruction in general education on reading fluency and reading comprehension. *Learn. Instr.* 56, 98–109. doi: 10.1016/j.learninstruc.2018.04.009

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

Fuchs, L. S., Fuchs, D., Hamlett, C. L., and Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement, using a reading maze task. *Except. Child.* 58, 436–450.

Fuchs, L. S., Butterworth, J. R., and Fuchs, D. (1989). Effects of ongoing curriculum-based measurement on student awareness of goals and progress. *Educ. Treat. Child.* 12, 63–72.

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., and Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A

meta-analysis of instructional components. *Rev. Educ. Res.* 79, 1202–1242. doi: 10.3102/0034654309334431

Gersten, R., Haymond, K., Newman-Gonchar, R., Dimino, J., and Jayanthi, M. (2020). Meta-analysis of the impact of reading interventions for students in the primary grades. *J. Res. Educ. Eff.* 13, 401–427. doi: 10.1080/19345747.2019.1689591

Glass, G. V., McGaw, B., and Smith, M. L. (1981). *Meta-analysis in social research*. Thousand Oaks, CA: SAGE Publications.

Graham, S., Hebert, M., and Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *Elem. Sch. J.* 115, 523–547. doi: 10.1086/681947

Guan, M., and Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychon. Bull. Rev.* 23, 74–86. doi: 10.3758/s13423-015-0868-6

Hall, T. E., Cohen, N., Vue, G., and Ganley, P. (2014). Addressing learning disabilities with udl and technology. *Learn. Disabil. Q.* 38, 72–83. doi: 10.1177/0731948714544375

Hartmeyer, R., Stevenson, M. P., and Bentsen, P. (2018). A systematic review of concept mapping-based formative assessment processes in primary and secondary science education. *Assess. Educ. Princ. Policy Pract.* 25, 598–619. doi: 10.1080/0969594X.2017.1377685

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: NY: Routledge.

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Heitink, M. C., Van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., and Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educ. Res. Rev.* 17, 50–62. doi: 10.1016/j.edurev.2015.12.002

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Thousand Oaks, CA: Corwin. doi: 10.4135/9781452219493

Higgins, R., Hartley, P., and Skelton, A. (2001). Getting the message across: The problem of communicating assessment feedback. *Teach. High. Educ.* 6, 269–274. doi: 10.1080/13562510120045230

Hu, G. (2002). Potential cultural resistance to pedagogical imports: The case of communicative language teaching in China. *Lang. Cult. Curric.* 15, 93–105. doi: 10.1080/07908310208666636

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

Ivanic, R., Clark, R., and Rimmershaw, R. (2000). *Student writing in higher education: New contexts*. Maidenhead: Open University Press.

Johnson, L., Graham, S., and 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. *J. Learn. Disabil.* 30, 80–91. doi: 10.1177/002221949703000107

Jönsson, A. (2020). Definitions of formative assessment need to make a distinction between a psychometric understanding of assessment and "evaluative judgment". *Front. Educ.* 5:2. doi: 10.3389/feduc.2020.00002

Kennedy, M. M. (2009). *Inside teaching*. Cambridge, MA: Harvard University Press.

Kingston, N., and Broaddus, A. (2017). The use of learning map systems to support the formative assessment in mathematics. *Educ. Sci.* 7:41. doi: 10.3390/educsci7010041

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

Kingston, N., and Nash, B. (2012). How many formative assessment angels can dance on the head of a meta-analytic pin: 0.2. *Educ. Meas. Issues Pract.* 31, 18–19. doi: 10.1111/j.1745-3992.2012.00254.x

Kintsch, W., and Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychol. Rev.* 85:363. doi: 10.1037/0033-295X.85.5.363

Kluger, A. N., and DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychol. Bull.* 119, 254–284. doi: 10.1037/0033-2909.119.2.254

Klute, M., Apthorp, H., Harlacher, J., and Reale, M. (2017). *Formative assessment and elementary school student academic achievement: A review of the evidence*. Washington, DC: Regional Educational Laboratory Central.

Konstantopoulos, S., Miller, S. R., van der Ploeg, A., and Li, W. (2016). Effects of interim assessments on student achievement: Evidence from a large-scale experiment. *J. Res. Educ. Eff.* 9, 188–208. doi: 10.1080/19345747.2015.1116031

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educ. Res.* 49, 241–253. doi: 10.3102/0013189X20912798

Lane, R., Parrila, R., Bower, M., Bull, R., Cavanagh, M., Forbes, A., et al. (2019). *Literature review: Formative assessement evidence and practice*. Melbourne, VI: AITSL.

Lau, Kl (2020). The effectiveness of self-regulated learning instruction on students' classical Chinese reading comprehension and motivation. *Read. Writ.* 33, 2001–2027. doi: 10.1007/s11145-020-10028-2

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., and Warschauer, M. (2020). The effectiveness and features of formative assessment in us k-12 education: A systematic review. *Appl. Meas. Educ.* 33, 124–140. doi: 10.1080/08957347.2020.1732383

Leighton, J., and Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511611186

Liao, Y.-K. C. (1999). Hypermedia and students'achievement: A meta-analysis. *EdMedia Innov. Learn.* 8, 1398–1399.

Lipsey, M. W., and Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: SAGE Publications.

López-López, J. A., Marín-Martínez, F., Sánchez-Meca, J., Van den Noortgate, W., and Viechtbauer, W. (2014). Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *Br. J. Math. Stat. Psychol.* 67, 30–48. doi: 10.1111/bmsp.12002

Luke, A., Woods, A., and Weir, K. (2013). *Curriculum, syllabus design, and equity: A primer and model*. Milton Park: Routledge. doi: 10.4324/9780203833452

Marcotte, A. M., and Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *J. Sch. Psychol.* 47, 315–335. doi: 10.1016/j.jsp.2009.04.003

Martens, B., Eckert, T., Begeny, J., Lewandowski, L., DiGennaro Reed, F., Montarello, S., et al. (2007). Effects of a fluency-building program on the reading performance of low-achieving second and third grade students. *J. Behav. Educ.* 16, 38–53. doi: 10.1007/s10864-006-9022-x

McCurdy, B. L., and Shapiro, E. S. (1992). A comparison of teacher-, peer-, and self-monitoring with curriculum-based measurement in reading among students with learning disabilities. *J. Spec. Educ.* 26, 162–180. doi: 10.1177/002246699202600203

McLaughlin, M. (2012). Reading comprehension: What every teacher needs to know. *Read. Teach.* 65, 432–440. doi: 10.1002/TRTR.01064

Miller, D. M., Scott, C. E., and McTigue, E. M. (2018). Writing in the secondary-level disciplines: A systematic review of context, cognition, and content. *Educ. Psychol. Rev.* 30, 83–120. doi: 10.1007/s10648-016-9393-z

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med.* 6:e1000097. doi: 10.1371/journal.pmed.1000097

Mok, M. M. C. (2012). "Assessment reform in the Asia-Pacific region: The theory and practice of self-directed learning oriented assessment," in *Self-directed learning oriented assessments in the Asia-Pacific. Education in the Asia-Pacific region: Issues, concerns and prospects*, ed. M. Mok (Dordrecht: Springer), 3–22. doi: 10.1007/978-94-007-4507-0_1

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

OECD (2008). *Assessment for learning formative assessment*. Paris: OECD.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *J. Educ. Stat.* 8, 157–159. doi: 10.2307/1164923

Palmer, E., and Devitt, P. (2014). The assessment of a structured online formative assessment program: A randomised controlled trial. *BMC Med. Edu.* 14:8. doi: 10.1186/1472-6920-14-8

Peters, M. T., Hebbecker, K., and Souvignier, E. (2021). Effects of providing teachers with tools for implementing assessment-based differentiated reading instruction in second grade. *Assess. Eff. Interv.* 47, 157–169. doi: 10.1177/15345084211014926

Pigott, T. D., and Polanin, J. R. (2019). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Rev. Educ. Res.* 90, 24–46. doi: 10.3102/0034654319877153

Polanin, J. R., and Pigott, T. D. (2015). The use of meta-analytic statistical significance testing. *Res. Synth. Methods* 6, 63–73. doi: 10.1002/jrsm.1124

Popham, W. J. (2008). Formative assessment: Seven stepping-stones to success. *Prin. Leadersh.* 9, 16–20.

Roskos, K., and Neuman, S. B. (2012). Formative assessment: Simply, no additives. *Read. Teach.* 65, 534–538.

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., and Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *J. Educ. Psychol.* 109, 1049–1066. doi: 10.1037/edu0000190

Sandelowski, M., Voils, C. I., Leeman, J., and Crandell, J. L. (2012). Mapping the mixed methods-mixed research synthesis terrain. *J. Mix. Methods Res.* 6, 317–331. doi: 10.1177/1558689811427913

Schmidt, F. L., Oh, I. S., and Hayes, T. L. (2009). Fixed-versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *Br. J. Math. Stat. Psychol.* 62, 97–128. doi: 10.1348/000711007X255327

Sedlmeier, P., and Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychol. Bull.* 105, 309–316. doi: 10.1037/0033-2909.105.2.309

See, B. H., Gorard, S., Lu, B., Dong, L., and Siddiqui, N. (2021). Is technology always helpful?: A critical review of the impact on learning outcomes of education technology in supporting formative assessment in schools. *Res. Pap. Educ.* 1–33. doi: 10.1080/02671522.2021.1907778

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton, Mifflin and Company.

Shimojima, Y., and Arimoto, M. (2017). Assessment for learning practices in Japan: Three steps forward, two steps back. *Assess. Matters* 11:2017. doi: 10.18296/am.0023

Simmons, D. C., Kim, M., Kwok, O. M., Coyne, M. D., Simmons, L. E., Oslund, E., et al. (2015). Examining the effects of linking student performance and progression in a Tier 2 kindergarten reading intervention. *J. Learn. Disabil.* 48, 255–270. doi: 10.1177/0022219413497097

Slavin, R. E. (2013). Effective programmes in reading and mathematics: Lessons from the best evidence encyclopaedia. *Sch. Eff. Sch. Improv.* 24, 383–391. doi: 10.1080/09243453.2013.797913

Slavin, R. E., and Cheung, A. C. K. (2019). Evidence-based reform in education: Responses to critics. *Sci. Insigt. Edu. Front.* 2, 65–69. doi: 10.15354/sief.19.ar027

Slavin, R., and Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educ. Eval. Policy Anal.* 31, 500–506.

Slavin, R. E., Cheung, A. C. K., Holmes, G., Madden, N. A., and Chamberlain, A. (2013). Effects of a data-driven district reform model on state assessment outcomes. *Am. Educ. Res. J.* 50, 371–396. doi: 10.3102/0002831212466909

Snow, R. E. (1986). Individual differences and the design of educational programs. *Am. Psychol.* 41:1029. doi: 10.1037/0003-066X.41.10.1029

Spector, J. M., Ifenthaler, D., Samspon, D., Yang, L., Mukama, E., Warusavitarana, A., et al. (2016). Technology enhanced formative assessment for 21st century learning. *Educ. Technol. Soc.* 19, 58–71.

Steiner-Khamsi, G. (2014). Cross-national policy borrowing: Understanding reception and translation. *Asia Pac. Educ. Rev.* 34, 153–167. doi: 10.1080/02188791.2013.875649

Sterne, J. A., Becker, B. J., and Egger, M. (2005). "The funnel plot," in *Publication bias in meta-analysis: Prevention, assessment and adjustments*, eds H. R. Rothstein, A. J. Sutton, and M. Borenstein (Chichester: Wiley), 75–98. doi: 10.1002/0470870168.ch5

Stigler, J. W., and Hiebert, J. (1998). Teaching is a cultural activity. *Teach. Educ.* 22, 4–11.

Stigler, J. W., and Hiebert, J. (2009). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York, NY: Simon and Schuster.

Sung, Y.-T., Chang, K.-E., and Liu, T.-C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Comput. Educ.* 94, 252–275. doi: 10.1016/j.compedu.2015.11.008

Tomasik, M. J., Berger, S., and Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Front. Psychol.* 9:2245. doi: 10.3389/fpsyg.2018.02245

Tomlinson, C. A. (2001). *How to differentiate instruction in mixed-ability classrooms*. Alexandria, VA: ASCD.

Topping, K. J., and Fisher, A. M. (2003). Computerised formative assessment of reading comprehension: Field trials in the UK. *J. Res. Read.* 26, 267–279. doi: 10.1111/1467-9817.00202

Tsai, F.-H., Tsai, C.-C., and Lin, K.-Y. (2015). The evaluation of different gaming modes and feedback types on game-based formative assessment in an online learning environment. *Comput. Educ.* 81, 259–269. doi: 10.1016/j.compedu.2014.10.013

Van der Kleij, F. M., Cumming, J. J., and Looney, A. (2017). Policy expectations and support for teacher formative assessment in Australian education reform. *Assess. Educ. Princ. Policy Pract.* 25, 620–637. doi: 10.1080/0969594X.2017.1374924

Wang, A., Firmender, J. M., Power, J. R., and Byrnes, J. P. (2016). Understanding the program effectiveness of early mathematics interventions for prekindergarten and kindergarten environments: A meta-analytic review. *Early Educ. Dev.* 27, 692–713. doi: 10.1080/10409289.2016.1116343

Wang, T. (2008). Web-based quiz-game-like formative assessment: Development and evaluation. *Comput. Educ.* 51, 1247–1263. doi: 10.1016/j.compedu.2007.11.011

What Works Clearinghouse (2020). *What works clearinghouse standards handbook, version 4.1 ed.* Washington, DC: Institute of Education Sciences: National Center for Education Evaluation and Regional Assistance.

Wiliam, D. (2011). What is assessment for learning? *Stud. Educ. Evaluation* 37, 3–14. doi: 10.1016/j.stueduc.2011.03.001

Wiliam, D., and Thompson, M. (2007). "Integrating assessment with learning. What will it take to make it work?," in *The future of assessment*, ed. C. A. Dwyer (New York, NY: Routledge), 53–82. doi: 10.4324/9781315086545-3

Wisniewski, B., Zierer, K., and Hattie, J. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Front. Psychol.* 10:3087. doi: 10.3389/fpsyg.2019.03087

Xu, Y., and Harfitt, G. (2018). Is assessment for learning feasible in large classes? Challenges and coping strategies from three case studies. *Asia Pacific J. Educ.* 47, 472–486. doi: 10.1080/1359866X.2018.1555790

Yan, Z., Chiu, M. M., and Ko, P. Y. (2020). Effects of self-assessment diaries on academic achievement, self-regulation, and motivation. *Assess. Educ. Princ. Policy Pract.* 27, 562–583. doi: 10.1080/0969594X.2020.1827221

Zhang, S., and Thompson, N. (2004). DIALANG: A diagnostic language assessment system. *Can. Mod. Lang. Rev.* 61, 290–293. doi: 10.1353/cml.2005.0011